Discover Public Health

RESEARCH Open Access

Associations between health system indicators and key public health outcomes using machine learning on WHO country level data

Pakiza Valizada[1*] and Ahmet Elbir[2]

*Correspondence:
Pakiza Valizada
pakiza.valizada@gmail.com; Pakiza_Valizada@cbar.az
[1]Central Bank of Azerbaijan, 167 Heydar Aliyev Avenue, Nizami District, AZ1029 Baku, Azerbaijan
[2]Yıldız Technical University, Davutpaşa Campus, Esenler, 34220 Istanbul, Türkiye

Abstract

This study examines how data analytics and machine learning (ML) techniques can be applied to country-level health indicators to generate descriptive and predictive insights into public health systems. While existing literature highlights the role of large-scale digital data in improving disease surveillance, policy design, and predictive health analytics, many low- and middle-income countries continue to rely on more limited datasets. This raises important questions regarding the extent to which structured, aggregated data can support similar analytical objectives. Using publicly available data from the World Health Organization (WHO) DataDot platform, this study explores relationships between key health system indicators, such as the "Universal Health Coverage (UHC) Service Coverage Index", healthcare workforce density, and "Domestic general government health expenditure (GGHE-D) as percentage of general government expenditure (GGE)", and major public health outcomes, including immunization coverage, noncommunicable disease (NCD) mortality, and life expectancy. A correlation analysis and ML models (Random Forest, XGBoost, and LightGBM) were used to detect patterns and evaluate predictive relationships. The findings identify descriptive associations and predictive patterns between selected health-system indicators and public health outcomes. Moreover, feature importance analysis was also conducted to emphasize the relative predictive contribution of interaction terms and multiple health system indicators. A conceptual framework is also proposed to illustrate how data analytics can inform public health system monitoring and decision-making. The framework is presented as a theoretical model to guide future empirical research. Overall, findings of the study are considered associational rather than causal.

Keywords Public health, Data analytics, Machine learning, Immunization coverage, WHO health indicators, Health system capacity

1 Introduction

Although "big data" has been an omnipresent term, there is no universally accepted definition for it. Gartner describes "big data" as information assets characterized by their high volume, high velocity, and high variety, which require novel processing techniques

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

to provide support for enhancing insights, decision-making, and automation processes [16]. In the healthcare context, "big data" includes a vast amount of data derived from various sources, such as medical imaging, EHRs, genomic sequencing and geo-mapping, IoT sensors, and wearable healthcare devices which are used for tracking personal health data. These data sources have been used to support predictive analytics, disease monitoring, and health system evaluation.

According to Adenyi, Okolo, Olorunsogo, & Babawarun, "big data" used in healthcare improves tracking of disease occurrences in the population, enables the use of various predictive analytics techniques and ML models by identifying patterns in historical data in order to anticipate the needs of patients and potential risks in their health status [1]. Although these approaches are often applied to large-scale data environments, many health systems, particularly in low- and middle-income countries, still operate with more limited datasets. This raises crucial questions about how structured and aggregated data can be analyzed to derive useful insights using ML approaches and data analytics.

By considering the pivotal role of data analytics in the public health sector, the following questions were identified in this study:

- To what extent can key health system indicators predict immunization coverage as an indicator of system responsiveness?
- How do system-level health inputs relate to health outcomes, especially "NCD mortality"?
- How do measures of health system capacity relate to immunization coverage across countries?
- How effective are ML techniques at identifying connections among the country-level health indicators related to public health outcomes?
- To what extent can the public health indicators be used in predicting life expectancy?

Based on these questions, the study proposed the following four analytical objectives:

1. To evaluate whether higher levels of major health system factors (e.g. "UHC Service Coverage Index", government health expenditure, and their interactions) are associated with and contribute to the prediction of improved immunization coverage, which can be considered a tracer indicator of health system performance and service delivery capacity.
2. To examine whether there are statistical associations between health system inputs (e.g. the "UHC Service Coverage Index", government health expenditure, healthcare workforce density) and health outcomes (e.g. NCD mortality rates). Such associations can be useful for developing data-driven policies.
3. To assess whether health system capacity metrics (such as the "UHC Service Coverage Index", healthcare workforce density, government health expenditure, and IHR core capacity score) are positively associated with immunization coverage.
4. To evaluate whether public health variables (e.g. immunization coverage, "NCD mortality rates", healthcare workforce density, government health expenditure, and tobacco use rate as one of the health-risk factor) contribute to the prediction of life expectancy.

Based on these analytical objectives, the empirical analysis was organized into four analytical components, and their corresponding analyses are presented in the "Data and

Methodology" section. In addition, the study also proposed a conceptual model for applying "data analytics" to public health, consisting of seven phases.

In this study, data were retrieved from the WHO's DataDot platform, which was created to serve as a centralized and user-friendly hub for accessing a wide range of global health datasets. The platform was initially designed to solve the fragmented global health reporting through a central health data repository. Today, the platform is mainly used for the purpose of collecting harmonized health indicators from member states and WHO monitoring systems, helping academic researchers and policymakers working in public health institutions to effectively track health system performance and international trends [24].

1.1 Literature review

Advances in data analytics and ML have significantly expanded the ability to analyze health-related data and identify patterns in public health systems. Modern analytical approaches are capable of identifying patterns and seasonality from historical datasets, supporting applications such as disease monitoring and predictive modeling.

Traditional disease surveillance systems, which primarily depend on manual data collection and reporting, are limited by time lags and missing information, especially in fast-moving public health emergencies [20]. Unlike conventional disease surveillance systems, recent progress in digital health technologies has resulted in more flexible and real-time surveillance systems. For example, digital tools such as Google Trends and the WHO's "Epidemic Intelligence from Open Sources" (EIOS) system have advanced this field by monitoring population-level health patterns [18]. Google Trends allows researchers to track search trends for public health-related terms for identifying patterns, predicting disease outbreaks, and monitoring public health trends [6–15]. Hong and Chen [20, 8] state that ML models play a crucial role in analyzing the large volumes of data collected from real-time health surveillance systems, social media, and electronic health records. These models improve real-time monitoring of viral infections and predict their progression. However, Telles, Lopes, and Franco claim that linear models may fail to observe certain aspects of viral diseases [19] (Fig. 1).

Similarly, traditional statistical models like the "Moving Average" (MA) and "Auto-Regressive Moving Average" (ARIMA) models assume that the relationship between the current, past and future values is linear and constant over time. However, epidemiological dynamics are considered nonlinear and non-constant since they depend on sudden external changes as a result of interventions, such as lockdowns, vaccinations, and viral mutations. Furthermore, the "ARIMA" models don't take into account exogenous variables (e.g., population mobility, vaccination rates), and in these types of models, it is assumed that the data is stationary, meaning its mean, variance, and autocorrelation stay constant over time [14–20]. Nevertheless, frequent structural outbreaks can change the case numbers of viral infections.

