PREDICTING THE SPREAD AND CONTROL OF DIPHTHERIA USING MACHINE LEARNING: A DATA-DRIVEN APPROACH TO DISEASE SURVEILLANCE AND INTERVENTION STRATEGY OPTIMIZATION

Abah Roseline Toyin[1], Leona Concord Diala[2*] & Agbo Christiana Ene[3]

[1,3]Dept of Mathematics, University of Abuja, Abuja.

[2]Dept of Mathematics, Margaret Lawrence University, Abuja.

[*]Corresponding author Email: dialaleona@gmail.com

Abstract

This study presents a comprehensive statistical analysis of diphtheria outbreaks across several African countries using a cleaned dataset from WHO data. Summary statistics reveal Nigeria as the most affected nation in terms of suspected cases and deaths, suggesting a possible correlation with population size, reporting efficiency, or outbreak severity. A notably high case fatality rate (CFR) in Mauritania may be attributed to limited healthcare access and delayed diagnoses. Correlation analysis indicates a strong positive association between deaths and both suspected and confirmed cases, with perfect correlations noted between deaths and both epidemiologically linked and clinical compatible cases. These findings suggest that increased diagnostic efforts may mitigate the CFR by identifying and managing cases earlier. A multiple linear regression model further underscores the significant contributions of lab-confirmed, epidemiologically linked, and clinical compatible cases to death outcomes, all with statistically significant p-values. The model's prediction accuracy is supported by a near-perfect regression fit. The analysis identifies Mauritania and Niger as high-risk countries for elevated CFRs, warranting urgent public health interventions. This study underscores the value of integrated data analytics in enhancing disease surveillance and informing evidence-based responses to diphtheria outbreaks.

Keywords: *Diphtheria, Machine learning, Infectious disease surveillance, Intervention strategies, Multiple Linear Regression*.

Introduction

Diphtheria is an acute bacterial infection caused by toxin-producing strains of Corynebacterium diphtheriae and, in some cases, also by *C.* ulcerans bacteria. It is transmitted from person to person through close physical and respiratory contact (Edwards et al., 2018). The early signs of respiratory diphtheria are sore throat, loss of appetite, and slight fever. Severe cases of diphtheria may develop a swollen neck and obstructed airway.

Despite the availability of effective vaccines, diphtheria continues to pose a significant public health threat in low- and middle-income countries, particularly where immunization coverage is suboptimal or where healthcare systems are under-resourced (WHO, 2023).

Recent outbreaks in parts of Africa, including Nigeria, Niger, and Mauritania, have underscored persistent vulnerability to diphtheria-related morbidity and mortality (WHO African Region, 2024). In Nigeria, these outbreaks involved over 18239 suspected cases, compounded by poor vaccination uptake, inadequate diagnostic capacity, weak surveillance, delayed response measures, community awareness and poor access to clean water and sanitation practices (Agrawal et al., 2023)

Analyzing the epidemiological patterns of diphtheria using statistical and predictive modeling approaches enables public health authorities to identify high-risk regions, understand causal factors, and allocate resources more efficiently. This study employs statistical correlation and multiple linear regression techniques to assess the relationships between suspected, confirmed, and fatal cases of diphtheria across selected African countries. The goal is to generate actionable insights that can inform public health strategies and enhance epidemic preparedness in vulnerable regions.

The effectiveness of the diphtheria toxoid vaccine has been well documented, with substantial reductions in disease burden in countries with high vaccination rates (Galazka, 2000). However, in low and middle income cuntries (LMICs), challenges such as vaccine hesitancy, limited access to health services, and logistical barriers contribute to sporadic but often severe outbreaks (Dandinarasaiah et al., 2013). For instance, Nigeria has witnessed repeated diphtheria outbreaks, linked to poor diagnostic capacity and delayed medical interventions (Oteri et al., 2020). Also, Socioeconomic factors play a significant role in the disease's awareness and dissemination. In Nigeria, the majority of patients with this illness are from the lower and moderate socioeconomic strata (Alege et al., 2025).

Diphtheria, although considered a vaccine preventable disease, has re-emerged as a public health threat in several low and middle income countries in recent years. Despite decades of global control, outbreaks in countries such as Nigeria, Yemen, India, and Venezuela since 2015 highlight persistent vulnerabilities in health systems and immunization programs (Dandinarasaiah et al., 2012). When large decontamination happens in the population and they are removed from the population, it will help to decreasing the number of persons exposed to diseases hence aiding in stopping the transmission of the virus (Abah et al., 2024). These outbreaks have been linked to lapses in vaccine coverage, conflict related disruptions, and underfunded surveillance systems.

