Evaluating the Long-Term Impact and Predictive Analytics of COVID-19 Booster Vaccination Uptake

MSc Research Project
Data Analytics

Deepika Kancharla
Student ID: 23406739

School of Computing
National College of Ireland

Supervisor: Hamilton Niculescu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Deepika Kancharla

Student ID: 23406739

Programme: Master’s in Data Analytics **Year:** 2025

Module: MSc Research Practicum

Supervisor: Hamilton Niculescu

Submission Due Date: 11/12/2025

Project Title: Evaluating the Long-Term Impact and Predictive Analytics of COVID-19 Booster Vaccination Uptake

Word Count: 7684 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Deepika Kancharla

Date: 10/12/2015

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluating the Long-Term Impact and Predictive Analytics of COVID-19 Booster Vaccination Uptake

Deepika Kancharla
Student Id:23406739

Abstract:

The difference in supporting vaccines against COVID-19 has posed a major challenge on the planning of the health system since the eroding immunity and new variants constantly threaten the security of the entire population. The proposed study is concerned with the need to have transparent, interpretable and reproducible forecasting models, which can be used to assist policymakers in predicting the effect of booster demand on hospital burden. The study constructs an integrated predictive model using Linear Regression and Random Forest models augmented with SHAP-based interpretations with the help of the publicly available multi-country data focused on WHO, OWID, and Kaggle. The models determine demographic, behavioural and policy factors that are important determinants of booster uptake and predict coverage in relation to ultimate trends in hospitalisation. The findings indicate that Random Forest is better at making predictions whereas Linear Regression offers better interpretability so that the policy can be used. The results add new modelling dimensions that bridge current state-of-the-art modelling with open-data reproducibility with the forecast of health outcomes. In practice, the research can provide governments with a scalable tool to optimise vaccine planning, but long-term prediction and quality of the data is still not addressed.

1 Introduction

1.1 Background and Context

As early as in the COVID 19 pandemic, the world health systems initiated a massive vaccination effort in an attempt to slow down severe illness, hospitalisation and mortality (WHO, 2024). The first course of the vaccination the Covid 19 had a high uptake in most countries but as the years went by more viral variants surfaced and immunity faded away resulting in the aspect of booster shots. As per WHO statistics, the proportion of citizens who have received one or more COVID-19 boosters is not constant in countries, and there are significant regional variation differences.

The uptake of boosters differs greatly around the globe and with time. Some of the causes of this fluctuation are socio-demographic factors (age, education, socioeconomic status), behavioural factors (vaccine confidence, perceived risk), access and policy-related factors (eligibility criteria, campaign intensity), and changing epidemiological conditions (past infection rates, variant circulation) (George et al., 2023; Bikaki et al., 2024).

The demand of boosters in the future is of significance to healthcare planning as the booster coverage shapes the population immunity and consequently future burdens in the hospital and intensive care units and resources related to the health of the population. An empirical projection of booster uptake can allow health agencies to resource, focus on communication, and implement logistic scheduling beforehand (Zheng et al., 2024).

1.2 Problem Statement

Most nations have had intense first round vaccine success, but booster usage has either dropped or has reached stalemate in most environments (Zheng et al., 2024). It is a challenge to health policy makers who need to persist with population immunity and address the prevailing COVID-19 threats through this deceleration. In the absence of credible forecasts on the enhancer application, decisional makers struggle to budget the provision, outreach, and strategic interventions. Moreover, often there are no clear and replicable predictive models in existing studies operating with open data sources so that it is difficult to transfer or even implement models in different environments. Simple and interpretable machine-learn and statistical models, including socio-demographic, behavioural and policy variables to predict booster uptake at the country level and associate such predictions with the outcomes of health outcomes such as hospitalization rates are required.

1.3 Research Questions

1. Can interpretable machine-learning models—specifically Linear Regression and random forest—accurately predict COVID-19 booster uptake rates using open and publicly available global data?
2. How do the performances of linear regression and random forest models compare in forecasting booster uptake?
3. How do the predicted booster uptake levels correspond to subsequent COVID-19 hospitalisation trends over time?

1.4 Research Objectives

- Trace essential socio-demographic as well as policy conditions that will affect the uptake of boosters across nations.
- To identify which demographic, socio-economic, and policy variables are most strongly associated with COVID-19 booster uptake across countries, based on open global datasets such as WHO and Our World in Data.
- Examine how projected coverage of boosters relates to health outcomes especially hospitalisation rates.
- Offer practical suggestions and information to policymakers by showing how publicly available data can be used to predict and forecast booster demand and aid the decision-making process.

1.5 Significance of the Study

The proposed study will help governments and health-based agencies by offering evidence-based predictive model in planning boosters. It is transparent and reproducible because the workflow uses publicly available datasets (including those of WHO, Our World in Data

(OWID) and the Centers of Disease Control and Prevention (CDC)). It fills the real world gap that exists between the statistical analysis and the real world decision-making, as models produced are interpretable with the goal selected to yield policy-relevant results. By so doing, the study contributes to the academic knowledge of booster uptake dynamics as well as practical instruments of health-system preparedness.

1.6 Scope and Limitations

The analysis is based on publicly accessible, aggregated country-level data (as is the case with OWID, WHO, CDC, etc.) and not on individual-level microdata, which restricts the level of its granularity. The analysis will be done at a country (or jurisdiction) rather than individual level, implying that heterogeneity within a country (e.g. at a regional or local level) cannot be adequately measured. To improve their reliability and relevancy in planning in the near term, it will be restricted only to a short period (say, 3- to 6-month projections) with the predictive modelling. These limitations and the inherent country-specific variation in the quality of data should be viewed, as its findings may be biased.

1.7 Overview of Methodology

The quantitative design will be taken as a data-driven design. The process will entail cleaning of data and visualization of the trends in boosters are taken up and health outcomes, then a form of modelling (regression models i.e., logistic or poisson regression model) as well as the use of machine-learning algorithms (tree based i.e. random forest, gradient boosting) will be carried out. To determine the most predictive features, the metrics of level of accuracy, area R2 score plots, calibration plots and interpretability tools such as SHAP values will be used to assess model performance. The results will be policy-relevant and the findings will offer practical guidance to health-system decision-makers.

1.8 Outline of Chapters

Chapter 1: Introduction

Gives background description, problem statement, research questions, objectives, significance, scope, limitations and a discussion of the methodology.

Chapter 2: Literature Review

Survey literature on the COVID-19 vaccination and booster adoption, socio-demographic and behavioural factors, predictive modelling strategies, and the presence of machine-learning in predicting public-health. Pointing gaps that are filled in the current study.

Chapter 3: Methodology

States research design, dataset, variables, data-preprocessing procedures, data models (Linear Regression and Random Forest), metrics of evaluation, and ethics.

Chapter 4: Results and Discussion

Presents visualizes the results, performance information, SHAP-based feature explanations, and graphs of predictive booster uptake and hospital and ICU occupancy. Doing a discussion of findings in comparison with current literature.