In order to overcome these constraints, Tanwar et al. [21] proposed a cloud-oriented AI-based framework, "D-espy" (i.e., "Disease-espy") for "COVID-19" disease detection and prevention. By using the "Stacked LSTM" model, it performs better than traditional models like "ARIMA" and incorporates a resource distribution mechanism to support the optimal allocation of critical health resources across various regions based on

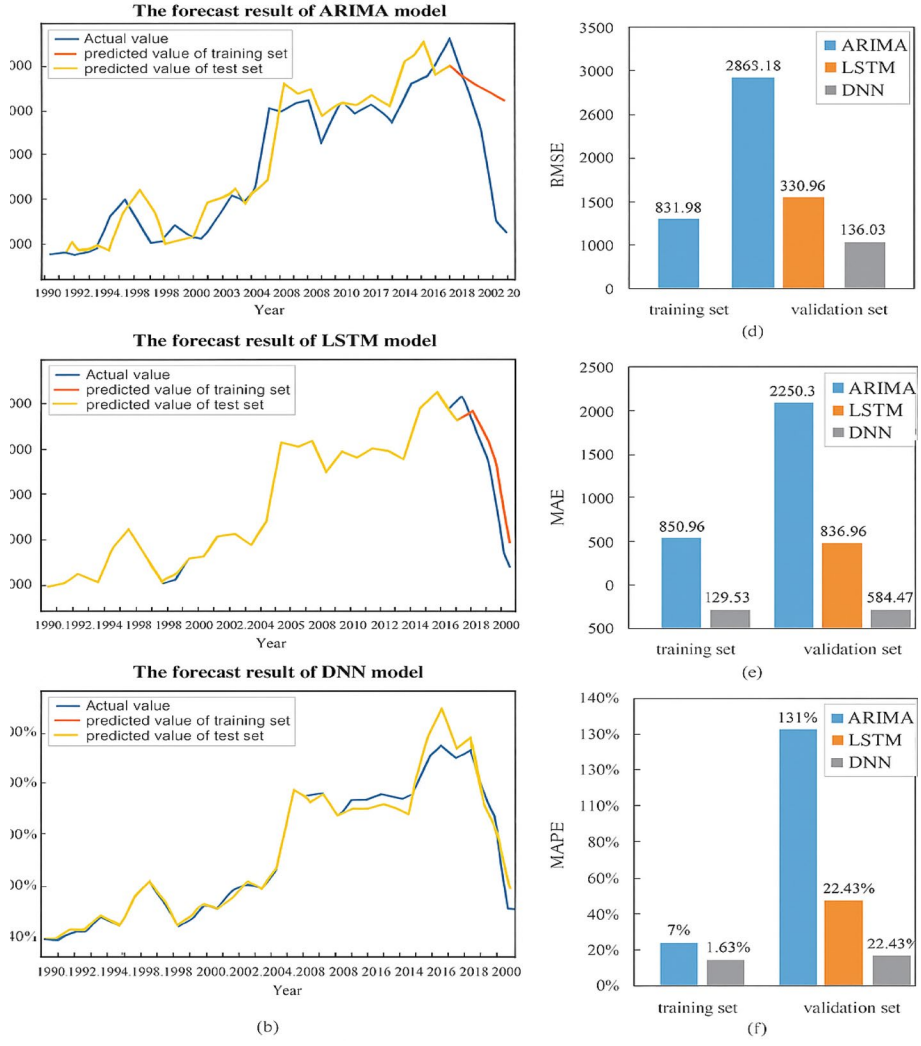

Fig. 1 **a** Comparison results of **a** ARIMA (0,2,1), **b** multivariate LSTM, and **c** DNN models; **d** RMSE, **e** MAE, and **f** MAPE performance metrics across the three models.
Source: Adapted from Lou HR, Wang X, Gao Y, et al. [25]

predicted disease growth. By supporting evidence-based decision-making, the "D-espy" enhances the effectiveness of public health strategies during viral infection outbreaks.

Recent research not only highlights the growing role of ML in public health for improving predictive analytics and supporting data-driven decision-making, but also addresses ethical concerns such as data privacy and bias. These insights underscore the value of applying ML techniques to analyze complex datasets, including country-level WHO indicators as examined in this study [26].

While these studies are mainly concerned with large scale or real time health data, there is increasing interest in using the same analytical techniques on structured, country-level datasets. This study contributes to this area by examining how data analytics and ML techniques can be used to analyze publicly available WHO health indicators.

Based on the existing literature and analytical approaches above, this study proposes a conceptual model to illustrate how data analytics can be systematically applied within public health systems.

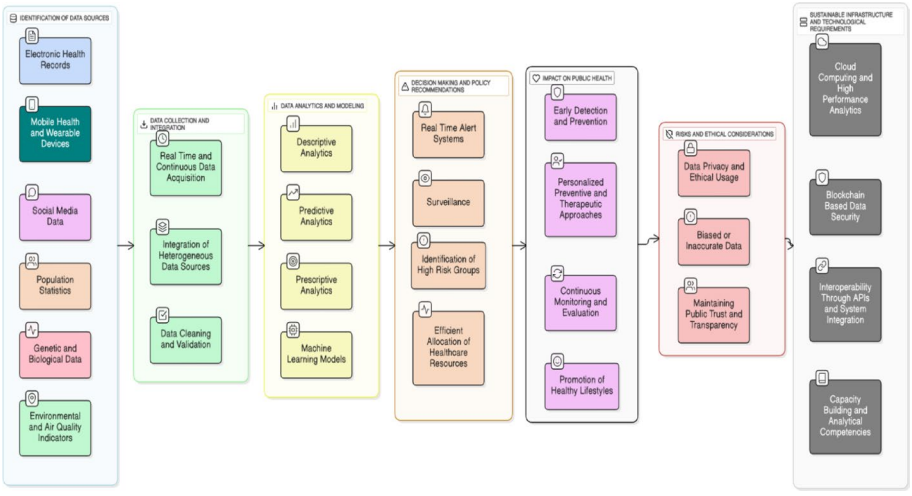

Fig. 2 Conceptual model for the application of "data analytics" in Public Health.
Source: Flowchart generated by the author using Eraser (https://app.eraser.io/)

Table 1 Initial phase: identification of data sources.

Electronic Health Records (EHR) [5]
"Mobile health" (mHealth) applications and "wearable devices" data [20].
Social media data.
Population statistics (census and surveys)
Genetic and biological data.
Environmental and air quality indicators (geospatial data)

Source: Compiled by the author based on various sources

Figure 2 serves as a conceptual illustration of the major components involved in the application of data-driven strategies within public health systems. The model emphasizes interconnected elements which are frequently highlighted in the existing literature. The conceptual model consists of seven main phases, which provides a thorough overview of how structured "health data" can be used to support public health outcomes by demonstrating these seven interrelated components. It is important to note that this conceptual model is derived from a synthesis of prior research and serves as a theoretical framework to provide direction for future research in data-driven public health (Table 1).

1.2 Conceptual model for the application of "data analytics" in public health

The literature reviews of the following articles were used to determine the stages for building a conceptual model for the application of "data analytics" in public health:

1.2.1 Data collection and integration

The next stage after identifying the data sources is to collect, integrate, clean and organize the data for modeling (Table 2).

1.2.2 Data analytics and modeling

Upon data collection and integration, the next phase involves applying various analytical methods to extract actionable insights and generate ML models (Fig. 3; Table 3).

Table 2 Data collection and integration.

Real-time and continuous data acquisition
Integration of heterogeneous data sources
Data cleaning and validation to ensure reliability

Source: Compiled by the author based on various sources

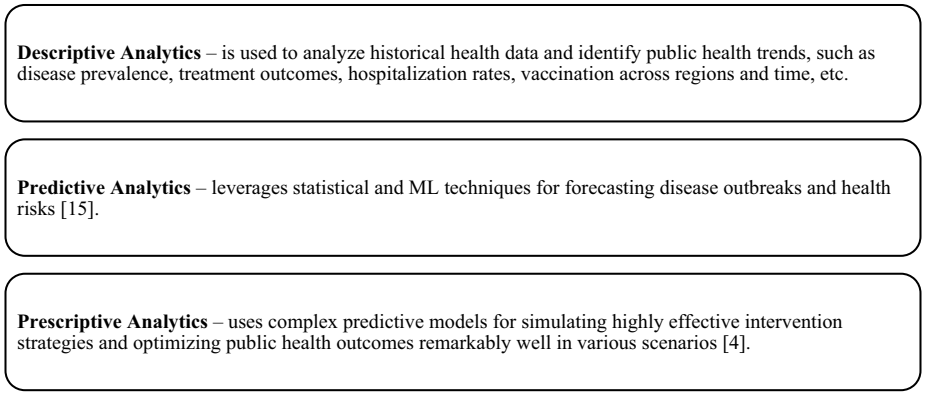

Fig. 3 Overview of descriptive, predictive, and prescriptive analytics in public health.
Source: Created by the author

Table 3 Machine learning models.