According to the World Health Organization (2023), diphtheria remains endemic in some regions due to delayed diagnosis, inadequate healthcare infrastructure, and poor laboratory confirmation capacity. The case fatality rate (CFR) can rise significantly in such contexts, as seen during recent outbreaks in Africa and Asia. For instance, a study by Dureab et al. (2019) on Yemen's diphtheria epidemic emphasized the critical role of conflict in disrupting healthcare delivery and immunization efforts, resulting in increased transmission and mortality. Political unrest and insurgency have negative impact on healthcare delivery by limiting access to primary healthcare centers (Witter et al., 2015).

The data from Africa suggest that diagnostic inefficiencies and delays in health seeking behavior contribute to high CFRs in countries like Mauritania and Niger. These findings align with M. J. Nsuami et al. (2013), who reported that strengthening disease surveillance systems particularly in resource limited settings is essential for early detection and response, noting that none of the initial diphtheria cases were recognized for seven months. The most effective way to stop the disease's transmission is still to isolate affected people and conduct a public education campaign (Musa et al., 2025).

As public health systems increasingly rely on data driven strategies, integrated analytics and real time modeling offer a practical approach to reducing the burden of diphtheria and similar vaccine preventable diseases. This study builds on recent literature by exploring the relationship between suspected and confirmed cases and mortality through regression based modeling, with the goal of enhancing epidemiological preparedness in Africa.

Literature Review

Diphtheria has not been eradicated worldwide despite extensive vaccination campaigns. It still resurfaces in areas with inadequate immunization programs, inadequate surveillance capabilities and societal unrest. Recent research shows that outbreaks are no longer isolated past events but are re-emerging public health threats, particularly in low- and middle-income countries. It is convenient to review related prior work in order to gain a better understanding of disease transmission and control through machine learning.

(Osarenren et. Al., 2024) conducted a comprehensive review of global strategies for addressing diphtheria resurgence. Epidemiological, clinical impact and prevention shows that diphtheria outbreaks continue to occur due to gaps in vaccination coverage, failures in health infrastructure and socio-political instability. Past and recent outbreaks in regions such as Bangladesh, Venezuela and Yemen show that even countries with some level of immunization history are vulnerable when routine immunization is incomplete. This review shows that diphtheria remains a public health concern wherever immunization systems are not robust and resilient. A study conducted by (Alege et. Al., 2025) in Nigeria involving hospitalized diphtheria cases from July 1, 2023, to April 30, 2024 found that the disease continues to inflict high mortality among children, particularly those with severe clinical presentations and incomplete vaccination status. Common predictors of death include respiratory abnormalities, hypoxemia and neck swelling with a reported mortality rate of 23.5% among confirmed cases. This study not only provides current epidemiological evidence from a major outbreak setting but also emphasizes the clinical severity of unchecked diphtheria in under-immunized populations.

A study conducted by (Hussein et. Al., 2025) reported a weekly surveillance data in Somalia in 2025 which found that 87% of suspected diphtheria cases had no documented immunization history. Children under five years of age accounted for a large proportion of cases. Despite efforts such as ring vaccination and diphtheria antitoxin distribution, constrained vaccine and antitoxin sullies, insecurity and access barriers significantly limited outbreak control. This situation illustrates how conflict and healthy system fragility worsen the risk of diphtheria transmission and associated mortality. (Okhuarobo et. Al., 2025) highlighted the importance of cultivating effective surveillance systems in evaluating diphtheria in Nigeria. The surveillance system in Kaduna state identified limitations in early case detection, reporting consistency, laboratory confirmation capacity and information feedback. All these factors hinder effective outbreak response. This aligns with broader literature showing that surveillance gaps and not just immunization gaps is crucial to determine outbreak trajectories and mortality outcomes.

A study of seasonal variations in public perception in Northern Nigeria was conducted by (Erim et. Al., 2025). It was found that fluctuating awareness and attitudes toward diphtheria and vaccination in that zone. Public perceptions and behaviors around diphtheria also affect disease outcomes and control strategies. Therefore, public awareness and risk communication are variables that can influence vaccine uptake and health-seeking behavior. These social factors are increasingly recognized as critical determinants of immunization coverage and outbreak control.

