Data Analytics for Vaccine Distribution

Chiemeka Nwakama, Eric Huang, Sam Konstan, and Nupur Kumar

Table of Contents

Introduction

By the end of 2023, more than 82 countries had successfully eradicated measles.[1] Vaccine research has produced a 93 percent-effective first vaccine dose, with that number increasing to 97 percent with an additional dose.[2] However, certain regions of the world continue to struggle to achieve herd immunity. In various parts of Africa, the Middle East, and Asia, measles remains a fatal risk, with a child mortality rate commonly ranging from 5 -10 percent and in some areas even as high as 20 percent.[3] As vaccines have grown more and more available, these nations have seen notable improvements—according to the World Health Organization's (WHO) in Africa, an estimated 4.5 million deaths were averted by vaccination from 2017 to 2021.[4] However, disparities in vaccine coverage persist, leaving many nations vulnerable.

This project focuses on West Africa, a region that exemplifies these disparities. By leveraging predictive models, we aim to identify trends in measles vaccination rates and determine which countries may require additional support. Our analysis uses historical data to explore the relationships between vaccination coverage and different factors to provide insights to guide future public health initiatives.

Methods

We began our analysis by aggregating datasets from two sources: the WHO Immunization Data for the Measles Vaccine[5] and UNICEF's Vaccine Coverage Survey Data[6]. The WHO dataset provided annual measles case counts from 1974 to 2023 across 213 countries. Unfortunately, the data was very sparse and contained high proportions of missing values, limiting its usefulness for analysis. The UNICEF dataset included MCV1 and MCV2 vaccine coverage rates from 2000 to 2023 for 189 countries. We only used the countries that showed up in both datasets and restricted the years to 2000-2023 to ensure consistency and relevance. We prioritized countries by calculating the mean vaccination coverage rate across all available years and identified the top five and bottom five performers. However, significant data gaps in several countries made them unsuitable. Across nations with adequate data, the top five were the United States, United Kingdom, Argentina, Germany, and Finland, while the Democratic Republic of the Congo, China, Nigeria, India, and Madagascar sat at the bottom.

As we advanced to modeling, we encountered limitations in the dataset's features. There were significant differences in economic, political, geographical, and healthcare factors among all these countries that made comparison and prediction difficult. To address the challenges of analyzing global data, we narrowed our focus to the West African region. This shift allowed us to incorporate new features such as healthcare access, infrastructure, and economic stability. We selected five countries based on data availability: Cameroon, Ghana, Liberia, Sierra Leone, and Nigeria. Using data from the World Bank Open Data[7] platform, we assembled a dataset by selecting variables we believed influenced vaccine distribution. Our dataset covered the years 2000 to 2022.

When the data set was exported, we encountered formatting challenges that made it difficult to analyze. Specifically, each variable was represented as a row and each year as a column, which complicated the data processing. To resolve this, we restructured the dataset, creating separate pivot tables for each country. In

[1] CDC, "Measles Cases Surge Worldwide."

[2] "Measles Vaccination Recommendations."

[3] Moss, "Measles Still Has a Devastating Impact."

[4] Masresha, "Progress Toward Measles Elimination."

[5] WHO, "Measles - Number of Reported Cases."

[6] UNICEF, *Immunization*

[7] World Bank Data, "World Development Indicators."

the revised format, each variable became a row and each year a column, significantly improving the readability of the data. This restructuring not only helped speed up the analysis process, but gave us a better understanding of the importance of data cleaning.

Our primary objective was to predict vaccination rates with “Immunization for measles (% of children ages 12-23 months)” as the response variable. The predictor variables included:

- Birth rate, crude (per 1,000 people),
- Current health expenditure (% of GDP)
- Death rate, crude (per 1,000 people)
- Mortality rate under-5 (per 1,000 live births)
- Number of under-five deaths
- Out-of-pocket expenditure (% of current health expenditure)
- People using at least basic drinking water services (% of population)
- Population, total
- GDP growth (annual %)
- Inflation, consumer prices (annual %)

We set aside 20 percent of data for testing, and we used two predictive models: regression trees and random forests. After evaluating the models using Mean Squared Error (MSE) and R-squared, we examined whether the predictors significantly influenced vaccine distribution. In an attempt to refine our analysis, we assessed the importance of each predictor variable and pruned off the insignificant ones (as detailed in Tables 3 and 5 in the Appendix). We re-ran the models post-pruning in an attempt to improve the MSE and R-squared values.