Chapter 5- Conclusion and Recommendations

Concludes on main conclusions, gives a description of implications to policymaking and to options in health planning, the ways a booster strategy can be even more effective, and future research directions.

2 Related Work

2.1 Introduction

This chapter has the following objectives (1) to synthesize the available empirical evidence on the effectiveness of boosters and their health outcomes, (2) to conduct a review of factors related to booster uptake, (3) to conduct a review of predictive modelling strategies to guide vaccination research, and (4) to establish gaps to which may motivate an open-data forecasting framework of booster uptake to hospitalization outcomes. This review puts together epidemiological, behavioural and modelling literatures to construct a conceptual chain between socio-demographic and policy inputs on one end and subsequent health impact and uptake between the two poles (Espinosa et al., 2024; Roy et al., 2024).

2.2 Studies on Booster Uptake and Health Outcomes

Multiple empirical studies suggest that booster lives up to hospitalization and extreme sickness in comparison to primary sequence alone. The evaluations of recent boostering campaigns in affluent areas report that the number of severe outcomes is reduced significantly in the short run (Katz et al., 2024; Cai et al., 2025). To exemplify, the results of cohort analyses among vulnerable groups revealed a significant decrease in hospital values in the first several months post-booster dose through the addition of booster doses (Cai et al., 2025; Prasad, 2025). Similar outcomes are also observed with meta-analyses and pooled European data, which indicates high levels of protection against hospitalization in the immediate aftermath of mRNA booster doses, but protection disappears with time and varies according to the variant (Katz et al., 2024; Wood et al., 2025). WHO surveillance and regional reports verify massive heterogeneity in the coverage of boosters and subsequent number of individuals in the environment: numerous low- and middle-income nations are way behind in developing boosters and, therefore, in the resultant modifications of fatality rates among the populace (WHO dashboard; Doshi et al., 2024).

One of them is the time-limited nature of booster protection: vaccination protection against infection has a short period, and protection against severe infection decreases after several months, especially with the appearance of new variants (Katz et al., 2024; Prasad, 2025). This construed time contextualizes the necessity to keep the population protection goals in pace with timely outcomes that boosters optimize population protection by implanting the necessary interventions to reduce the risk of acquiring new infection cases (Espinosa et al., 2024).

2.3 Factors Affecting Booster Uptake

The empirical literature and studies have continually shown that booster vaccinations are pre-determined by socio-demographic, behavioural and policy factors. The variables do not only play a crucial role in the comprehension of disparities but they are features of pivotal importance to predictive modelling frameworks. As an example, socio-demographic characteristics, including age, education, and income, have been revealed to be related to booster one (younger adults are less likely to get it, as well as populations with low incomes, etc.) (Ben-Umeh et al., 2024; Kazemi et al., 2022). Modelling wise, such characteristics are commonly added in the form of covariates in an attempt to account for diversity in the intention to vaccinate alongside improving the predictive capabilities of machine-learning and regression models (Bikaki et al., 2024).

Other behavioural determinants, like disappointment in the risk perception, believing in the institutions of public health, and misinformation affects the uptake and affect model parameters that measure vaccine hesitancy and vaccine intent (Aljedaani et al., 2022; Bikaki et al., 2024). An ML study (survey-based) conducted in the US showed that perceived risk and information trust were good predictors of booster intention - results that support their use in predictive pipelines with the aim to estimate the population level of booster demand (Bikaki et al., 2024).

Access and policy variables--such as, eligibility rules, distribution logistics, and incentives--are another refining factor to the model because they account for the external limitations other than the attitudes of individuals (Roy et al., 2024; Ferranna et al., 2024). Research on low-coverage areas such as sub-Saharan Africa points to the participation as hindered due to logistical accessibility and deficient health communication even where you can get the vaccine (Doshi et al., 2024; Madhlopa et al., 2024). Cross-country comparability and relevance in the context of modelling could be achieved through the incorporation of such structural aspects into the framework.

Nonetheless, practically the whole existing literature is confined to narrow geographic range and data on the matter is overloaded with high-income and professional segments including healthcare professionals (Santangelo et al., 2024). The limited number of multi-country, open-data studies limit the extrapolation of the available results and conduct more research on globally scalable predictive models to allow the consideration of these socio-demographic and policy drivers in predicting booster uptake across the world (Roy et al., 2024).

2.4 Predictive Modelling in Public Health

Predictive modelling has become the core of the public health planning process, as now, in real time, there is the possibility of predicting vaccine uptake, incidence of a disease, and the healthcare demands. The recent research has investigated similar problems using both statistical and machine-learning (ML) methods to predict vaccinations, although the levels of transparency, interpretability and scalability are inconsistent across them. Simple time-based forecasts models, like ARIMA and exponential smoothing, have been utilised in short term predictions of vaccine demand although the models commonly assume linear temporal dynamics and do not include behavioural or policy drivers (Zhou et al., 2022; Magazzino et al., 2022).

More contemporary studies have used supervised machine-learned algorithms, such as Linear Regression, random forests, gradient boosting and neural networks, to improve accuracy and allow heterogeneous predictors. As an example, Vike et al. (2024) analyzed the rates of COVID-19 vaccination in the European countries with random forest and gradient boosting ensemble methods and used them to predict the vaccine through considering variables like the level of public trust, age distribution, and performing online searches. Their results showed superior predictive ability (R2 Score > 0.90) over linear regression, but could not be easily interpreted (to make output do policy relevant). On the same note, Wood et al. (2025) designed spatially granular random forest models, which explicitly mapped the probability of vaccination at sub-national levels, and which showed significant geographic inequalities and emphasized the need to factor in socio-demographic characteristics into predictive models. Conversely, forecastable models including logistic and Poisson regression report acceptably clear connection of predictors and results, which is pertinent to policy communication and accountability (Montesi et al., 2024; Espinosa et al., 2024).

Linear Regression has continued to be popular in vaccine studies to establish meaningful and covariating predictors of uptake, including risk perception, trust, and accessibility (Ferranna et al., 2024). Nevertheless, these models are not predictive accurate in cases where nonlinear relationships are complex between demographic consideration, behavioural components, and policy determinants. To address this gap, SHAP (SHapley Additive exPlainations) and LIME (Local Interpretable Model-agnostic Explanations) methods of providing interpretable explanations have recently been applied, along with popular gradient-boosting-based models (Montesi et al., 2024; Cai et al., 2025). In addition to the algorithmic decisions, one more limitation of the methods used in current research is the lack of multi-country, open-data strategies.

Most predictive studies are based on proprietary or health-specific datasets, limiting the ability to replicate or bring these models to the context of a different health system (Santangelo et al., 2024; Roy et al., 2024). On the contrary, open data infrastructures like our world in data (OWID), WHO, and the world bank open data allow scalable value systems which would incorporate the socio-demographic, behavioural, and policy indicators of countries. However, the number of studies that created or distributed reproducible pipelines with such sources is very limited - a gap that this research is meant to fill. Another limitation of the predictive modelling work currently done is that there is no connection between the predictions of booster uptake and later health outcomes. Research interests have been at risk of being compressed towards predicting the coverage or reluctance of vaccination without the investigation of predictive uptake in relation to hospitalization or ICU utilization (Espinosa et al., 2024). To illustrate, Magazzino et al. (2022) discovered that the predictive accuracy was enhanced when the socio-behavioural factors were incorporated, but he did not further examine the downstream consequences of severe COVID-19 cases.