Decision-Making and Policy Recommendations [9]
Real-time alert systems (early warning)
Digital epidemiological surveillance
Identification of high-risk groups [2]
More efficient allocation of healthcare resources
Impact on Public Health
Early detection and prevention of disease spread [3]
Personalized preventive and therapeutic approaches [17]
Continuous monitoring and evaluation of health programs
Promotion of healthy lifestyles and behavior change
Risks and Ethical Considerations
Data privacy and ethical usage of sensitive health information [8]
Biased or inaccurate data leading to flawed decisions
Maintaining public trust and transparency
Sustainable Infrastructure and Technological Requirements
Cloud computing and high-performance analytics
Blockchain-based data security
Interoperability through APIs and system integration [12]
Capacity building and development of analytical competencies

Source: Compiled by the author based on various sources

2 Methodology and data

This study examines the relationships between key health system indicators and public health outcomes using publicly available country-level data. Instead of determining the direct impact of "big data technologies", the analysis focuses on identifying any associations and predictive patterns that can be helpful for making data-driven decisions in public health.

The dataset employed in this study was retrieved from the WHO's DataDot platform, which provides country-level health indicators. The data analysis in the study is meant to

be illustrative and conceptual, which highlights how patterns and relationships detected within smaller datasets can inform evidence-based decision-making that large-scale data systems seek to optimize. In this context, the analysis serves as an illustrative application of how ML techniques can be used to extract meaningful insights from aggregated datasets.

All data processing and analysis were conducted using the R statistical programming environment. Rows with excessive null values were excluded to minimize artificial bias and mitigate noise while preserving the reliability of statistical inference. Remaining missing values in numeric variables were imputed using the country-specific median [22, 27]. The methodological approach combines conventional predictive ML models (including Random Forest, XGBoost, and LightGBM) with correlation analysis to explore relationships between system-level inputs, such as the "UHC Service Coverage Index", government health expenditure, and healthcare workforce density, and key health outcomes, including immunization coverage, NCD mortality, and life expectancy. Different methods are employed to address the analytical objectives, since this study combines descriptive correlation analysis with ML models that are focused on predictive performance. For predictive modeling components, evaluation is based on model accuracy and predictive contribution. For correlation analyses, correlation coefficients were calculated using pooled country-year observations and are reported only as descriptive point estimates. P-values are not reported because repeated observations within countries may inflate statistical significance.

The dataset has a country-year panel structure. For that reason, observations from the same country across years were not treated as fully independent. For predictive modelling, the train/test split and cross-validation folds were constructed at the country level to ensure that the same country didn't appear in both training and testing sets. In addition, likely structural zeros in measles-containing-vaccine second-dose (MCV2) and third dose of pneumococcal conjugate vaccine (PCV3) were treated as missing before reconstructing the immunization coverage outcome, since these values may reflect non-introduction of the vaccine rather than true 0% coverage. Missing numeric predictors were imputed using training-only country-specific medians, with training-set global medians used as a fallback where country-specific medians were unavailable [22, 27].

After outcome exclusions and predictor imputation, Analytical Objectives 1 and 2 used the full balanced sample of 4,026 country-year observations from 183 countries covering 2000–2021. Analytical Objectives 3 and 4 were restricted to 2010–2021 because the IHR Core Capacity Score was available only from 2010 onward, resulting in 2,196 country-year observations from 183 countries in each analysis.

It is important to note that the findings should be interpreted as associational rather than causal, as the study is based on observational data and doesn't incorporate experimental or quasi-experimental design.

Despite the global authority and broad scope of the WHO DataDot platform, there are some limitations that should be taken into account in analyzing the health data, as outlined in Table 4 below:

For clarity and consistency, the technical variable names used in the regression models, figures, and feature-importance outputs correspond to the following public health indicators. "UHC_INDEX" refers to the "UHC Service Coverage Index", "GOV_EXP_ON_HEALTH_PER_100_N" refers to "GGHE-D" as a percentage of "GGE", and

Table 4 Constraints related to analyzing health data in this study.

Constraint domain	Description
Data Standardization	Variations in diagnostic criteria of patients and reporting standards across countries can result in measurement bias or comparability issues.
Ethical Constraints	Micro-level analysis is hampered by anonymization and restricted access to datasets, which also restricts the use of these resources by independent researchers.
Political Bias	The accuracy of health reports may be hampered by institutional constraints or political interests in the member states.
Analytical Limitations	The statistical analysis in this study was aimed at identifying relationships between selected variables. However, there can be omitted variables which can affect the robustness and validity of the regression models.

Source: Created by the author

"HealthWorkers_Density" refers to healthcare workforce density, calculated as the combined density of physicians, nurses, and pharmacists per 10,000 population. The dependent variable "immunization_coverage" or "immunization_coverage_rate" refers to the average coverage of three vaccines, DTP3, MCV2 and PCV3, which are used in this study as tracer indicators of immunization system performance and service delivery capacity. "NCD Death Rate" or "NCD_RATE_PER_100_N" refers to the age-standardized mortality rate from major noncommunicable diseases; and "IHR Core Capacity Score" or "IHR_capacity_score" refers to the average score across core capacities for detecting, reporting and responding to public health threats under the International Health Regulations. In the life expectancy model, "TOBACCO_USE_RATE" refers to the percentage of population using tobacco products, and "DTP3_RATE_PER_100_N", "MCV2_RATE_PER_100_N" and "PCV3_RATE_PER_100_N" refer to the corresponding vaccine coverage rates. "DENSITY_PHY_RATE_PER_10000_N", "DENSITY_NURSE_RATE_PER_10000_N" and "DENSITY_PHARM_RATE_PER_10000_N" refer to the density of physicians, nurses and pharmacists per 10,000 population. The interaction terms "UHC_x_GOV", "UHC_x_HW", "GOV_x_HW", and "UHC_GOV_HW" represent the combined effects between the main predictor variables used in the models.

2.1 Analytical objective 1: prediction of immunization coverage based on health system factors and interaction terms

This analysis evaluates whether major health system indicators, including the "UHC Service Coverage Index", "GGHE-D as a percentage of GGE", healthcare workforce density, and their interaction terms, contribute to the prediction of immunization coverage as a tracer indicator of health system performance and service delivery capacity.

"LightGBM", "Random Forest", and "XGBoost" regression models were utilized to predict immunization coverage using a combination of healthcare system indicators and interaction features. The models were tuned via cross-validation, and the final models were trained with optimized hyperparameters to enhance prediction accuracy.

In the regression models, the "immunization_coverage" variable was used as a dependent variable and was measured as the average of three key vaccines-DTP3, MCV2, and PCV3 - which are recognized as tracer indicators of public health system performance and service delivery capacity (Tables 5 and 6) [28].

We can see the definitions of the predictor variables and interaction terms used in the models in the tables below:

Based on the results of the Fig. 4 and the skewness test distribution, the predicted "immunization coverage rate" variable yields a skewness coefficient of -0.503, which

Table 5 Predictor variables used in the "LightGBM", "Random Forest", and "XGBoost" models.

Predictor variables	Description
UHC_INDEX	"UHC Service Coverage" Index;
GOV_EXP_ON_HEALTH_PER_100_N	"GGHE-D" as a percentage of "GGE"
HealthWorkers_Density	Total density rate of healthcare system workers; computed by summing the density rates of nurses, physicians, and pharmacists per 10,000 population
-DENSITY_NURSE_RATE_PER_10000_N	Number of nurses per 10,000 population.
-DENSITY_PHY_RATE_PER_10000_N	Number of physicians per 10,000 population.
-DENSITY_PHARM_RATE_PER_10000_N	Number of pharmacists per 10,000 population.

Source: Created by the author

Table 6 Interaction terms used in the "LightGBM", "Random Forest", and "XGBoost" models.