A LightGBM model was developed by (Mc Cord—De Iaco et al., 2023) to create two sets of predictive models for each pathogen: one based solely on clinical symptoms and another combining symptoms with laboratory data. The models' performance was assessed by metrics such as recall, F1 score, and AUC, with results showing fair accuracy; for instance, pertussis recall was around 0.72–0.74, and respiratory syncytial virus (RSV) recall was approximately 0.68–0.71, with similar F1 scores for models with and without lab data. (Salim et al., 2024) examined the spread and trends of diphtheria in Nigeria, particularly focusing on the significant out break that began in December 2022. The study utilized desktop research, sourcing data from secondary sources like the Nigeria Centre for Disease Control and the WHO Global Health Library, and analyzed confirmed cases using descriptive statistics. Results indicated 111 confirmed cases and 22 deaths across four states, highlighting a concerning case fatality rate and suggesting that the current outbreak is one of the worst in recent.

(Melchane et. al., 2024) made use of machine learning (ML) and deep learning (DL) to evaluate the role of artificial intelligence (AI) in predicting and preventing infectious diseases. The study categorized AI-based research efforts into three primary areas namely, predictions using public health data to anticipate disease spread, predictions using individual medical data for disease detection, and integrated predictions that combine both data types to estimate outbreak magnitudes The results show that AI models have been success fully used in various infectious disease contexts (e.g., COVID-19, influenza, dengue) with considerable accuracy.

A robust predictive model for infectious disease case numbers, using diphtheria as a representative case with a focus on improving the accuracy and interpretability of disease forecasts in the presence of data uncertainty was developed by (Musa et al., 2025). The methodology involves collecting real diphtheria case data from the World Health Organization (WHO) spanning 1980 to 2020 and implementing an adaptive network based fuzzy inference system (ANFIS) using Type-2 fuzzy sets, comparing its performance against classical Type-1 fuzzy systems and other statistical/machine learning approaches. The results demonstrate that the Type-2 fuzzy system outperforms both Type-1 systems and ARIMA in terms of root mean square error (RMSE) and mean absolute percentage error (MAPE), particularly in settings with higher data uncertainty.

Methodology

This study adopts a quantitative, cross-sectional, data-driven analytical design to investigate the epidemiological patterns and mortality determinants of diphtheria across selected African countries. The analysis integrates descriptive statistics, correlation analysis, and multiple linear regression modeling to explore relationships between suspected cases, confirmed case categories, and deaths due to diphtheria. The dataset used in this study was obtained from World Health Organization (WHO) surveillance records on diphtheria outbreaks across selected African countries. Prior to analysis, the dataset was examined, no missing values were detected. Summary statistics were computed to assess data distribution, variability and central tendencies.

Descriptive statistics, correlation analysis and multilinear regression model were the statistical tools used for analysis. R software version 4.4.2 was used to analyze the data.

Results

```
> print(numeric_summary)
 Suspectec_cases      Deaths             CFR          Lab_confirmed  Epidemiologically_linked Clinical_compatible
 Min.   :   20   Min.   :  1.0   Min.   : 1.30   Min.   :  0   Min.   :  0.0            Min.   :    0
 1st Qu.:   79   1st Qu.:  6.0   1st Qu.: 2.60   1st Qu.:  9   1st Qu.:  0.0            1st Qu.:   20
 Median : 2721   Median : 87.0   Median : 3.20   Median : 39   Median : 94.0            Median : 2155
 Mean   : 5598   Mean   :165.6   Mean   : 8.46   Mean   :117   Mean   :141.4            Mean   : 3450
 3rd Qu.: 2754   3rd Qu.:141.0   3rd Qu.: 5.20   3rd Qu.:191   3rd Qu.:102.0            3rd Qu.: 2453
 Max.   :22417   Max.   :593.0   Max.   :30.00   Max.   :346   Max.   :511.0            Max.   :12620
 Total_confirmed
 Min.   :    9
 1st Qu.:   20
 Median : 2448
 Mean   : 3708
 3rd Qu.: 2586
 Max.   :13477
```

Figure 1

Figure 1 shows the summary statistics of the variables within the diphtheria dataset, including metrics like mean, median, standard deviation and range for each variable. The absence of missing values indicates that the dataset is clean and ready for analysis.