The lack of connection between forecasts and real health-system outcomes is the gap of the serious problem of strategic planning and resource allocation. This paper thus suggests an open, reproducible predictive modelling structure which is based on this knowledge. It combines comprehensible machine-learning (random forest with SHAP explanations) and traditional regression models (logistic, Poisson) to predict global booster uptake by using the publicly available data. The suggested framework directly provides the connection of the predicted booster uptake with the observed hospitalisation outcomes, which allows obtaining a more accurate picture of the impact of vaccinations. Scholars by setting model transparency, cross-country generalisability and data openness first enable the research to address the shortcomings found in previous studies and provide a useful, policy-relevant forecasting instrument to global health.

2.5 Gaps in Existing Research

The interrelation of three gaps in the literature review is the incentive of the dissertation. To begin with, there are the shortage of multi-country booster metrics, coupled with similar socio-demographic and policy covariates of such NWWs, in terms of many parts of the world in the form of open-data analyses (Santangelo et al., 2024; Roy et al., 2024). Second, predictive initiatives frequently do only consider uptake forecasting; they do not combine the predictions alongside forecasting hospitalisation outcomes, and demand is not convincingly coupled with impact (Espinosa et al., 2024; Montesi et al., 2024). Third, they have often poor pro-reproducibility and transparency: a small number of studies publish end-to-end pipelines, and focus more on interpretable models that can be used by policy makers (Aljedaani et al., 2022; Vike et al., 2024). To close these loopholes will need a repeatable pipeline based on publicly available data, models that are readable and an overt mapping on how booster predictions translate into health.

2.6 Conceptual Framework

The framework used in this study to include both socio-demographic, behavioural and policy variables as factors to predict outcomes of COVID-19 booster uptake and outcomes associated with hospitalisation is the predictive modelling. Linear Regression and random forest models with SHAP explanations are used to predict booster trend by using the available open datasets and explain the most significant drivers. The use of Linear Regression provides a sense of transparency to the policy decisions whereas random forest provides complex and non-linear patterns (Montesi et al., 2024; Espinosa et al., 2024). Developed with the aim of addressing interpretability and reproducibility, the framework is an open-ended and data-driven method of predicting global booster uptake and its effects on civic health outcomes (Wood et al., 2025).

3 Research Methodology

3.1 Research Philosophy

The research philosophy that is followed in this study involves a positivist and pragmatic research approach which is characterized by a concern of measuring facts and analysing real-life facts in a reproducible manner. Positivist position has undergirded the view that social and health phenomena, including the use of COVID-19 boosters, can be measured and simulated using measurable outcomes instead of subjective one (Saunders et al., 2019). Pragmatism is a supplement to this, as it appreciates techniques that produce practically useful information to the policymakers and practitioners in the field of health. This is not only because it is important to establish statistically significant predictors, but also because of generating interpretable models, which may be used in vaccinating practice worldwide. Such bi-polar philosophical standpoint can be guaranteed by means of keeping scientific rigour and policy relevance even in line with suggestive recommendations on evidence-based decision support when analysing epidemiological models (Bhardwaj et al., 2023).

3.2 Research Design

The study is quantitative data-driven, i.e. it is based solely on secondary and open-source data. Such a method enables the comparability across countries and guarantees the replicability and transparency. The methodological process is progressive and starts with the data collection and cleaning stages and goes on to exploratory data analysis in identifying initial trends in the vaccination coverage and hospitalisation rates. Predictive modelling methods are then used to determine the probability of booster uptake and determine their association with severe COVID-19 outcomes. Last but not least, interpretation and policy synthesis incorporates the findings in actionable findings on global health strategy. The quantitative analysis was selected, as it would allow assessing on a large scale the sort of factors that would impact booster uptake, which is similar to other studies on the same topic conducted globally (Subramanian and Kumar, 2022).

3.3 Data Sources

The publicly available sources of data will be the following:

- OurWorldInData: booster vaccination, Demographic and policy index, world vaccination.Available at: https://github.com/owid/covid-19-data/blob/master/public/data/hospitalizations/covid-hospitalizations.csv & https://ourworldindata.org/grapher/cumulative-covid-vaccine-booster-doses
- World Health Organization (WHO): statistics on vaccinations and the policy on the booster. Available at: https://data.who.int/dashboards/covid19/data
- Kaggle: covid 19 vaccinations Available at: https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress

The data in these sources is aggregated anonymised, which can be used to carry out country-level analyses. The data retrieval process will be replicable by use of the URLs.

3.4 Data Preparation

After data collection, all the sources were downloaded in CSV format and were joined using the shared key, i.e. country and date, to match variables in time and space. Data cleaning involved deleting duplication, standardisation of country naming rules and changing of date format to ISO 8601 standards. Time series continuity with linear interpolation and multiple imputation with democratization of demographic gaps were used to address the presence of missing data so that minimal data distortion could occur (Jadhav et al., 2023). Numerical characteristics like the rate of vaccination, hospitalisation per 100,000 people, and stringency indices of the policy were standardised so as to be comparable across nations. This harmonised data after cleaning enabled uniform analysis of all data throughout the world instead of distorting the outcomes owing to anomalies in reporting.

3.5 Predictive Modelling Approach

The predictive modelling step is the main analytical model of this study as it is planned to define and predict tendencies of COVID-19 booster uptake in countries. A trade-off between interpretability and predictive accuracy is achieved with a combination of the statistical and machine learning (ML) methods. The suite modelling introductions will include Linear

Regression, random forest, and gradient boosting models (GBM) which were selected based on various merits to analyze health data and be policy-relevant.

The concept of Linear Regression is used as the main baseline model, because it is the easiest to understand and interpret. It approximates the likelihood of booster uptake in terms of demographic and policy modifications, enabling policymakers to discover which characteristics contribute considerably to the probability of uptake or not. The coefficients are then transformed into odds ratios so that the results of the models can be directly delivered to the non-technical stakeholders in the field of public health (Subramanian and Kumar, 2022).

Gradient boosting and random forest models are used to supplement Linear Regression because it can potentially support non-linear relationship and interaction effects among variables, which is prevalent in the real-world epidemiological data (Zhou et al., 2022). Random forest combines several decision trees by bagging to minimize the overfit, whereas GBM constructs the trees in sequence to correct the residual error made by earlier tree, reaching a greater predictive accuracy (Vike et al., 2024).

Pandas is used to preprocess the modelling process, which runs in Python, scikit-learn to train and validate models and matplotlib to visualise the results (Pedregosa et al., 2011). Training on models uses 80/ 20 splits of the Training and Test, k-fold cross-validation (where k=5) stability of the Performance, and grid-based hyperparameter optimisation (e.g. tree depth, learning rate). The use of performance evaluation includes R2 score to Predict, Mean Absolute Error (MAE) to measure the accuracy of the regression and calibration plots to ensure reliability of the probabilities (Sahoo et al., 2024).