Interaction term	Description
UHC_x_GOV	Captures the combined effect of UHC and "GGHE-D" as percentage of "GGE".
UHC_x_HW	Explores the interaction between UHC and availability of healthcare workforce density.
GOV_x_HW	Evaluates the joint effect of "GGHE-D" as percentage of "GGE" and healthcare workforce capacity.
UHC_GOV_HW	Three-way interaction to evaluate the amplified effect when all three predictors are considered together.

Source: Created by the author

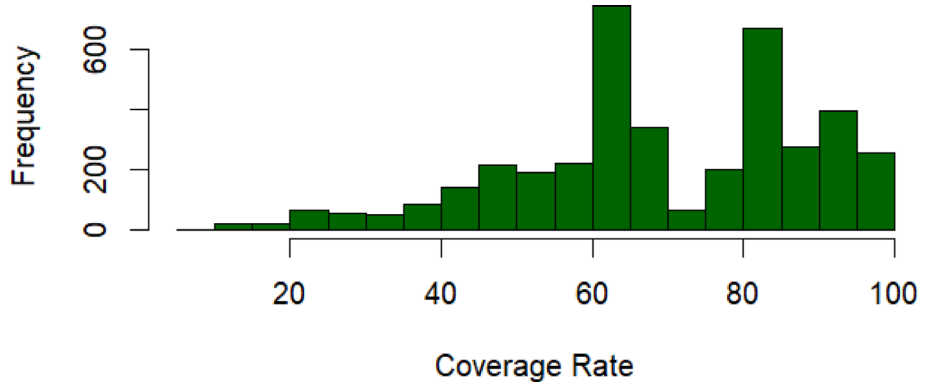

Fig. 4 Histogram of immunization coverage rate.
Source: Created by the author using R (ggplot2 package)

indicates a moderately left-skewed distribution, though the degree of skewness is not extreme [22].

2.2 Model evaluation and performance metrics

The performance of the three models was evaluated on a held-out test set using a country-level, leakage-free validation strategy. In order to address the panel structure of the dataset and avoid leakage from country-year observations, the train/test split and 5-fold cross-validation were constructed using country-level grouping rather than purely random row-wise splitting. This ensured that observations from the same country didn't appear in both sets. In addition, 5-fold cross-validation was grouped by country, so that all years from the same country remained in the same validation fold. This approach provides a more conservative assessment of predictive performance than a row-wise random split and directly addresses the panel structure of the data. Under this validation

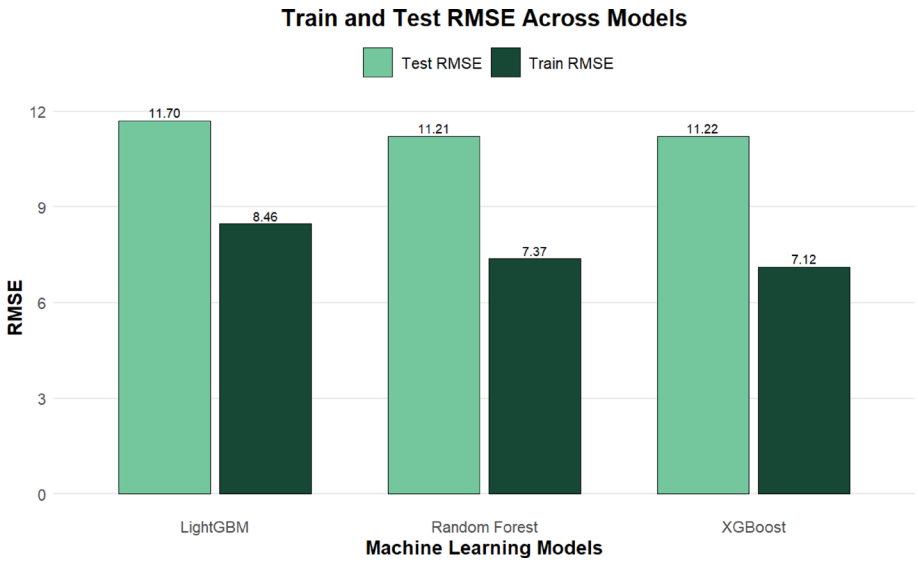

Fig. 5 Train and test RMSEs for the "Random Forest", "XGBoost", and "LightGBM" Models.
Source: Created by the author using R (ggplot2 package)

Table 7 Results of the ML models.

Model	CV RMSE mean	CV RMSE SD	Train RMSE	Test RMSE	Train R^2	Test R^2	Train MAE	Test MAE	Over-fitting gap
Random Forest (Ranger)	9.86	3.29	7.37	11.21	0.68	0.50	5.34	8.54	3.83
XGBoost	10.24	3.00	7.12	11.22	0.70	0.49	5.36	8.59	4.09
LightGBM	10.01	3.23	8.46	11.70	0.60	0.47	6.49	9.13	3.23

Source: Created by the author using R

approach, three ML models were compared using "Root Mean Squared Error" (RMSE), "Mean Absolute Error" (MAE) and "Coefficient of Determination" (R^2) performance metrics. The "Random Forest (Ranger)" model showed better performance by obtaining the lowest RMSE, the lowest MAE, and the highest R^2. The model obtained an R^2 of 0.50, an MAE of 8.54, and an RMSE of 11.21 on the held-out test set. The test R^2 value of 0.50 indicates moderate agreement between observed and predicted immunization coverage values under the country-level validation framework. These results indicate moderate predictive performance under a stricter panel-aware validation framework (Fig. 5; Table 7).

2.3 Feature importance

In order to determine which features have the most influence on the predicted variable, the feature importance scores were examined based on the permutation importance from the "Random Forest (Ranger)" model, which is used for measuring the decrease in the model performance when the values of a feature are randomly shuffled. As we can see from Fig. 6 below, the "UHC Service Coverage Index" (UHC_INDEX) exhibits the highest importance score by a substantial margin. This is followed by the interaction between the "UHC Service Coverage Index and healthcare workforce density"

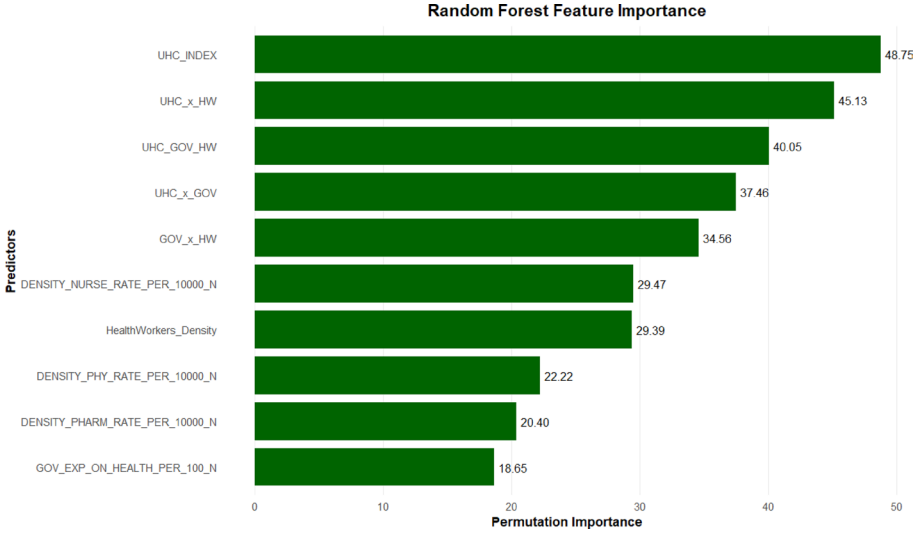

Fig. 6 Permutation importance plot of "Random Forest (ranger)" Model.
Source: Created by the author using R (ggplot2 package)

(UHC_x_HW), while the three-way interaction term "UHC Service Coverage Index × GGHE-D as percentage of GGE × healthcare workforce density" (UHC_GOV_HW) ranks next.