When choosing the best regression model for predicting measles vaccine coverage, we used polynomial regression instead of a linear model. We saw significant variability in vaccine coverage, with nothing to indicate that it may be a linear function of our predictors. By using a high-degree polynomial function for our model, we were able to account for the data’s complexities and make more accurate predictions.

In each of our models, the variable x represents the number of years since the last observed measles coverage data, with 2022 as the baseline or starting year. For x=0, the model outputs the coverage percentage observed in 2022. In polynomial regression, this comes out to B_0, the y-intercept. Setting x=1 would correspond to the model's prediction for 2023. Each country has its own polynomial regression model that is tailored to its unique circumstances. While we used the same set of predictors across all countries, the relationships between these predictors and coverage may vary, resulting in a distinct model for each nation. This approach ensures that each country's unique characteristics are reflected in its model. It acknowledges that vaccination coverage trends are shaped by factors specific to each country, highlighting the limitations that using a single global model would pose. Using our models, we predicted measles vaccine coverage for the five years following 2022. We then combined our findings to compare and analyze trends and differences in coverage among the countries.

Results

In our first round of testing, we compared the accuracy of a random forest model and a regression tree model. This test found that the random forest models yielded overall poor results (Table 1). Cameroon was the one exception, yielding an R-squared value larger than 0.7. Our regression trees produced mixed results. The R-squared values for Cameroon, Nigeria, and Sierra Leone all improved by a large margin, whereas those for Ghana and Liberia grew more negative (Table 2). Upon inspection, Ghana’s and Liberia’s coverage rates consistently fluctuated even with little change in the predictors. We decided to move on with the more accurate model despite these anomalies.

After choosing to pursue a regression tree, we began to prune our models. Because each model had slightly different variable importance, we added them up between all countries. The result of this showed "Deaths Under 5" to be the most significant by a large margin, followed by the other 2 death-related predictors and "Population". Furthermore, a few direct healthcare-related predictors had underperformed such as "Health Expenditure" and "GDP Growth" (Table 3). We decided to cut off the 2 lesser death-related predictors due to a high correlation with "Deaths Under 5" as well as to prune off the bottom 3 predictors (Table 4). Our pruned model contained the following variables: Deaths Under Age 5, Population, Birth Rates, Basic Drinking Water, and Out-of-Pocket Expenditure.

The newly pruned list of predictors was then used to reform the previous 2 models. Pruning the random forest model exacerbated the problem we'd run into earlier, marginally improving R-squared values for Cameroon, Nigeria, and Sierra Leone while reducing already-poor performance for Ghana and Liberia (Table 5). Overall, pruning the tree did not make a significant and beneficial difference. Our regression trees saw a much larger difference after pruning. Liberia's models fared notably better—its MSE dropped by nearly 45 percent—while the remaining four countries performed much worse (Table 6). Among these results, we found that the hierarchy of predictors had changed. The importance of "Basic Drinking Water" had increased dramatically, now becoming the highest in importance. The rest of the variables saw a condensed importance range, though still maintaining the same order.

Lastly, we used polynomial regression with all variables to predict immunization rates for each country from 2023 to 2027. From this, it was found that all countries excluding Liberia are expected to have increased immunization rates in the near future (Figure 1). However, for Liberia and Ghana, these results should be further scrutinized due to the previous R-squared values, suggesting that the predictors used do not fit their response (Table 2). The overall projected immunization rates for all 5 countries were found to increase linearly throughout the next several years (Figure 2). Despite the negative R-squared values of Ghana and Liberia, the region of West Africa still faces quite a few similarities, so the other countries' results may still be relevant.