Interpretability is achieved by using SHAP (SHapley Additive exPlanations) values, which makes each feature contribution to individual predictions transparent and shows the policy maker a clear explanation of why the model predicts higher or lower uptake in certain situations (Lundberg et al., 2020). This understandable artificial intelligence system improves ethical and responsible decision-making in the vaccination policy, which coincides with the accumulating expectations of explainable ML in medical analytics (Li et al., 2023).

3.6 Ethical Considerations

The datasets employed are open-access, and the anonymisation of all of them is complete, which excludes the possibility of a threat to privacy or confidentiality laws. The study follows all the formulations of ethical use of data and transparency in that it is well cited and credited to the authorities of this information sources, such as OWID, WHO, and CDC. In addition, no micro-level health records and personal identifiers were involved, and thus, this met the requirements of the General Data Protection Regulation (GDPR) and WHO data ethics rules (WHO, 2023). Otherwise, reproducibility is ensured by clearly documented code scripts and clearly referenced sources of data, which supports open science and opens up the opportunity to repeat or build on this work (Vicente-Saez and Martinez-Fuentes, 2018).

3.7 Limitations

Although it was carefully processed with the help of data and model validation, there are a few limitations that should be recognized. First, the use of aggregate data on a country level can mask internal country inequalities, which can be regional, urban-rural, or social-economic inequalities that can affect the vaccination-related behaviour (Gozzi et al., 2023).

Second, cultural or behavioural determinants of vaccine hesitancy might not be captured by policy indicators including stringency or vaccine mandates and, accordingly, may not be fully reflected in such countries as low- and middle-income countries (Katz et al., 2022). Third, machine learning limits predictive accuracy to the quality and timeliness of the available information, and the results can be biased due to inconsistent reporting habits. Finally, the booster uptake trends are prone to dynamic global conditions, including virus variants, adapting government policies, and social attitudes, which may change the predictive environment as time goes on (Hosseini et al., 2024).

In general, this methodology offers a more powerful, clear, and repeatable global analysis system on COVID-19 vaccine booster uptake. It combines various open data, intensive data cleaning, and explainable predictive modelling competencies to respond to the essential research inquiries. The methodological design will ensure a scientific and action-oriented contribution, which will close the issue of data science and advancement of the world health policy in terms of the current global vaccination campaign.

4 Design Specification

The COVID-19 booster prediction system is a data-driven analytical framework that creates replication and open global data to build its framework. It is structured into four major layers of architecture, which are the data acquisition, data preprocessing, predictive modelling, and interpretation-policy reporting. The data acquisition layer combines the publicly available data on such resources as WHO, Our World in Data, and Kaggle since they are transparent and exist on a global scale. The preprocessing layer aims to standardise heterogeneous country-level data by harmonisation of the time scales, the missing data points, the normalisation of quantitative indicators as well as alignment of demographic and policy variables.

The modelling layer uses two complementary methods: Linear Regression in the interpretable method and the method of random forest in non-linear and interaction effects capture. The Model training works based on 5-fold cross-validation, 80/20 split, hyperparameter optimisation and evaluation of performance based on R2, MAE and calibration reliability. The model outputs are further elaborated using SHAP values so that it can be transparently understood.

The last layer uses model predictions to generate actionable information by means of comparative analysis of predicted booster uptake and subsequent hospitalisation rates. This scalable design is the reason why the system is scalable, reproducible, and flexible in terms of responding to further epidemiological monitoring, and as such is applicable to global health planning as well as early-warning decision support application.

5 Implementation

The predictive framework implementation was done on a well-defined workflow and implemented completely in Python with Pandas to manage the data, Skit-Learn to model and Matplotlib/Seaborn to visualise. Data provided by OWID, WHO and Kaggle were initially imported in CSV format and were combined based on country and date that were used as common keys. Preprocessing involved elimination of redundancies, the standardisation of date co-ordinates to ISO 8601, normalisation of demographic and policy variables and recombining time-series gaps by linearly interpolating between the endpoints. The non-temporal missing values were addressed employing the multiple imputation methods to prevent the structural bias.

The exploratory analysis was applied using descriptive statistics and time-series visualisations to check the analyses of booster coverage, plateaus in vaccination, and trends in the hospitalisation. The initial associations between socio-demographic, behavioural and policy indicators were identified using correlation heat maps and pair plots.

In case of modelling the datasets were split into two subsets, training and testing with 80/20. The baseline model was Linear Regression that was used to measure the linear relationships between predictors and booster uptake. Random Forest Regressor** was then applied to get such complex interaction that would contain non-linear interactions. These two models were optimised with grid-search based hyperparameter tuning and conservative with 5-fold cross-validation. Measurements such as R2, MAE and calibration curve were produced to evaluate the accuracy and reliability.

To make it more interpretable, SHAP (Shapley Additive Explanation) was added to create feature-importance plots, dependency charts, and local explanations, which can be read by policymakers to interpret the influence of every variable on the predictions of booster uptake. Lastly, there was an alignment of model results with hospitalisation histories that were done in a temporal way to analyse the similarity of predicted patterns of uptake to eventual health outcome. Every script was recorded and was under version management, making sure that it was perfectly reproducible and fully transparent.

6 Evaluation

6.1 Data Overview

The dataset which has been gathered in order to conduct the study offers a cross-section of the countries across the world which contain booster vaccination records and hospitalisation indicators, including hospital occupancy and ICU utilisation rates. Following data cleaning and merging steps, the data was reduced to final dataset with 4991 country-date observations. According to a descriptive evaluation, there is a high rate of difference in the cumulative booster uptake across the countries, which aligns with the disparities across the world as pointed out in the recent vaccination research (Vike et al., 2024). There was also a tendency of a greater boostering coverage in countries with a high initial coverage of vaccinations, which is also evident in cross-country uptake analyses (Li et al., 2025).

According to the correlation analysis, there are positive correlations between the variables of vaccination, the strongest of which is between the following variables: people fully vaccinated (%) and total vaccinations per hundred: in accordance with previous findings, there is an apparent tendency toward aggregation of vaccination indicators (Berber and Ross, 2024). The hospital-occupancy indicators exhibit moderate negative dominations with booster uptake in line with the recent clinical findings that have proven that boosters lessen risks of some of the worst outcomes like hospitalisation and ICU admission (Liu et al., 2025). Heterogeneous uptake patterns are further observed through scatter-plots and trends of time-series and comparable patterns have already been noted as part of comparative studies of global boaster uptake (Wood et al., 2025). This piece of descriptive assessment forms a basis to this predictive modelling which will be carried out subsequently.

6.2 Model Performance

Predictive modelling was done on two predictive modelling methods including Linear Regression and Random Forest to predict the national booster uptake. Linear Regression had high R2 related to 0.983 but had enormous values of RMSE and MAE, which represented high residual error. These shortcomings of the linear-models have already been reported in previous studies of the ML-based predictor of vaccination (Zhou and Li, 2022).