Other interaction terms contribute to the model to a moderate extent, including "UHC Service Coverage Index × GGHE-D as percentage of GGE" (UHC_x_GOV), and "GGHE-D as percentage of GGE × workforce density" (GOV_x_HW). Among the standalone predictors, "GGHE-D as percentage of GGE" (GOV_EXP_ON_HEALTH_PER_100_N) shows the lowest contribution, whereas healthcare workforce density (HealthWorkers_Density) contributes more substantially to the model after the "UHC Service Coverage Index".

Overall, the results show that both the "UHC Service Coverage Index" and the interaction terms contribute importantly to the prediction of immunization coverage across countries. The permutation importance results should be interpreted as indicators of predictive contribution within the Random Forest model, rather than as evidence of causal effects.

2.4 Analytical objective 2: examining whether stronger health system inputs are associated with lower NCD death rates

This analysis examines whether selected health system inputs, including the "UHC Service Coverage Index", "GGHE-D" as a percentage of "GGE", and healthcare workforce density, are associated with NCD death rates across countries. The purpose is to identify whether stronger health system inputs are linked to better health outcomes and may provide useful insights for data-informed policy development.

Correlation analysis was used to determine relationships between various structural inputs (such as the UHC Index, government spending on healthcare, and physician density) and actual health outcomes (e.g., NCD death rates), which can be useful for providing useful insights for data-informed policy development. For this analysis, four variables were employed as follows:

- NCD Death Rate - The age-standardized (30–70) mortality rate for major noncommunicable diseases (e.g., cardiovascular diseases, cancers, chronic respiratory diseases, diabetes).
- "UHC Service Coverage" Index (SDG 3.8.1), labeled as "UHC Index" in figures, a composite score ranging from 0 to 100, based on 14 indicators across four domains - Reproductive, Maternal, Newborn and Child Health (RMNCH), infectious diseases, NCDs, and service capacity and access.
- GGHE-D as a percentage of GGE, labeled as "Government Health Expenditure".
- Healthcare Workforce Density (per 10,000 population) - total number of healthcare workers (physicians, nurses, and pharmacists) per 10,000 population. Indicates the availability of medical professionals, labeled as "Healthcare Workers Density" [24].

In order to evaluate the linearity of relationships among the selected variables, scatterplots with Linear Fit and LOESS ("Locally Estimated Scatterplot Smoothing") curves were used. As we can see from the Fig. 7, while certain portions of the data exhibit a linear trend, others deviate from linearity. For this reason, three types of correlation tests were employed: Pearson, Spearman, and Kendall's Tau.

In their research article, El-Hashash and Shiekh [13] indicate that, generally, the Pearson correlation is considered a parametric measure that quantifies the linear relationship between the variables, whereas Spearman correlation is a non-parametric method used for assessing monotonic relationships. Spearman's rank correlation is also robust for outliers and non-linearity. Kendall's tau is also a non-parametric measure, but contrary to Spearman and Pearson correlation tests, it makes no assumptions about the data distribution.

Correlation strengths were indicated based on the following criteria:

- Coefficients with an absolute value of less than 0.40 indicate weak correlations.
- Coefficients with an absolute value between 0.40 and 0.69 indicate moderate correlations.
- Coefficients with an absolute value of 0.70 or greater indicate strong correlations [23].

As we can see from the table mentioned below, all three predictor variables are negatively correlated with the "NCD death rate". Furthermore, compared to other independent variables, the "UHC index" has the strongest negative correlation with the "NCD death rate", which indicates that a higher "UHC Index" is strongly associated with a lower "NCD death rate." Government health expenditure and healthcare workforce density also show a negative correlation. In general, selected predictor variables show moderate and weak negative correlations with the "NCD death rate". However, it should be underscored that correlation doesn't imply causation: a correlation between two variables doesn't imply that one variable causes another variable to occur. Of course, one variable may have a causal effect on the other, but it can be a result of confounding (Table 8).

Considering the factors above, these results support the idea that better health system inputs (such as "UHC Service Coverage Index", "GGHE-D" as a percentage of "GGE", and healthcare workforce density) are associated with lower "NCD death rates".

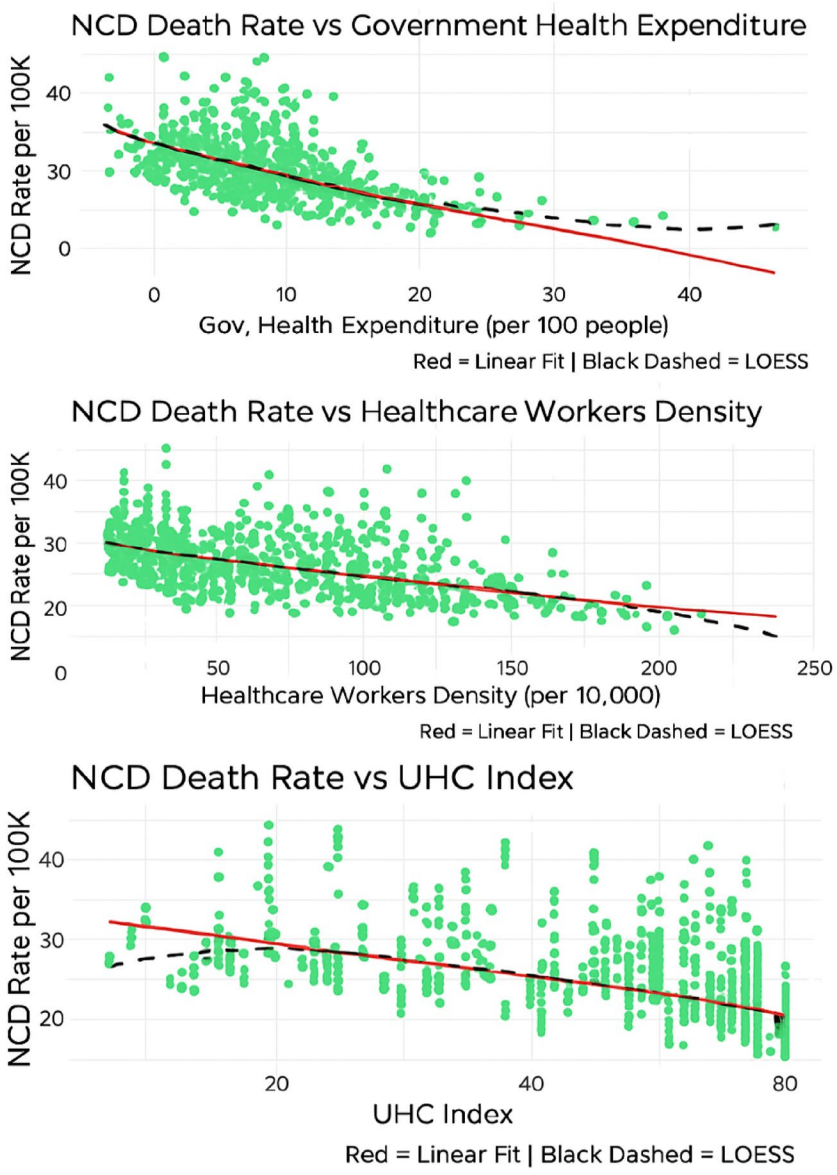

Fig. 7 Scatter plots for "NCD Death Rate" and selected variables. While some portions of the data follow a linear trend, the overall patterns reveal a consistent negative relationship-indicating that stronger health system factors are associated with lower "NCD mortality".
Source: Created by the author using R (ggplot2 package)

Table 8 Summary of correlation results.