Discussion

Because our models for certain countries, such asGhana and Liberia, had significantly lower accuracies, we were not able to identify an overarching implementable takeaway. Using these results to decide which countries need the most aid would require higher model accuracy across countries. The negative R-squared values for Ghana and Liberia indicate that the factors currently included in our models provide limitations for vaccine distribution in these countries, highlighting the need to explore additional variables for vaccine coverage. For instance, Liberia's 2022 measles outbreak exposed critical gaps in immunization services, with vaccination rates falling below 95% herd immunity threshold.[8] Disparities in coverage across regions, inadequate healthcare, and pandemic disruption could have further brought barriers to achieving high vaccination rates. While we cannot present solidified conclusions, we have identified two key considerations to improve future models.

First, we treated each year as an independent datapoint. As a result of our limited collective experience using time series data, we missed certain relationships between years. There are three ways to make up for this gap. First, future trials could introduce more year-to-year relationships as additional variables. For example, one datapoint could be the year 2014 with variables for 2014 vaccination rates, 1-year-prior vaccination rates, and 2-years prior vaccination rates. This technique would help simpler models to embed smaller trends as additional predictors. Second, there are models designed specifically for managing time-series data that would be a better fit for this type of project. Third, using a more complex model such as a

[8] Shibayo et al., "Descriptive Analysis of Measles Outbreak in Liberia."

deep neural network would do a better job of capturing relationships between neighboring years. With such limited data for a specific country, though, overfitting would be a considerable issue.

Our second consideration is finding ways to add in more data. Ideally, we would have a finer partition of datapoints (6-12 datapoints per year instead of just one). Even without that, there are many ways in which additional data could be incorporated. The easiest would be to add more demographic information—geography, age distribution, wealth distribution, etc as we expect that these would all have some bearing on vaccination rates. Alternatively, we could attempt to work in more complicated and incomplete datatypes such as political and economic events, international alliances and aid organizations, and more. In that case, the struggles would be a) finding the proper datasets, b) assigning a numerical system to non-quantitative events, and c) choosing only among models that can handle missing data.

Beyond the prediction task, one future area of research is to identify risk factors for countries with limited data. By creating a model with several predictors such as the ones mentioned above, we could analyze variable loadings and determine what values indicate low vaccination rates. That would allow us to predict that a country may need help with vaccinations even if we don't have that nation's actual vaccination data. Suggested factors to investigate include but could not be found/were unavailable include:

- Accessibility of healthcare services in rural and urban areas
- Maternal education levels
- Socioeconomic status, specifically affecting healthcare
- Nutritional status of children
- Number of hospital beds (per 1,000 people)
- Number of physicians (per 1,000 people)

Conclusion

This project explored the predictors of measles vaccine coverage in Cameroon, Ghana, Liberia, Sierra Leone, and Nigeria, using each year from 2000 to 2022 as a datapoint. In order to identify what countries require aid for measles vaccination, we first ran a regression tree and random tree to try to figure out which predictors were significant. Running both models, we attempted to remove insignificant predictors based off predictor importance as well as R-squared values; however, when rerunning both models with some predictors that looked insignificant, we did not see enough of an improvement trend amongst all countries to justify keeping these predictors dropped, so we decided to keep all the predictors we originally had. While our models showed some promise, with Cameroon achieving an R-squared value of 0.9 in a regression tree model, the overall accuracy across countries was inconsistent, Ghana and Liberia obtained negative R-squared values indicating that the predictors used were insufficient to describe vaccine distribution in these countries. Once we finalized the predictors, we trained the data for each country individually, creating five unique polynomial regression models, one for each country. Using these models allowed us to predict measles vaccine immunization coverage for the next five years, from 2022 to 2027, by incrementing the x-value to represent the desired number of years past 2022.

This project highlights the importance of improving data quality and adding variables for better vaccine coverage modeling. Using time-series models or deep learning algorithms could capture year-to-year relationships more effectively. Future research could focus on adding predictors for low vaccination coverage in areas where data is limited and health boost public health strategies. While our models provide a good foundation for understanding measles immunization trends in West Africa, there is still a need for better data in order to better understand how to boost vaccine coverage and ensure fair access for all.