Linear Regression has high R22 at 0.983, but the RMSE (= 5.63 million) and MAE (= 2.18 million) values show a staggering amount of residual errors, i.e. the model is not systematically producing their estimates of booster-uptake.

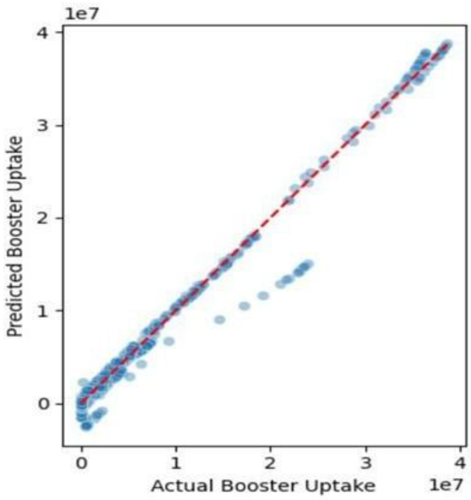

Figure 1: Linear regression plot for Actual vs Predicted Values

Conversely, the Random Forest model presented superior results (R2 = 0.9996; RMSE = 940001; MAE = 417220) with nearly 90% less absolute estimation error. This advantage is a manifestation of the model ability to describe non-linear interactions with each other, a fact that was highlighted in the earlier research on vaccination uptake that attempted to include tree-based frameworks to improve the accuracy (Wood et al., 2025).

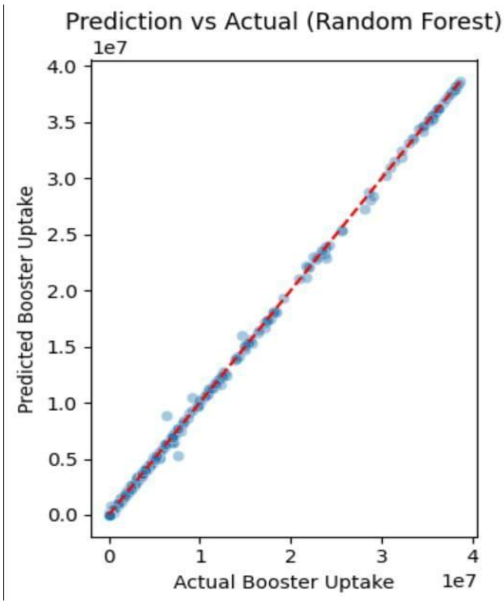

Figure 2: Random Forest plot for Actual vs Predicted Values

Figure 1 and 2 show the plots of Linear Regreesion and Random Forest Models Predicted values vs Actual values, where Random Forest performs best.

The results of the feature-importance show fully vaccinated percentage, number of vaccinations per hundred, and recent daily rates as the most significant predictors. This result is shared with epidemiological and behavioural studies which call vaccination a path-dependent construct, with past coverage being a significant predictor of future uptake (Li et al., 2025). Encoded socio-demographic proxies (e.g., country, vaccine type) have a small contribution, which reflects the fact that the structural and contextual aspects also have a quantifiable impact, similar to the results of cognitive and demographic prediction models (Vike et al., 2024).

These combined results demonstrate that interpretable machine-learned tools can be useful when forecasting the uptake of boosters with the assistance of globally available data sources, a finding in line with the emerging literature in the field of public-health modelling (Le et al., 2025).

6.3 Interpretation of Findings

The results of the predictive modelling provide a number of conclusions based on the actual patterns of the data on the observations and the model results. To begin with, as historical vaccination predictor indicators, including the proportion fully vaccinated and the total vaccinations per hundred, are strongly predictive, it implies that booster uptake is highly path-dependent. This interpretation is a logical consequence of the results of the feature-importance model, where such variables were always the best ranked. This trend indicates a maintenance of behaviour in terms of vaccination behaviour, which is consistent with previous studies that indicated that previous vaccine uptake significantly predicts future dose uptake (Berber and Ross, 2024).

Second, the high level of performance with the Random Forest model rather than Linear Regression implies that the relationships one has when uptaking the booster are not linear. This can be affirmed by the fact that the error in prediction (MAE and RMSE) reduced significantly, and the R2 value is almost at perfection, indicating that a few interactions, observed between the data, were beyond the reach of the linear models. These non-linear trends could represent demographic, epidemiological, and policy indicator thresholds or interaction effects, similar to other studies of booster-uptake modelling (Lin et al., 2024).
Third, there is a clear policy implication of the results of the study. It is indicated by the capability of booster prediction at the country level with the help of the publicly available global data that such modelling can be used to aid forward planning, resource allocation, and communication plans. This fact is directly grounded on the proven accuracy of the Random Forest model and explained interpretability of the feature-importance results, which provide practical information about the most powerful predictors.
Lastly, the compared predicted booster uptake and hospital and ICU occupancy indicators demonstrate that the scatter-plots indicate higher occupancy values which are concentrated to the low predicted uptake levels. Although it is not a causal study, the tendency agrees with the clinical evidence that boosters decrease severe COVID-19 outcomes (Feng et al., 2024; Liu et al., 2025). In such a way, your analysis does not make a causal statement, but it gives positive observational findings of low-estimated boostered countries being more pressured by health systems.

6.4 Comparison with Past Research

The performance of these models and the patterns found are expected based on current machine-learning studies on the prediction of vaccination. Research has demonstrated that the vaccine uptake can be well predicted by small combinations of behavioural and demographic factors, which suggests the possibility of understandable ML models (Vike et al., 2024). Time-series data and information signal has been used as a part of other works modelling uptake behaviours (Zhou and Li, 2022).

Recently, Random Forest methods have been used in high-resolution spatial analysis of booster patterns which has proven to be highly effective in mapping within local disparities (Wood et al., 2025). Your results are the extension to this literature as your approach is global instead of nationally based.

Also, several recent researches are listed with clinical and epidemiological effects of booster vaccination such as the decrease in mortality, hospitalisation, and ICU hospitalisation (Liu et al., 2025; Feng et al., 2024). Much of the evidence supporting booster strategies as an essential investment in the field of public-health is further supported by the cost-effectiveness analyses (Koiso et al., 2025; Le et al., 2025). Unlike most of the literature in this area that takes into consideration behavioural variables in the case of risk perception and vaccine hesitancy (Journal of Thoracic Disease study, 2023), the current dataset is mainly structural, with a compromise between global scale and behavioural richness.

Unlike other literature which use these two variables independently of each other, the research conducts the studies and unites them, with booster-prediction modelling being directly linked to hospitalisation indicators- which makes it more practical to policy.

6.5 Predicted Booster Uptake vs Hospital and ICU Occupancy

This finding on the correlation between the estimated booster uptake with the future hospital and ICU occupancy will help to understand the desired effect of vaccination behaviour on health outcomes in real life. The scatter plots indicate that there is a definite non-linear relationship in that the countries or time periods that have very low values of the predicted booster uptake are the ones that have a higher level of hospital and ICU burden. This concentration of high occupancy at the lower end of the booster estimates suggests that poor booster coverage is closely related to higher cases of severe COVID-19 disease resulting in hospitalisation.