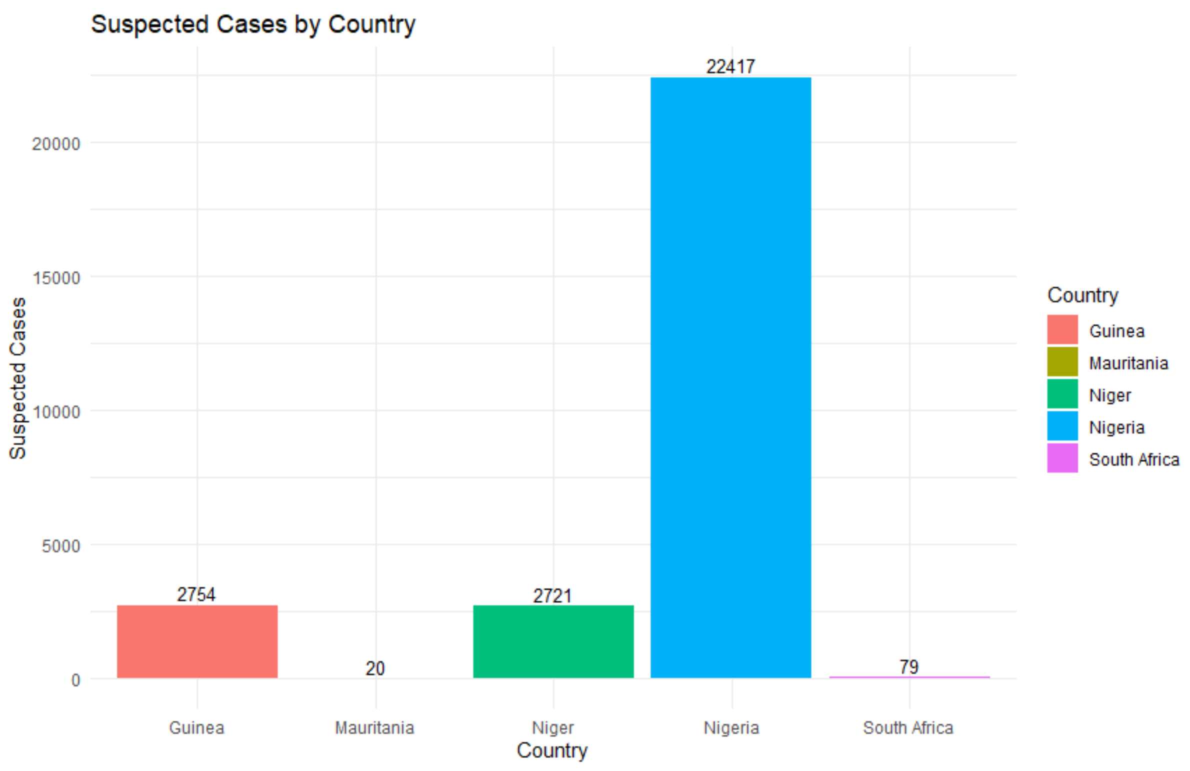

Figure 2

The figure above indicates that Nigeria is the country most affected by diphtheria in terms of suspected cases, which could be linked to factors such as population size, reporting practices or outbreak intensity while Mauritania has the least number of susceptible cases.

