Greyling, Talita; Rossouw, Stephanié

Working Paper

Vaccination uptake, happiness and emotions: using a supervised machine learning approach.

GLO Discussion Paper, No. 1482

Provided in Cooperation with:
Global Labor Organization (GLO)

Vaccination uptake, happiness and emotions: using a supervised machine learning approach.

Talita Greyling[1], Stephanié Rossouw[2]

Abstract

The COVID-19 pandemic is an example of an immense global failure to curb the spread of a pathogen and save lives. To indirectly protect people against a deadly virus, a population needs to achieve herd immunity, which is attained either through vaccination or prior infection. However, achieving herd immunity by vaccination is preferable as it limits the health risks of disease. As the coronavirus mutated, vaccination estimates for achieving herd immunity went from 70% to 90%. In this study, we investigate the order of the importance of the variables to identify those factors that contribute most to achieving high vaccination rates. Secondly, we consider if subjective measures, including the level of happiness and different collective emotions of populations, contribute to higher vaccine uptake. We employ an XGBoost machine learning model (and, as robustness tests, Random Forest and Decision Tree models) to train our data. Our target output variable is the number of people vaccinated as a percentage of the population. We consider two thresholds of our output variable, the first at 70% of a country's population, corresponding to the initial suggestions to achieve herd immunity, and the second with a threshold of 90%, suggested later due to the highly infectious virus. We use a dataset that includes ten countries in the Northern and Southern Hemisphere and variables related to COVID-19, vaccines, country characteristics and the level of happiness and collective emotions within countries. The most important variables listed in reaching the 70% and 90% thresholds are similar. These include the implemented vaccination policy, international travel controls, the percentage of the population in rural areas, the average temperature, and the happiness levels within countries. It is remarkable how the importance of subjective measures of people's emotions and moods play a role in attaining higher vaccination levels. As the vaccine threshold increases, the importance of subjective well-being variables rises. Therefore, not only the implemented policies and country characteristics but also the happiness levels and emotions play a role in compliance and achieving higher vaccination thresholds. Our results provide actionable policy insights to increase vaccination rates. Additionally, we highlight the importance of subjective measures such as happiness and collective emotions to increase vaccination rates and assist governments to be better prepared for the next global pandemic.

JEL codes: C55; I10; I31; H12; N40

Keywords: COVID-19; vaccine; happiness; emotions, supervised machine learning

[1] College of Business and Economics, University of Johannesburg, Gauteng, South Africa, Honorary Adjunct Academic, School of Social Sciences & Humanities, Auckland University of Technology, New Zealand, Research Fellow with the Global Labor Organization, ORCID: 0000-0002-3777-7003.

[2] Corresponding author: School of Social Sciences & Humanities, Faculty of Culture and Society, Auckland University of Technology, Auckland, New Zealand, Senior Research Associate, College of Business and Economics, University of Johannesburg, Johannesburg, South Africa, Research Fellow with the Global Labor Organization, ORCID: 0000-0003-3538-9215.

1. Introduction

The COVID-19 pandemic is an example of an immense global and national failure to curb the spread of the virus and save lives. The sheer magnitude of this failure becomes evident when we consider the death toll due to COVID-19. As of 20 July 2023, the World Health Organisation (WHO) (2023) reported that there had been a total of 768,237,788 confirmed cases of COVID-19, including 6,951,677 deaths. Europe has been the hardest hit region, with 2,245,217 deaths, and Africa has the least recorded deaths, with 175,408 deaths (Africa faces doubts regarding the accuracy of its data). The enormity of this death toll (lagging only behind the Spanish flu and HIV/AIDS) and the economic damage to countries, industries and individuals are unmeasurable (Baldwin, 2020; Ludvigson et al., 2020; Lu et al., 2020; Fetzer et al., 2020).

Furthermore, COVID-19 not only affected health but also had a profound impact on family functioning and well-being. For example, New Zealand found a significant increase in family violence reports to police, which ranged from 345 to 645 a day, compared to 271 to 478 a day in the same period in 2019 (Mental Health and Wellbeing Commission, 2023). Andrade et al. (2022) note that the fear and uncertainty of health risks, the stress from restrictions and constraints on everyday life, and financial concerns impacted emotional well-being.

During a pandemic, the aim is to stop the spread of the disease and protect individuals against a specific pathogen. We know that globalisation, the geography of economic relations, and international travel pose significant challenges in stopping the spread of a virus. A population must achieve herd immunity to protect people from the disease indirectly. Herd immunity is achieved when a population is immune through vaccination or immunity developed through previous infection. However, the World Health Organisation (WHO) supports achieving herd immunity through vaccination rather than exposing them to the pathogen. To safely achieve herd immunity against COVID-19, it was estimated at the early stages of the pandemic that a vaccination threshold of 70% should be achieved (Randolph & Barreiro, 2020; Bartsch et al., 2020; Goldblatt et al., 2022). However, as COVID-19 evolved, the virus mutated and became more infectious, and the estimated vaccination threshold increased to 90% (Plans-Rubió, 2022). According to Bloom et al. (2021), high vaccination uptake yields sizable and diverse health, economic, and social benefits, including herd protection, increased work hours and productivity, and potentially improved social equity. In other words, the faster the uptake, the fewer lives are lost, and the potentially devastating economic and social impact is minimised.