Contributions

Team Member	Contribution
Eric Huang	Code for Modeling, Results.
Sam Konstan	Introduction, Discussion, Conclusion.
Nupur Kumar	Data Collection + Cleaning, Methods.
Chiemeka Nwakama	Code for Modeling, Methods.

Works Cited

CDC. "Measles Cases Surge Worldwide, Infecting 10.3 Million People in 2023." CDC Newsroom, November 19, 2024. https://www.cdc.gov/media/releases/2024/p1114-measles-cases.html.

Immunization. UNICEF, July 2024. https://data.unicef.org/topic/child-health/immunization/.

Masresha, Balcha G. "Progress Toward Measles Elimination — African Region, 2017–2021." *MMWR. Morbidity and Mortality Weekly Report* 72 (2023). https://doi.org/10.15585/mmwr.mm7236a3.

"Measles - Number of Reported Cases." Accessed December 16, 2024. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/measles---number-of-reported-cases.

"Measles Vaccination Recommendations - MN Dept. of Health." Accessed December 16, 2024. https://www.health.state.mn.us/diseases/measles/hcp/vaxrecs.html#:~:text=MMR%20vaccine%20is%20about%2093,if%20exposed%20to%20the%20virus.

Moss, William J. "Measles Still Has a Devastating Impact in Unvaccinated Populations." *PLoS Medicine* 4, no. 1 (January 2007): e24. https://doi.org/10.1371/journal.pmed.0040024.

Shobayo, Bode, Chukwuma David Umeokonkwo, Ralph Weah Jetoh, Julius S. M. Gilayeneh, Godwin Akpan, Maame Amo-Addae, Jane Macauley, and Rachel T. Idowu. "Descriptive Analysis of Measles Outbreak in Liberia, 2022." *IJID Regions* 10 (March 1, 2024): 200–206. https://doi.org/10.1016/j.ijregi.2024.01.008.

"World Development Indicators | DataBank." Accessed December 16, 2024. https://databank.worldbank.org/source/world-development-indicators#.

Appendix

Table 1. Random Forest Results (All Predictors).

Countries	**MSE**	**R^2**
Cameroon	32.7	0.72
Ghana	33.30	-2.38
Liberia	99.09	-0.07
Nigeria	50.98	0.55
Sierra Leone	157.87	0.42

Table 2. Regression Tree Results (All Predictors).

Countries	**MSE**	**R^2**
Cameroon	12.25	0.9
Ghana	36.81	-3.76
Liberia	149.30	-1.57
Nigeria	21.57	0.81
Sierra Leone	78.28	0.71

Table 3. Predictor Importance (All Predictors).

Predictor	**Importance**
Deaths (Under Age 5)	0.919
Death Rate	0.654
Population	0.632
Mortality Rate (Under Age 5)	0.615
Birth Rates	0.550
Basic Drinking Water	0.502
OOP Expenditure	0.472
GDP Growth	0.249

Health Expenditure	0.222
Inflation	0.185

Table 4. Predictor Importance (Remaining Predictors).

Predictor	**Importance**
Deaths (Under Age 5)	1.079
Population	1.066
Birth Rates	0.936
Basic Drinking Water	1.148
OOP Expenditure	0.769

Table 5. Random Forest Results (Pruned Predictors).

Countries	**MSE**	**R^2**
Cameroon	28.48	0.76
Ghana	34.98	-2.55
Liberia	115.35	-0.98
Nigeria	56.30	0.51
Sierra Leone	131.97	0.52

Table 6. Regression Tree Results (Pruned Predictors).

Countries	**MSE**	**R^2**
Cameroon	35.00	0.70
Ghana	46.81	-3.76
Liberia	83.34	-0.43
Nigeria	29.57	0.74
Sierra Leone	80.52	0.70

Table 7. Predicted Rates.