Variable	Correlation type	Correlation value	Strength
NCD Death Rate vs. UHC Service Coverage Index	Pearson	-0.523	Moderate
	Spearman	-0.649	Moderate
	Kendall	-0.478	Moderate
NCD Death Rate vs. Healthcare Workforce Density	Pearson	-0.418	Moderate
	Spearman	-0.449	Moderate
	Kendall	-0.314	Weak
NCD Death Rate vs. GGHE-D as a percentage of GGE	Pearson	-0.531	Moderate
	Spearman	-0.577	Moderate
	Kendall	-0.413	Moderate

Source: Created by the author using R

2.5 Analytical objective 3: assessing associations between health system capacity indicators and immunization coverage

This analysis assesses whether health system capacity indicators, including the "UHC Service Coverage Index", healthcare workforce density, "GGHE-D" as a percentage of "GGE", and IHR Core Capacity Score, are associated with immunization coverage across countries.

Correlation analysis was conducted for the years 2010–2021 to explore the relationship between health system capacity indicators and immunization coverage as there was no data related to the IHR Core Capacity score before 2010. Immunization coverage was used as an indicator of health system performance and service delivery capacity (Table 9).

For the correlation analysis, these variables were employed as follows:

Spearman correlation was selected based on two key features of the dataset:

- The variables demonstrated a non-linear relationship which made using linear-based methods (i.e., Pearson correlation) less appropriate.
- And non-normal distribution of selected variables influenced by social, economic, and demographic disparities.

As we can see from the Figs. 8 and 9, the scatterplot matrix demonstrates descriptive associations between immunization coverage and several major health system indicators, emphasizing the relevance of integrated health system capacity in relation to public health outcomes. Based on the results of the correlation analysis, there is a moderate positive correlation between the "UHC Service Coverage" Index and "immunization coverage rate" (Spearman $\rho = 0.55$), indicating that countries which have a higher rate of access to essential health services are associated with higher immunization coverage rates. A similar strength of association is also observed for the "Health Workers Density" indicator ($\rho = 0.55$), indicating that greater availability of healthcare personnel is associated with higher vaccination coverage.

On the other hand, there is a weak correlation between "GGHE-D as a percentage of GGE", labeled in the model outputs as GOV_EXP_ON_HEALTH_PER_100_N", and the "immunization coverage rate" ($\rho = 0.38$). This means that higher public investment (in other words, "GGHE-D as a percentage of GGE") in healthcare is associated with better immunization coverage. Additionally, the "IHR Capacity Score" has a lower correlation ($\rho = 0.37$) meaning that while the capacity to detect public health emergencies

Table 9 Variables used in the correlation analysis.

Variable	Description
Immunization Coverage Indicator	Average of DTP3, MCV2, and PCV3 vaccine coverage rates; used as an indicator of immunization system performance and service delivery capacity.
"UHC Service Coverage" Index	Represents the delivery of key health services across populations, used in this study as a measure of overall health system capacity.
Health Workers Density Indicator	Sum of the density rates of nurses, physicians, and pharmacists per 10,000 population.
IHR Core Capacity Score (SPAR Average)	Average score across 15 core capacities, indicating a country's ability to identify, report, and respond to public health threats in accordance with International Health Regulations (IHR) standards.
GGHE-D as percentage of GGE	Indicates fiscal transparency and political commitment to health; higher values suggest stronger investments in infrastructure, data systems, and service delivery, supporting accountability in health systems.

Source: Created by the author

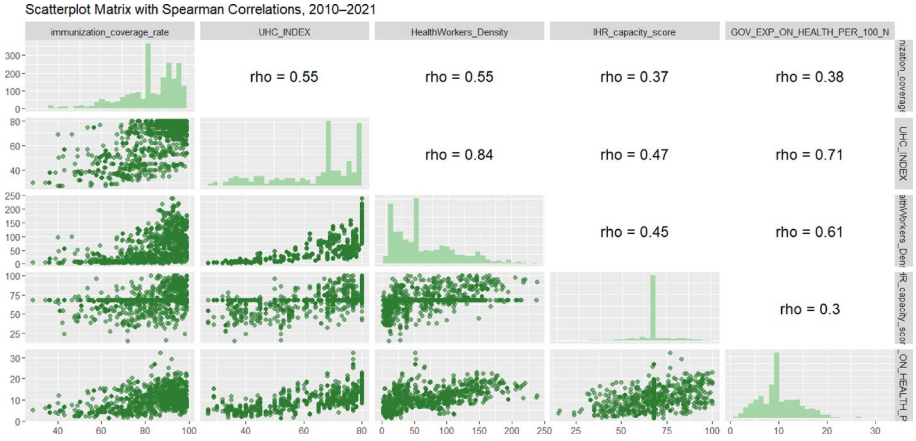

Fig. 8 Scatter plots for "immunization coverage rate" and selected variables.
Source: Created by the author using R (ggplot2 package)

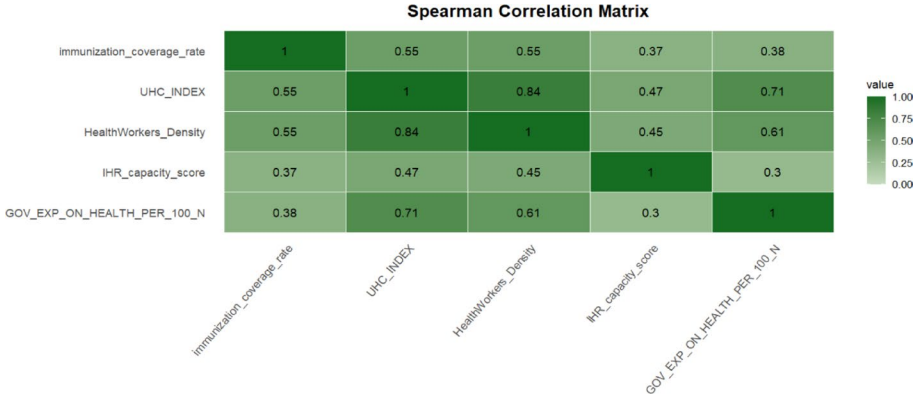

Fig. 9 Correlation Matrix of Key Health System Indicators. Immunization rates across populations in relation to GGHE-D as a percentage of GGE, the UHC service coverage index, and healthcare workforce density-key factors associated with vaccination coverage.
Source: Created by the author using R (ggplot2 package)

and threats is important to the resilience of the health system in general, it may not be as closely associated with the success of routine immunization programs. These results show that government investment in the healthcare system, high level of general "UHC Service Coverage" and the healthcare workforce density are positively correlated with immunization coverage in the pooled-country-year data. Though the strength of these correlations varies by indicators, the findings show that higher health system capacity is associated with higher vaccination coverage rates.

2.6 Analytical objective 4: assessing whether selected public health indicators contribute to the prediction of life expectancy

This analysis evaluates whether selected public health indicators, including immunization coverage, NCD death rate, healthcare workforce density, "GGHE-D" as a percentage of "GGE", tobacco use rate, and IHR Core Capacity Score, contribute to the prediction of average life expectancy.

In order to determine how the public health indicators explain variation in life expectancy, a "Random Forest" regression analysis was conducted for the years 2010–2021, as there were no data related to the "IHR Core Capacity score" before 2010.

The following variables were included in the "Random Forest" regression:

- Average life expectancy.
- IHR Core Capacity Score.
- "Diphtheria-tetanus-pertussis" (DTP3) immunization coverage rate among 1-year-olds.
- "Measles-containing-vaccine second-dose" (MCV2) immunization coverage rate.
- "Pneumococcal Conjugate vaccines" (PCV3) immunization coverage rate among 1-year-olds.
- Density of physicians (per 10,000 population).
- Density of nurses (per 10,000 population).
- Density of pharmacists (per 10,000 population).
- "GGHE-D" as a percentage of GGE per 100 population (% of total government spending).
- "tobacco_use_rate": percent of the population who use tobacco products.
- NCD Death Rate.

From the variables above, the "average life expectancy" indicator was used as dependent variable, and the rest of the variables were included in the model as predictor variables. Continuous predictors were retained on their original scales, as tree-based models are invariant to monotonic transformations and do not require normalization or standardization.