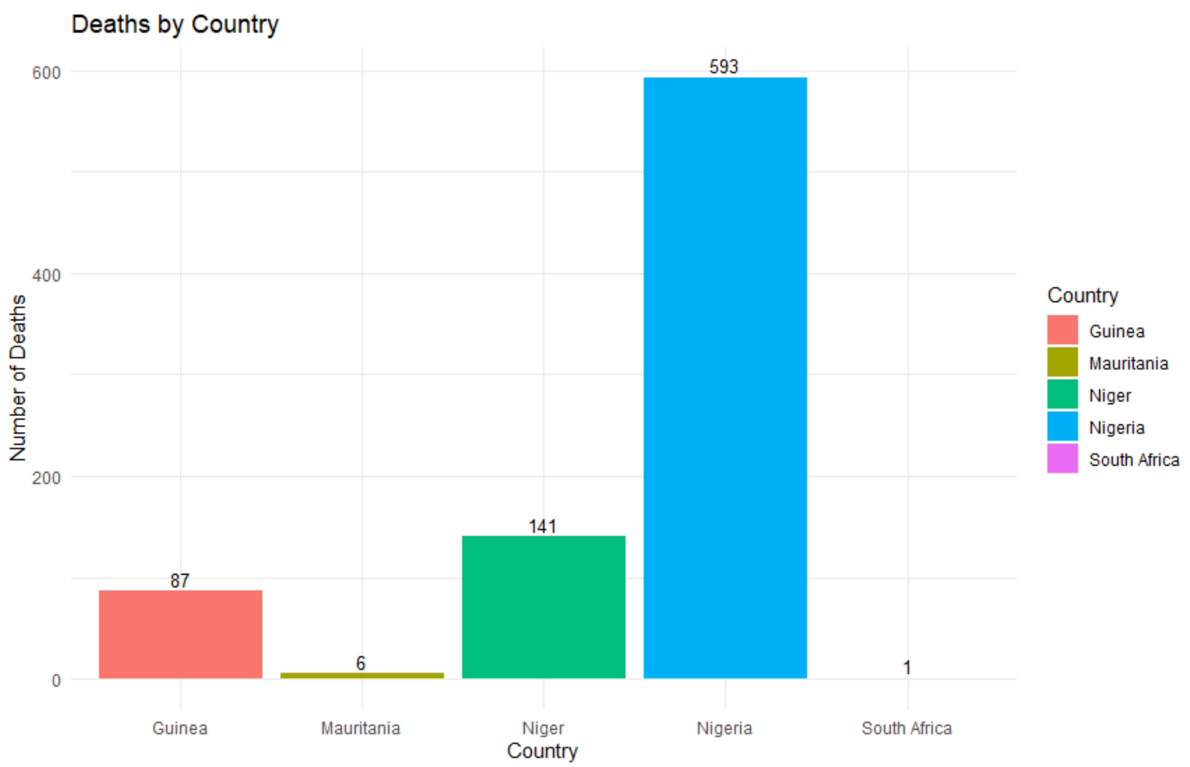

Figure 3

Nigeria has the highest number of deaths, followed by Niger and Guinea. South Africa has the lowest number of deaths. The higher death toll in Nigeria and Niger can be a result of severe outbreaks or poor timely medical intervention.

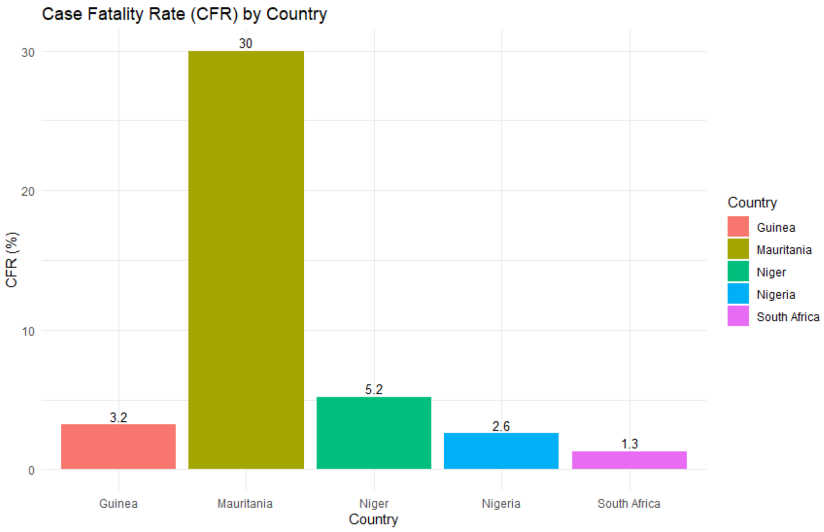

Figure 4

High CFR in Mauritania may indicate limited healthcare resources, late diagnosis or delayed treatment, leading to a higher proportion of fatalities relative to confirmed cases.