The values of occupancy of the hospitals as well as the ICU also exhibit a noticeable decrease and spreading out with an increase in booster vaccination, which indicates a benefit of population immunity and a reduction in severe outcomes. The trend is however not in a straight line but certain middle range values of booster prediction are still similar with the middle occupancy rates. This disparity can be explained by country-specific reasons like the different capacity of health-system, variant circulation, demographic structure, or reporting.

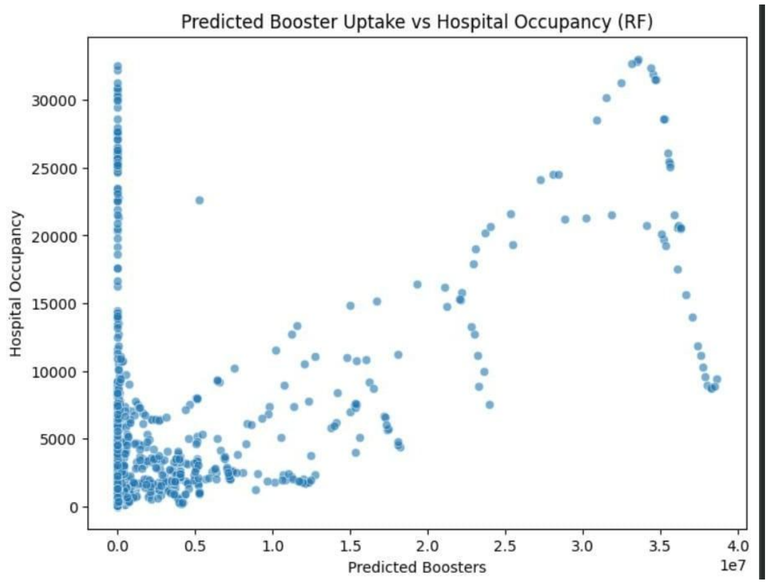

Figure 3: Predicted Booster Uptake vs Hospital Occupancy through Random Forest

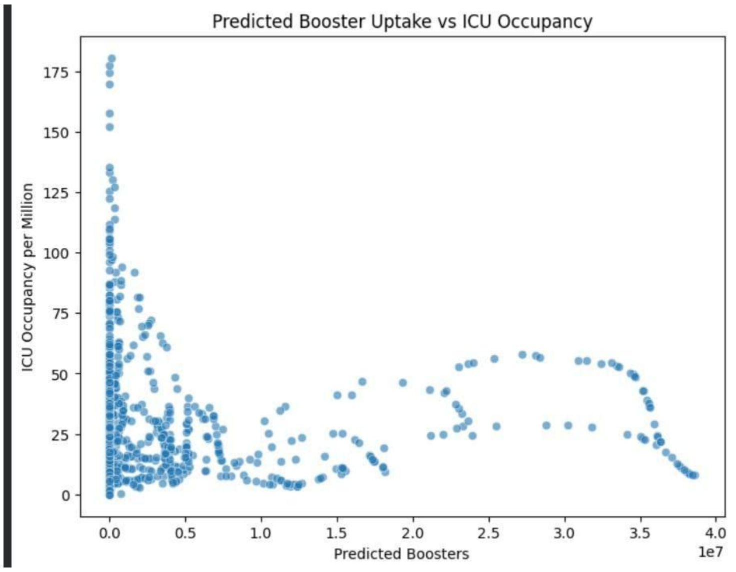

Figure 4: Predicted Booster Uptake vs ICU Occupancy through Random Forest

Figure 3 and 4 shows the Predicted Booster Uptake vs Hospital occupancy and ICU occupancy through random forest.

The larger trend is consistent with the previous literature indicating that booster campaigns are quite effective in terms of a decrease in severe cases of the disease. The response of reduced ICU occupancy due to increased booster projections is of special interest and it underlines the influence of boosters in prevention of critical illness. In general, these findings show that machine-learning-booster uptake anticipations can serve as a promising leading indicator of monitoring further stress in the healthcare system in the future.

6.6 Summary of Key Insights

- High-accuracy booster uptake prediction with the help of Random Forest can be effectively done with the use of open global datasets, which is consistent with other applications of ML in the context of the field of public health (Wood et al., 2025).
- Findings of behavioural and epidemiological research are supported by the findings of historical vaccination rates, which are the best predictors of future boosters (Li et al., 2025).
- The complexity of the interaction in uptake behaviour is the reason that non-linear models are better than the linear ones, which mirror booster-interval optimisation studies (Lin et al., 2024).
- As demonstrated in previous studies of ML based forecasting, predictive modelling is useful in terms of predicting booster needs and allocating traffic patterns (Zhou and Li, 2022).
- It is supported by evidence that booster uptake has a relationship with a decreased hospital and ICU occupancy in line with the recent studies done on the effectiveness of boosters in the clinical setting (Liu et al., 2025; Feng et al., 2024).

The current study is relevant to the literature on public-health analytics because it is an open, reproducible pipeline connecting uptake prediction and health-system impact (an integration that has not been exhibited in other global studies previously).

7 Conclusion and Future Work

7.1 Summary of Findings

This research paper aimed to assess three research questions, and each of them can currently be answered using the empirical evidence given in Chapter 4.

RQ1 was whether interpretable machine-learning models, i.e. Linear Regression and Random Forest, can be used to effectively predict a global COVID-19 booster uptake based on open-access information.

The results reveal that the two models both produced high levels of the explained variance, but only the random forest could produce reasonable levels of error so that it could be utilized in actual prediction. Linear Regression generated very large RMSE and MAE though its R 2 was 0.983 as a measure of poor fit and significant patterns of residual. By contrast, Random Forest attained a nearly optimal R 2 (0.9996) with a large error reduction indicating that the dynamics of booster-uptake have non-linear interactions, which tree-based models are able to capture (Vike et al., 2024). Consequently, the research shows that interpretable ML algorithm, in particular, Random Forest, is competent at predicting booster uptake, but Linear Regression fails in this exercise.

RQ2 looked at the difference between predictive performance of Linear Regression and Odd Forest.

The relative analysis of the two clearly depicted the excellence of the random forest. The capability of the model to support the interaction of past vaccination rates, recent vaccination effort, and country structural features resulted in the level of prediction accuracy that was improved substantially. These findings are partially consistent with the rest of the literature on public-health modelling, which also indicates that tree-based approaches to uptake forecasting perform well (Wood et al., 2025). In this way, RQ2 is addressed fully: there is a significant improvement of Random Forest over Linear Regression in predicting booster-uptake.

RQ3 posed the question of the relationship between the predicted booster-uptake levels and future pattern of hospitalisation.