2.7 Model evaluation and performance metrics

The performance of the "Random Forest" regression model was evaluated on a held-out test set using RMSE and R^2 as the primary metrics (Table 10).

Table 10 shows that the model's RMSE and R^2 on the held-out test set were 4.13 and 0.74, respectively. The chosen predictor variables account for about 74% of the variance in "average life expectancy", according to the R^2 value of 0.74.

Additionally, a diagnostic plot of predicted and actual average life expectancy values (with a 45° reference line) was generated in order to visually inspect prediction accuracy, bias, and variance patterns. The results indicate that health system indicators contribute to the accurate prediction of life expectancy.

From the Fig. 10 above, we can see that many observations located near the 45-degree reference line which demonstrates that predicted and actual values are positively aligned. This indicates that the Random Forest model captures the general pattern in life expectancy reasonably well. Although some deviations are visible, the overall distribution of

Table 10 The results of the Random Forest model.

Model	CV RMSE mean	CV RMSE SD	Train RMSE	Test RMSE	Train R^2	Test R^2	Train MAE	Test MAE	Overfitting gap
Random Forest (Ranger)	4.05	0.64	2.94	4.13	0.85	0.74	1.53	2.89	1.20

Source: Created by the author using R

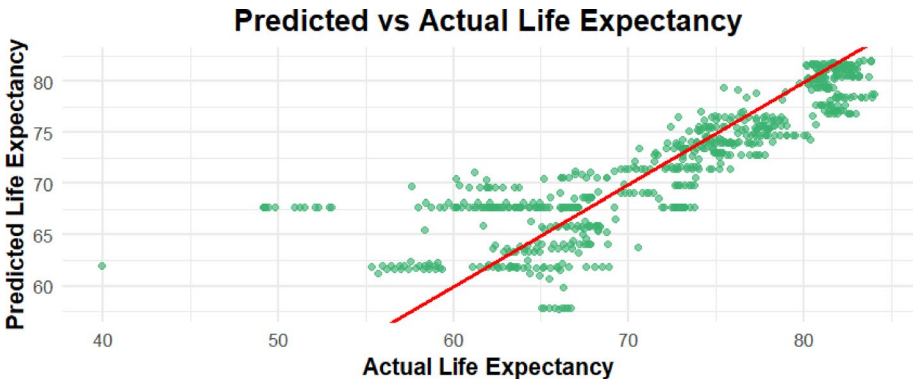

Fig. 10 Predicted vs. Actual Life Expectancy based on "Random Forest" Model.
Source: Created by the author using R (ggplot2 package)

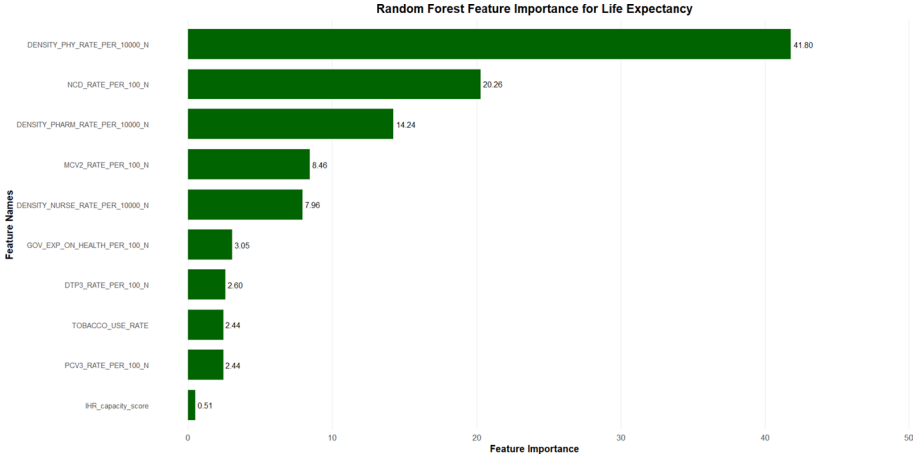

Fig. 11 Feature importance analysis based on "Random Forest" regression model.
Source: Created by the author using R (ggplot2 package)

points supports the model's good predictive performance under the country-level, leakage-free validation framework.

A "Random Forest" regression model trained to predict average life expectancy from a set of public health indicators was used to evaluate feature importance in Fig. 11. With the largest mean decrease in impurity across ensemble trees, the model determined that DENSITY_PHY_RATE_PER_10000_N ranked highest in feature importance within the model. NCD_RATE_PER_100_N and DENSITY_PHARM_RATE_PER_10000_N also showed relatively high importance in the prediction of life expectancy. MCV2_RATE_PER_100_N had an importance score of 8.46, followed by DENSITY_NURSE_RATE_PER_10000_N (7.96). Other predictors including GOV_EXP_ON_HEALTH_PER_100_N (3.05), DTP3_RATE_PER_100_N (2.60), PCV3_RATE_PER_100_N (2.44) and TOBACCO_USE_RATE (2.44) showed lower relative importance. The IHR_capacity_score had the lowest importance score (0.51) in this model.

In conclusion, by providing flexibility to capture interactions and nonlinearities, the "Random Forest" model was able to achieve high predictive accuracy. The model's findings demonstrate that country-level health system data can support accurate prediction of life expectancy under a panel-aware, leakage-free validation framework. The results

obtained from the model should be interpreted as indicators of predictive contribution within the model rather than as evidence of causal effects.

3 Results

The findings from this study indicate that the use of data analytics and ML methods to country-level health indicators can reveal meaningful patterns and associations in public health system. The predictive modeling of immunization coverage shows how interactions among health system variables can be effectively captured using ML approaches.

For Analytical Objective 1, model performance evaluation indicates that the "Random Forest (Ranger)" model achieved the best predictive performance among the evaluated models, with higher test R-squared values and relatively lower prediction error under the country-level validation framework. These results suggest that the selected independent variables capture variation in immunization coverage across countries. The feature importance analysis shows that interaction terms between the "UHC Service Coverage Index", "GGHE-D as percentage of GGE", and healthcare workforce density ranked highly in relative predictive importance, emphasizing that combined system-level factors contributed to the prediction of immunization coverage.

For Analytical Objective 2, correlation analyses were used descriptively to examine the direction and strength of associations between health system indicators and NCD death rates. Because the data consist of repeated observations within countries, p-values are not reported or emphasized. Instead, Pearson, Spearman, and Kendall correlation coefficients are interpreted as descriptive point estimates. The results show negative associations between the "UHC Service Coverage Index", "GGHE-D as percentage of GGE", healthcare workforce density, and NCD death rates. The "UHC Service Coverage Index" correlation was found to be stronger than the other two variables, suggesting that higher services coverage level is associated with lower NCD mortality rates.

For Analytical Objective 3, Spearman correlation results show positive associations between health system capacity indicators and immunization coverage. More specifically, the "UHC Service Coverage Index" ($\rho = 0.55$) and healthcare workforce density ($\rho = 0.55$) showed moderate positive associations with immunization coverage, while the IHR Core Capacity Score ($\rho = 0.37$) and government health expenditure ($\rho = 0.38$) demonstrated a weaker positive association. These results suggest that service coverage and health workforce availability were more closely associated with immunization coverage than government health expenditure or IHR capacity in the pooled country-year data.

For Analytical Objective 4, the "Random Forest" feature-importance results show that physician density, NCD death rate, and pharmacist density ranked among the highest contributors to life expectancy prediction within the model.

4 Discussion

The results demonstrate the potential of data-driven analytical methods to support the interpretation of complex associations within public health systems. In this context, the conceptual framework proposed in the study provides a structured way of how data can be systematically utilized to interpret the performance of health systems.

The findings also suggest that key health system capacity indicators, such as the "UHC Service Coverage Index", healthcare workforce density, and "GGHE-D as percentage of GGE" are associated with selected public health outcomes. Especially, the results show

that interaction terms among these variables also ranked highly in relative predictive importance, suggesting that combined system-level factors may contribute to prediction beyond individual indicators alone.