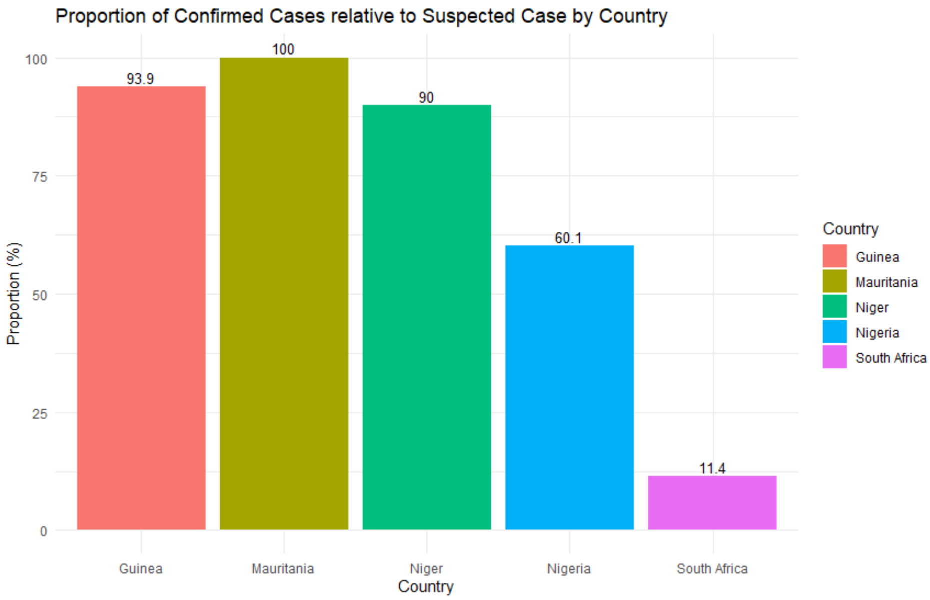

Figure 5

Figure 5, shows the proportion of confirmed cases (lab confirmed, epidemiologically linked, and clinical compatible) relative to the total suspected cases for each country. Countries with higher proportions of confirmed cases (relative to suspected cases) demonstrate poor diagnostic capacity while lower proportions may indicate better lab confirmation or access to healthcare.

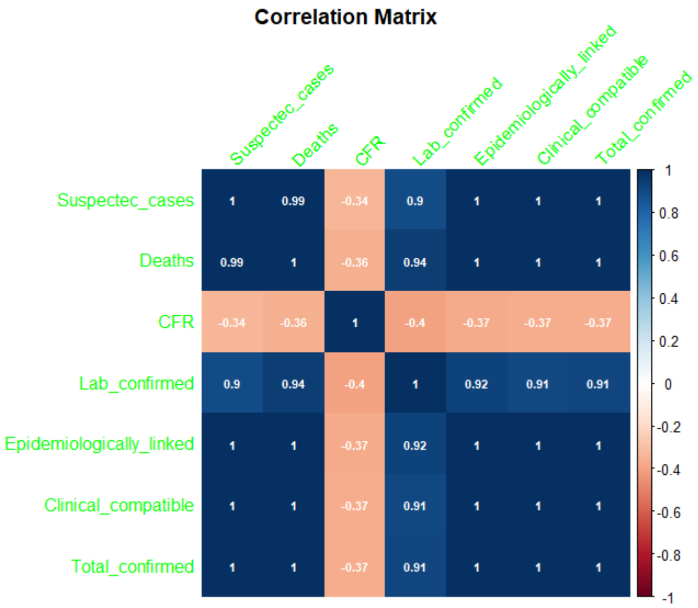

Figure 6

The figure above shows the correlation of different variables.

There is a strong positive correlation between deaths with suspected cases and lab confirmed cases which suggests that countries with suspected cases and lab confirmed cases experience higher death counts. There is perfect correlation between deaths with epidemiologically linked and clinical compatible cases which indicates that the later could cause higher severity, leading to more deaths. The moderate negative correlation between CFR with suspected cases, deaths, lab confirmed, epidemiologically linked and clinical compatible might show that more proper diagnostic exercise reduces the proportion of severe cases, thereby lowering CFR. Epidemiologically linked and clinical compatible variables are perfectly correlated with suspected cases and deaths.

```
> summary(regression_model)

Call:
lm(formula = Deaths ~ Lab_confirmed + Epidemiologically_linked + 
    Clinical_compatible, data = Diphtheria)

Residuals:
       1        2        3        4        5 
-0.05436  0.18460  0.10223  0.64701 -0.87947 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)              17.32264    1.56442   11.07   0.0573 .
Lab_confirmed            -1.71591    0.13122  -13.08   0.0486 *
Epidemiologically_linked 17.06906    1.07628   15.86   0.0401 *
Clinical_compatible      -0.59848    0.04009  -14.93   0.0426 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.113 on 1 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 6.508e+04 on 3 and 1 DF,  p-value: 0.002882
```

Figure 7

The figure above is the summary of multiple linear regression model that shows how the independent variables (Lab confirmed, epidemiologically linked and Clinical compatible) can contribute to the number of deaths. Each independent variable has a p-value which is less than 0.05, which indicates the variables have significant impact on death cases.