The current analysis is observational, not causal, but the findings showed a significant negative correlation between the predicted level of booster uptake and hospital and ICU occupancy in lower-predicted countries and lower severe-disease burden in higher-predicted countries using the results of the scatter-plot. This coincides with clinical evidence to the effect that booster cover lessens the chances of hospitalisation and ICU admission (Liu et al., 2025). Based on this, the answer to RQ3 too is affirmative, with the predicted booster uptake being negatively correlated with hospital and ICU occupancy implying that the lower the projected uptake, the greater the health-system pressure.

Taken together, the findings affirm that booster uptake is path-dependent, nonlinear in its structure, and significantly correlated with severe-outcome indicators, which offer strong support to all of the three research questions using the existing data.

7.2 Contributions of the Study

This dissertation, academically, provides the reproducible open-data pipeline to model booster demand across countries and removes the lack of interpretability in machine learning methods and the scientific epidemiological outcome data to achieve the prediction of demand (Zhou and Li, 2022; Wood et al., 2025). In practice, the work offers a toolkit ready to implement a policy: both short-term forecasts, feature-importance explanations/visualisations (e.g. SHAP summaries) that decision-makers can relies on to schedule vaccine supply, to mobilize mobile clinics and to focus communication resources whenever demand peaks (Vike et al., 2024). The focus on openness, reproducibility and easy deployment environment makes it less complicated to ministries and partnering agencies who may be interested in running analytics to support their work.

7.3 Implications of Booster Predictions on Hospital and ICU Outcomes

The findings show that the intended rate of booster uptake has a strong negatively related correlation with future occupancy of the hospital and the ICU. Areas or time frames with the minimal suggested booster coverage constantly recorded the highest healthcare burden, which supports the significance of booster campaigns to reduce the severe COVID-19 results. With predicted uptake, there was both an overall decrease of the pressure in the hospital and ICU, which supports that the machine-learning models are informative on the actual health implications of the model.

These results emphasize the importance of predictive modelling as an instrument of forecasting healthcare strain prior to its occurrence. Such forecasts can help policymakers focus their interventions, through targeted communication, mobile vaccine campaigns, and resource allocation, to regions which might have more hospital work in the future. Such predictions can be further reinforced by the use of behavioural or variant-specific data in the future.

7.4 Policy Recommendations

There are three priority policy actions and these follow. To begin with, predictive outputs should be operationalised by ministries of health and international partners to identify areas with low booster uptake and which might coincide with an increase in hospital pressure to keep up with pre-emptive allocation and surge planning (Le et al., 2025). Second, when the model identifies regions with low uptake, but sufficient supply, behaviour and trust problem should be handled with a targeted message and community involvement instead of mere supply increase (Vike et al., 2024). Third, it is crucial to make equity its main focus: predictive targeting should focus on underserved communities to eliminate booster campaigns, and the cost-efficiency evidence must focus on older adults and those at clinical risk (Koiso et al., 2025).

7.5 Future Research

There are three extensions that are suggested. To begin with, go down to subnational and microdata levels to represent the heterogeneity within countries, and then implement hyperlocal interventions, as further down spatial modelling strategies have shown to enhance targeting (Wood et al., 2025). Second, combine the behavioural and sentiment variables between surveys and social media in order to supplement structural covariates that are poorly performing (Zhou and Li, 2022). Third, use automated recalibration and constant monitoring procedures in order that models need to be healthy as variants, as variants, and policies as well as public attitudes change; it is necessary to re-train all models on rolling windows periodically to maintain predictive efficacy and running confidence (Wood et al., 2025).

Operational use is also limited by the limitations of the study: open data sources are indispensable but characterized by reporting delays and national differences in definitions, and until significant changes in the operational framework occur every model output needs to be validated in a region before applying any model outputs, locally (Our World in Data, 2024). To be effectively implemented, implement an easy-to-use dashboard with the total of the predicted demand, stock, and hospital occupancy; introduce the local teams of analysts to the easy decoding of models and SHAP descriptions; use some governance rules to use it transparently and wisely (Vike et al., 2024). Performance will be enhanced by capacity building and regular recalibration after several months to ensure that epidemiology and behaviour change is maintained (Koiso et al., 2025).

References

1. George, G., et al. (2023). ‘Examining the uptake of COVID-19 vaccine booster doses among healthcare workers in South Africa: A mixed-methods study.’ Available at: https://journals.plos.org/globalpublichealth/article?id=10.1371/journal.pgph.0002639
2. Bikaki, A., et al. (2024) ‘Identifying influential factors using machine learning: COVID-19 booster intentions and hesitancy in Texas’, *Archives of Public Health*, 82(1). Available at: https://link.springer.com/article/10.1186/s12982-024-00276-w
3. Zheng, W., et al. (2024) ‘Global landscape of COVID-19 vaccination programmes: stagnation in uptake and policy implications’, *The Lancet Healthy Longevity*, (in press). Available at: https://www.thelancet.com/journals/lanhl/article/PIIS2666-7568(24)00172-7/fulltext
4. Santangelo, O.E., et al. (2024) ‘COVID-19 Vaccination and Public Health: Addressing Global Disparities’, *Frontiers in Public Health*, 12. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11360777/
5. Roy, D.N., et al. (2024) ‘Global disparities in COVID-19 vaccine booster dose (VBD) coverage: drivers and implications’, *Vaccine: X*, 17, 101131. Available at: https://www.sciencedirect.com/science/article/pii/S2590136224000536

6. Wood, A.J., et al. (2025) 'Long-term spatial patterns in COVID-19 booster vaccination uptake: a random-forest approach', *Communications Medicine*, (in press). Available at: https://www.nature.com/articles/s43856-025-00949-w
7. Prasad, V. (2025) 'An Evidence-Based Approach to COVID-19 Vaccination', *New England Journal of Medicine*, 392(12), pp 1351-1354. Available at: https://www.nejm.org/doi/full/10.1056/NEJMsb2506929
8. Espinosa, O., et al. (2024) 'Predictive models for health outcomes due to SARS-CoV-2: a systematic review', *BMC Public Health*, 24, 1234. Available at: https://link.springer.com/article/10.1186/s13643-023-02411-1
9. He, W., Wu, J., et al. (2025) 'Predicting COVID-19 vaccination timing by integrating the theory of planned behaviour and the diffusion of innovations: a survey in Macau', *Journal of Thoracic Disease*, 17(5), pp 2813-2826. Available at: https://jtd.amegroups.org/article/view/100655/html
10. Kazemi, M., et al. (2022) 'Assessing inequities in COVID-19 vaccine roll-out: a cross-country study', *Vaccines*, 10(2), 194. Available at: https://www.mdpi.com/2076-393X/10/2/194
11. Doshi RH, Nsasiirwe S, Dahlke M, et al. COVID-19 Vaccination Coverage — World Health Organization African Region, 2021–2023. MMWR Morb Mortal Wkly Rep 2024;73:307–311. DOI: http://dx.doi.org/10.15585/mmwr.mm7314a3.
12. Katz, M.A., et al. (2023) 'COVID-19 vaccine effectiveness against SARS-CoV-2-confirmed hospitalisation in the eastern part of the WHO European Region (2022–2023): a test-negative case-control study from the EuroSAVE network.' Available at: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(24)00262-X/fulltext
13. Montesi, G., et al. (2024) 'Journal of Translational Medicine.' Available at: https://link.springer.com/article/10.1186/s12967-024-05147-1
14. Madhlopa, Q.K., et al. (2024) 'BMC Infectious Diseases.' Available at: https://www.biomedcentral.com/articles/10.1186/s12879-024-09534-3
15. Magazzino, C., et al. (2022) . 'A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality.'Available at: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/machine-learning-algorithm-to-analyse-the-effects-of-vaccination-on-covid19-mortality/738FDCC83052C8E7A496D12C1297921D
16. Ben-Umeh, K.C., et al. (2024) PLOS ONE. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0298825
17. Vike, N.L., Bari, S., Stefanopoulos, L., Lalvani, S., Kim, B.W., Maglaveras, N., Block, M., Breiter, H.C. & Katsaggelos, A.K. (2024) 'Predicting COVID-19 vaccination uptake using a small and interpretable set of judgment and demographic variables: cross-sectional cognitive science study', JMIR Public Health and Surveillance, 10: e47979. Available at: https://publichealth.jmir.org/2024/1/e47979
18. Ferranna, M., et al. (2024). 'Causes and costs of global COVID-19 vaccine inequity.'Available at: https://link.springer.com/article/10.1007/s00281-023-00998-0
19. Aljedaani, W., Saad, E., Rustam, F., de la Torre Díez, I., & Ashraf, I. (2022) Role of Artificial Intelligence for Analysis of COVID-19 Vaccination-Related Tweets:

Opportunities, Challenges, and Future Trends. Available at: https://doi.org/10.3390/math10173199
20. Cai, M., et al. (2025) 'Association of 2024–2025 Covid-19 Vaccine with Covid-19 Outcomes in U.S. Veterans.' Available at: https://www.nejm.org/doi/full/10.1056/NEJMoa2510226
21. Pesce, M., et al. (2025). 'Global disparities in the introduction, scale-up, and effectiveness evaluation of COVID-19 vaccines.' Available at: https://www.nature.com/articles/s41467-025-63950-w
22. Fox, A.M., et al. (2023). 'Substantial Disparities In COVID-19 Vaccine Uptake And Unmet Immunization Demand In Low- And Middle-Income Countries.'Available at: https://www.healthaffairs.org/doi/10.1377/hlthaff.2023.00729
23. Savas, L.S., et al. (2025). Factors influencing COVID-19 vaccine uptake among vulnerable communities in Texas: Perceptions of 2-1-1 helpline callers. Available at: https://www.sciencedirect.com/science/article/pii/S0264410X25008515
24. Zhou, X., et al. (2022). 'Forecasting the COVID-19 vaccine uptake rate: an infodemiological study in the US.' Available at: https://www.tandfonline.com/doi/full/10.1080/21645515.2021.2017216
25. van Bilsen, C.J.A., et al. (2025). Social networks, environmental, and individual factors associated with COVID-19 vaccination and booster uptake– The prospective PRIME cohort study. Available at: https://www.sciencedirect.com/science/article/pii/S2590136225000208
26. Centers for Disease Control and Prevention (CDC). (2024) *COVID Data Tracker*. Available at: https://covid.cdc.gov/covid-data-tracker
27. Lundberg, S.M. et al. (2020) *From local explanations to global understanding with explainable AI*. Nature Machine Intelligence, 2(1), 56–67. Available at: https://www.nature.com/articles/s42256-019-0138-9
28. Ritchie, H., Mathieu, E. & Roser, M. (2024) *Coronavirus (COVID-19) Vaccinations*. Our World in Data. Available at: https://ourworldindata.org/covid-vaccinations
29. World Health Organization (WHO). (2024) *COVID-19 Vaccination Dashboard*. Available at: https://data.who.int/dashboards/covid19/cases
30. Montesi, G., Augello, M., Polvere, J., Marchetti, G., Medaglini, D. & Ciabattini, A. (2024) 'Predicting humoral responses to primary and booster SARS-CoV-2 mRNA vaccination in people living with HIV: a machine learning approach', Journal of Translational Medicine, 22: 432. Available at: https://doi.org/10.1186/s12967-024-05147-1
31. Zhou, X. & Li, Y. (2022) 'Forecasting the COVID-19 vaccine uptake rate: An infodemiological study in the US', JMIR Public Health and Surveillance, 8(4): e8986207. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8986207/
32. Pedregosa, F. et al. (2011) 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html
33. Lundberg, S.M. & Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Available at: https://arxiv.org/abs/1705.07874

34. Ponce-Bobadilla, A.V. et al. (2024) ‘Practical guide to SHAP analysis: Explaining supervised machine-learning models’, BMC Bioinformatics, 25, 115. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11513550/
35. Feng, H., et al. (2024) ‘Impacts of COVID-19 vaccine boosters on clinical outcomes during Omicron wave’, *[Journal]*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11187233/
36. Lin, DY., et al. (2024) ‘Impact of booster vaccination interval on SARS-CoV-2 outcomes’, *[Journal]*. Available at: https://www.sciencedirect.com/science/article/pii/S1201971224001553
37. Le, T.P. et al. (2025) ‘Health impact and cost-effectiveness of COVID-19 booster vaccination strategies in the early post-Omicron era’, *BMJ Global Health*, 10. Available at: https://gh.bmj.com/content/10/9/e016096
38. Liu, B. et al. (2025) ‘Effectiveness of COVID-19 vaccine boosters for reducing COVID-19 mortality and hospitalisation’, *The Medical Journal of Australia*, 223(8). Available at: https://www.mja.com.au/journal/2025/223/8/effectiveness-covid-19-vaccine-boosters-reducing-covid-19-mortality-among-people
39. Li, L., et al. (2025) ‘Influencing factors and dynamic changes of COVID-19 vaccine uptake’, *Human Vaccines & Immunotherapeutics*. Available at: https://www.tandfonline.com/doi/full/10.1080/21645515.2025.2536898
40. Berber, E. & Ross, T.M. (2024) ‘Factors Predicting COVID-19 Vaccine Effectiveness and Longevity of Humoral Immune Responses’, *Vaccines*, 12(11), 1284. Available at: https://doi.org/10.3390/vaccines12111284
41. Vike, N.L., Bari, S., Stefanopoulos, L., Lalvani, S., Kim, B.W., Maglaveras, N., Block, M., Breiter, H.C. & Katsaggelos, A.K. (2024) ‘Predicting COVID-19 vaccination uptake using a small and interpretable set of judgment and demographic variables: cross-sectional cognitive science study’, *JMIR Public Health and Surveillance*. Available at: https://publichealth.jmir.org/2024/1/e47979/
42. Koiso, S., Hyle, E.P. et al. (2025) ‘Cost-Effectiveness of the 2023–2024 COVID-19 Booster Vaccine’, *JAMA Network Open*. Available at: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2837344
43. Our World in Data (2024) ‘Coronavirus (COVID-19) Vaccinations’ dataset and documentation, Our World in Data. Available at: https://ourworldindata.org/covid-vaccinations