Additionally, the ML models show the potential of using multiple health system indicators to predict public health outcomes under a panel-aware validation framework. However, these results should be interpreted as predictive and associational rather than causal.

The conceptual model suggests that if data are correctly collected, analyzed, and transformed into actionable insights, public health systems can support making decisions effectively.

5 Conclusion and suggestions

The key message that can be derived from the findings of this paper is that ML algorithms combined with country-level health data can produce descriptive and predictive insights about public health outcomes. However, because of the observational nature of the data, these findings should be interpreted as identifying statistical associations and predictive patterns, and not in a causal sense.

This study explores how data analytics can be applied to structured health indicators to support public health analysis, despite of using country-level datasets. The predictive performance of the "XGBoost" and "Random Forest" models shows how ML algorithms can reveal patterns and interactions among various public health indicators. Especially, the "UHC Service Coverage" index, and the interaction terms involving health indicators, such as the "UHC Service Coverage" index, healthcare workforce density, and "GGHE-D as percentage of GGE" also ranked highly in relative predictive importance for immunization coverage. These findings highlight the importance of integrated health system components when analyzing public health outcomes.

By using various ML techniques and correlation analyses on publicly available country-level indicators, the findings show that the "UHC Service Coverage" index, "GGHE-D as a percentage of GGE", and "healthcare workforce density" are moderately associated with improved "immunization coverage rate", lower "NCD mortality", and higher "average life expectancy" across countries. These associations show the possible usefulness of data analytics in formulating evidence-based policies, even in scenarios where data are limited. It is acknowledged that the results are based on observational data and should be interpreted as associational rather than causal.

Moreover, feature importance analysis from the "Random Forest" regression model indicates that variables such as physician density, NCD death rate, and pharmacist density ranked among the highest contributors to life expectancy prediction within the model. This highlights the value of combining multiple health indicators when modeling population health outcomes.

Based on the key findings, the following recommendations are proposed:

- Promoting integrated health system approaches by considering combined effects of multiple indicators.
- Using ML models to support evidence-based decision-making.
- A high level of investment in the healthcare systems including workforce capacity and service coverage.
- Building analytical capacity for the effective use of health data in policy and planning.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12982-026-02675-7.

Supplementary Material 1.

Author contributions

Pakiza Valizada: Conceptualization of the research idea, development of the theoretical framework, methodology design, data collection and organization, primary analysis, interpretation of findings, and drafting and finalizing the manuscript. Ahmet Elbir: Provided methodological support in data analysis and provided theoretical guidance.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability

The datasets analyzed during the current study are included in the Supplementary Materials accompanying this article.

Declarations

Ethics approval and consent to participate

Not applicable, as this study uses publicly available aggregated data and does not involve human participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 5 March 2026 / Accepted: 23 July 2026

Published online: 05 October 2026

References

1. Adenyi AO, Okolo CA, Olorunsogo T, Babawarun O. Leveraging big data and analytics for enhanced public health decision-making: A global review. GSC Adv Res Rev. 2024;18(2):450–6. https://doi.org/10.30574/gscarr.2024.18.2.0078.
2. Ayemoro T, Ekundayo T. AI-driven risk stratification for targeted public health interventions. World J Biol Pharm Health Sci. 2024;20(2):988–94. https://doi.org/10.30574/wjbphs.2024.20.2.0901.
3. Badidi E. Edge AI for early detection of chronic diseases and the spread of infectious diseases: Opportunities, challenges, and future directions. Future Internet. 2023;15(11):370. https://doi.org/10.3390/fi15110370.
4. Bertsimas D, Boussioux L, Wright RC, Delarue A, Digalakis V Jr, Jacquillat A, Zeng C. From predictions to prescriptions: A data-driven response to COVID-19. arXiv. 2020. Available from: https://arxiv.org/abs/2006.16509
5. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, Specchia ML. The impact of electronic health records on healthcare quality: A systematic review and meta-analysis. Eur J Public Health. 2016;26(1):60–4. https://doi.org/10.1093/eurpub/ckv122.
6. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System. 2023. Available from: https://www.cdc.gov/brfss/
7. Centers for Disease Control and Prevention. National Health Interview Survey. 2023. Available from: https://www.cdc.gov/nchs/nhis/index.htm
8. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123–44. https://doi.org/10.1146/annurev-biodatasci-092820-114757.
9. Chow JCL, Li K. Large language models in medical chatbots: Opportunities, challenges, and the need to address AI risks. Information. 2025;16(7):549. https://doi.org/10.3390/info16070549.
10. Conderino S, Anthopolos R, Albrecht SS, Farley SM, Divers J, Titus AR, Thorpe LE. Addressing information biases within electronic health record data to improve the examination of epidemiologic associations with diabetes prevalence among young adults: Cross-sectional study. JMIR Med Inf. 2024;12(1):e58085. https://doi.org/10.2196/58085.
11. Cunningham SA, Muir JA. Data cleaning. In: Mehl AR, Brown SD, editors. The Cambridge handbook of research methods and statistics for the social and behavioral sciences. Volume 1. Cambridge: Cambridge University Press; 2023. pp. 443–67. https://doi.org/10.1017/9781009010054.022.
12. Dixon BE. Health information exchange: Navigating and managing a network of health information systems. Cambridge: Academic; 2016.
13. El-Hashash EF, Shiekh RHA. A comparison of the Pearson, Spearman rank and Kendall tau correlation coefficients using quantitative variables. Asian J Probab Stat. 2022;20(3):36–48. https://doi.org/10.9734/ajpas/2022/v20i3425.
14. Eraser. Eraser. https://app.eraser.io/
15. Ezeh CJ, Anioke SC, Oyewole S, David MG. The role of predictive analytics in enhancing public health surveillance: Proactive and data-driven interventions. World J Adv Res Rev. 2024;24(3):3059–77. https://doi.org/10.30574/wjarr.2024.24.3.3909.
16. Gartner. IT glossary – Big Data. 2011. https://www.gartner.com/en/information-technology/glossary/big-data.
17. Goetz LH, Schork NJ. Personalized medicine: Motivation, challenges, and progress. Fertil Steril. 2018;109(6):952–63. https://doi.org/10.1016/j.fertnstert.2018.05.006.
18. Google. Google Trends. https://trends.google.com/
19. Health Level Seven International. Overview-arch – FHIR v6.0.0-ballot2. https://build.fhir.org/overview-arch.html.

20. Hong W. Advances and opportunities of mobile health in the postpandemic era: Smartphonization of wearable devices and wearable deviceization of smartphones. JMIR mHealth uHealth. 2024;12(1):e48803. https://doi.org/10.2196/48803.
21. Tanwar S, Kumari A, Vekaria D, Kumar N, Sharma R. An AI-based disease detection and prevention scheme for COVID-19. Comput Electr Eng. 2022;103:108352. https://doi.org/10.1016/j.compeleceng.2022.108352.
22. Van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2018.
23. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018;126(5):1763–8.
24. World Health Organization. Indicators. Geneva: World Health Organization; 2026. https://data.who.int/indicators Accessed 2026 May 29
25. Lou HR, Wang X, Gao Y, et al. Comparison of ARIMA, DNN, and LSTM models for prediction of infectious disease trends. BMC Public Health. 2022;22:14642. https://doi.org/10.1186/s12889-022-14642-3.
26. Chow JCL, Li K. Advancement in public health through machine learning a narrative review of opportunities and ethical considerations. J Big Data. 2025;12:101. https://doi.org/10.1186/s40537-025-01201-x.
27. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.
28. Patel C, Rendell N, Sargent GR, Ali A, Morgan C, Fields R, Sheel M. Measuring national immunization system performance: a systematic assessment of available resources. Glob Health Sci Pract. 2023;11(3):e220055. https://doi.org/10.9745/GHSP-D-22-00555.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.