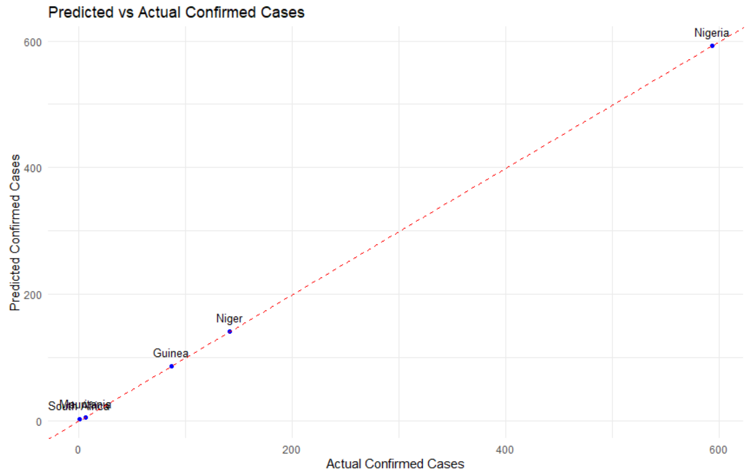

Figure 8

This is the graph fitting of multi-linear regression plot. Since the points are on the line, it indicates that the multi-linear model is accurately predicting deaths based on the independent variables aforementioned.

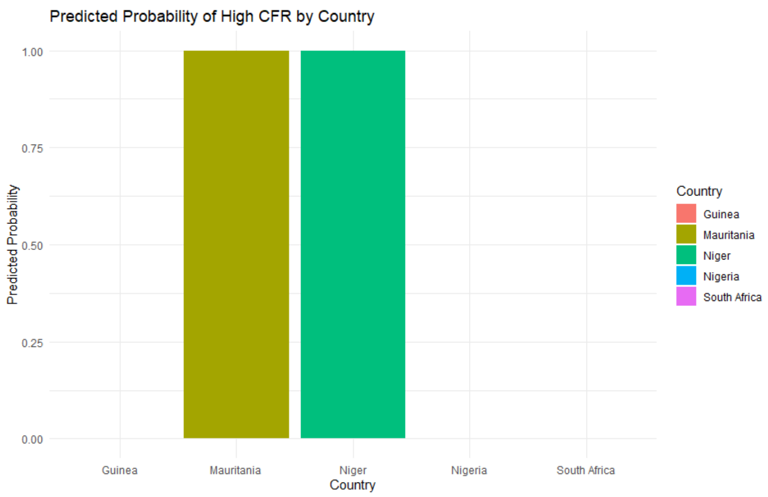

Figure 9

The model predicts that Mauritania and Niger have a higher probability of a high CFR compared to other countries and may require interventions to reduce the mortality rates.

Discussion of Result

The findings emphasize the urgent need to strengthen diagnostic systems and healthcare capacity in African countries where diphtheria outbreaks persist. The strong correlation between confirmed cases and mortality suggests that early and accurate diagnosis is critical to reducing fatalities. High CFRs in countries like Mauritania signal systemic gaps in outbreak response and access to timely medical care.

The regression model not only confirms the significance of various types of confirmed cases in predicting deaths but also serves as a practical tool for identifying high-risk regions. Such predictive modeling supports proactive public health planning and resource allocation, especially in contexts with limited diagnostic reach.

This study also highlights how machine learning and statistical techniques can enhance traditional surveillance methods by providing actionable insights. Integrating these approaches into national disease monitoring frameworks could significantly improve diphtheria response strategies.

Overall, this analysis provides evidence-based recommendations for improving infectious disease surveillance and controlling diphtheria spread through targeted interventions, especially in high-risk and resource-limited settings.

Conclusion

This analysis reveals critical insights into the epidemiology of diphtheria across African countries. The data highlights Nigeria as the most severely affected, both in suspected cases and mortality, while Mauritania's elevated CFR points to systemic healthcare challenges.

Correlation and regression analyses affirm that confirmed case categories significantly influence death rates, with strong statistical relevance. The multi-linear regression model accurately predicts mortality trends and identifies Mauritania and Niger as priority countries for intervention. These findings reinforce the importance of early diagnosis, improved laboratory capacity, and strengthened healthcare systems in reducing diphtheria-related deaths.

References

Abah, R. T., Zhiri, A. B., Oshinubi, K., & Adeniji, A. (2024). Mathematical analysis and simulation of Ebola virus disease spread incorporating mitigation measures. *Franklin Open*, 100066. https://doi.org/10.1016/j.fraope.2023.100066.

Alege, A., Ibrahim, O. R., Ibraheem, R. M., Aladesua, O., Lugga, A. S., Yahaya, Y. Y., ... & Suleiman, B. M. (2025). Clinical presentation and predictors of hospital mortality of diphtheria in Nigeria, July 2023 to April 2024: a single-center study. *BMC Infectious Diseases*, *25*(1), 8.