Countries	2023	2024	2025	2026	2027
Cameroon	59.15	58.72	58.82	59.5	61.00
Ghana	65.88	98.07	99.75	99.75	99.75
Liberia	65.86	64.86	63.78	62.64	61.46
Nigeria	67.35	71.97	77.31	83.41	90.35
Sierra Leone	89.74	90.04	90.31	90.57	90.83

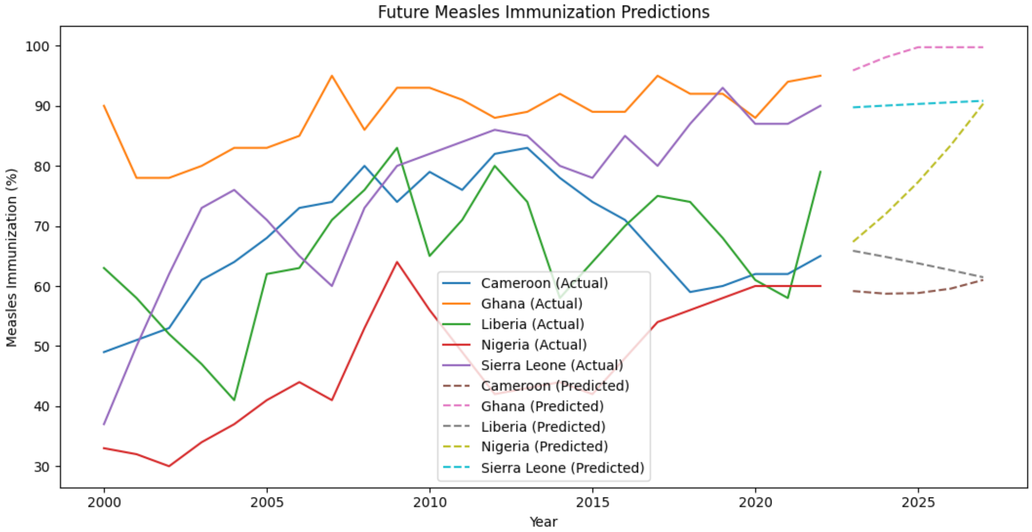

Figure 1. Future Measles Immunization Predictions.

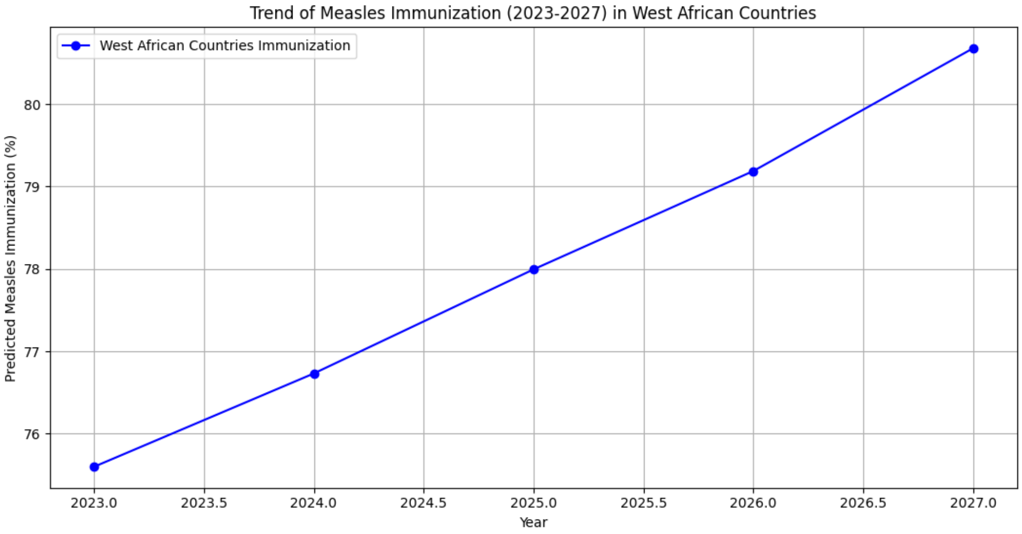

Figure 2. Predicted Trend of Measles Immunization For All Five West African Countries.