Dureab, F., Al-Sakkaf, M., Ismail, O., Kuunibe, N., Krisam, J., Müller, O., & Jahn, A. (2019). Diphtheria outbreak in Yemen: the impact of conflict on a fragile health system. *Conflict and Health*, *1*. https://doi.org/10.1186/s13031-019-0204-2

Dandinarasaiah, M., Vikram, B. K., Krishnamurthy, N., Chetan, A. C., & Jain, A. (2012).

Diphtheria Re-emergence: Problems Faced by Developing Countries. *Indian Journal of Otolaryngology and Head & Neck Surgery*, *4*, 314–318. doi.org/10.1007/s12070-012-0518-5

Edwards, D., Kent, D., Lester, C., Brown, C. S., Murphy, M. E., Brown, N. M., Sule, O., Itani, A., Chand, M., Trindall, A., Pearson, C., Roddick, I., Fry, N. K., Hoffmann, J., Iyanger, N., Kemp, L., White, J., Javid, B., Ramsay, I. D., … Reacher, M. (2018). Transmission of toxigenic Corynebacterium diphtheriae by a fully immunised resident returning from a visit to West Africa, United Kingdom, 2017. *Eurosurveillance*, *39*. https://doi.org/10.2807/1560-7917.es.2018.23.39.1700681

Erim, A., Lawal, H. O., Onyeaghala, C. A., Chukwu-Mba, C. C., Ajibade, O. O., Okafor, U. G., & Nzedibe, O. E. (2025). Seasonal variations in public perceptions of diphtheria in Northern Nigeria. *BMC Public Health*, *25*(1), 2146.

Galazka, A. (2000). The Changing Epidemiology of Diphtheria in the Vaccine Era. *The Journal of Infectious Diseases*, *s1*, S2–S9. https://doi.org/10.1086/315533.

Hussein, S. A., Osman, M. M., Hassan, M. M., Hassan, Y. S. A., Hussein, A. A., Adem, R., ... & Ali, A. A. (2025). Diphtheria outbreak in Somalia: a weekly sitrep on the recent health crisis-2025. *Tropical Medicine and Health*, *53*(1), 167.

Mc Cord—De Iaco, K. A., Gesualdo, F., Pandolfi, E., Croci, I., & Tozzi, A. E. (2023). Machine learning clinical decision support systems for surveillance: a case study on pertussis and RSV in children. *Frontiers in Pediatrics*, *11*, 1112074.

Melchane, S., Elmir, Y., Kacimi, F., & Boubchir, L. (2024). Artificial Intelligence for Infectious Disease Prediction and Prevention: A Comprehensive Review. *arXiv preprint arXiv:2411.10486*.

M.J, U., D.A, L., & E, O. (2019). Strengthening surveillance systems for outbreak detection in low-income countries: A scoping review. *Tropical Medicine and Health, 47, 39.* https://doi.org/https://doi.org/10.1186/s41182-019-0180-4

Musa, S., Usaini, S., Ahmed, I., Kiataramkul, C., & Tariboon, J. (2025). Dynamics of the Diphtheria Epidemic in Nigeria: Insights from the Kano State Outbreak Data. *Mathematics*, *13*(7), 1189.

Okhuarobo, U. J., Owoicho, S. A., Daikwo, J., Shehu, I. H., Omomoh, E., Ibara, M. O. P., ... & Saleh, F. (2025). Evaluation of diphtheria surveillance system in Kaduna State, Nigeria, July 2023–December 2023. *Journal of Public Health in Africa*, *16*(1), 1379.

Osarenren, J., Omosigho, P. O., & Okesanya, O. J. (2024). Global strategies for addressing diphtheria resurgence epidemiology clinical impact and prevention. *Discover Public Health*, *21*(1), 219.

Oteri, A., Ilesanmi, O. S., & Afolabi, A. A. (2020). Challenges of infectious disease surveillance in West Africa: a focus on diphtheria. *Pan African Medical Journal*, 36, 12.

Salim, S. M., & Hamza, F. M. (2024, January). Trends and geographical distribution of diphtheria in Nigeria: A re-emerging disease. In *Biology and Life Sciences Forum* (Vol. 31, No. 1, p. 22). MDPI.

Witter, S., Falisse, J. B., Bertone, M. P., Alonso-Garbayo, A., Martins, J. S., Salehi, A. S., ... & Martineau, T. (2015). State-building and human resources for health in fragile and conflict-affected states: exploring the linkages. *Human resources for health*, *13*, 1-15.

World Health Organization (WHO). (2023). *Diphtheria – Key facts*. https://www.who.int/news-room/fact-sheets/detail/diphtheria