As of 22 July 2023, a total of 13,474,265,907 vaccine doses have been administered. This translates into 64.8% of the world population being fully vaccinated[3] (WHO, 2023). However, when we disaggregate the data, we see the stark inequality between high-income countries, 74.32%, and low-income countries, 27.54% (Mathieu et al., 2021). Despite global partnerships like COVAX, these low vaccination rates in developing and underdeveloped countries highlight the lack of international support and cooperation. As Sheikh et al. (2021) noted, most developing nations lack the financial and technological resources to invest in vaccine development. Therefore, relying on developed nations through global cooperation was instrumental in vaccinating their people. Unfortunately, in a shameful show of 'individuality', developed nations, constituting only 16% of the world population, bought more than half of the vaccines available at the start of 2021. This glaring absence of international support and cooperation is seen as one of the biggest failures of the COVID-19 pandemic.

Greyling and Rossouw (2022) also argue that this immense failure is partly due to the inability at a global and national level to distribute and administer vaccines efficiently. Furthermore, at the national level, governments and the public health care systems not only failed at stopping the spread of the virus and protecting human lives but also failed to adhere to basic norms of institutional rationality and transparency, breeding mistrust in governments (Paul et al., 2021; Sallam, 2021).

Considering the abovementioned, our primary aim is to retrospectively evaluate the COVID-19 pandemic and determine the most important factors to reach vaccination thresholds. Therefore, we will determine the most important factors for achieving herd immunity at the 70% vaccination threshold, estimated at the beginning of the COVID-19 pandemic and the 90% vaccination threshold, as estimated later in the pandemic. A secondary aim lies in determining those factors that differ between the 70% to 90% vaccination threshold to see which factors are responsible for advancing a population's decision to reach the higher vaccination level. Special consideration will be given to whether subjective well-being measures played a role in the decision to be vaccinated since we know that negative emotions, such as fear of the side effects of vaccines, influence peoples' attitudes towards receiving the vaccine (Greyling & Rossouw, 2022) and that happier people make better health-related decisions (Anik et al., 2009; Lyubomirsky et al., 2005).

To achieve the aforementioned, we use data from four datasets. The first dataset is extracted from Google COVID-19 Open Data[4]. It provides us with abundant information related to COVID-19 and

---

[3] Total number of people who received all doses prescribed by the initial vaccination protocol, divided by the total population of the country.

[4] Available from https://health.google.com/covid-19/open-data/explorer

information on population, geographical location, the economy, general health and climate. The other three time series datasets are derived from tweets and form part of the *Gross National Happiness.today* project[5]. These three unique datasets reflect i) the general sentiment and emotions within countries, ii) the sentiment and emotions towards vaccines and iii) the sentiment and emotions towards government institutions.

We use an eXtreme Gradient Boosting (XGBoost) model to determine the most important factors that can predict reaching vaccination thresholds. We chose the XGBoost model since it is more efficient, computationally much lighter and has been shown to outperform most supervised algorithms (Abdurrahim et al., 2020; Nielsen, 2016). However, we construct two other models using Random Forest and Decision Tree as robustness tests. To evaluate the accuracy of our models, we use the mean squared error (MSE), mean absolute error (MAE) and root mean square error (RMSE). The XGBoost outperforms random forest and decision tree predictions in line with expectations. Consequently, we discuss the results of the XGBoost model. However, we also present the results of the other models in Supplementary Information C.

Our results on the importance of the factors that increase vaccine uptake at a 70% threshold and 90% threshold overlap with the following factors: vaccination policy implemented, international travel controls, the percentage of the population in rural areas and the average temperature. Interestingly, we find that the importance of happiness differs between the two thresholds. Happiness is less important in achieving the 70% threshold. However, to increase the threshold to 90%, the importance of happiness cannot be ignored. The results clearly show that if governments want higher levels of compliance and vaccine uptake, subjective well-being measures such as mood and emotions must be prioritised. Addressing how people feel, in general, towards vaccines and governments is vitally important when policymakers want to push beyond the lower 70% vaccine threshold and achieve the "golden standard" of 90% fully vaccinated.

Our study makes several contributions to the existing literature. First, this is the first study conducting a post-COVID-19 cross-country analysis of the most important variables to increase vaccine uptake. Second, we are the first study to include subjective measures of well-being in our estimations, such as happiness levels, people's emotions and their perceptions towards vaccines and governments, to establish whether subjective measures play a role in increasing vaccination uptake. Third, we are the first to apply supervised machine learning models to determine which factors matter most to achieve

[5] Available from https://gnh.today/

different vaccination thresholds (please note that our dependent variable is continuous, thus different to models in which a binary, mostly a "yes-no" response, is used). Our XGBoost model can be used as a benchmark for future research related to the most important factors for increasing vaccination uptake. Furthermore, this study offers some actionable insights for policymakers on increasing vaccination rates to curb pandemics' health and economic and political effects.

The rest of the paper is structured as follows. The next section contains a literature review of studies investigating factors influencing COVID-19 vaccination rates. Section 3 describes the data and the selected variables, while section 4 outlines the methodology. The results and discussion follow in sections 5 and 6, while the paper concludes in section 7.

2. Literature review

Since increasing the uptake of the COVID-19 vaccine was fundamentally important to decrease the harm caused to human lives and livelihoods, many studies have focused on predicting factors associated with the uptake. However, few studies have used machine learning to determine the factors contributing to higher vaccine uptake. Therefore, the literature review mainly discusses studies that rely on survey data and traditional empirical analysis, which also informs our discussion. Studies that used machine learning in their approach conclude this section.

2.1 Factors associated with vaccination uptake: Evidence from survey data

Regarding individual European country studies, Bajos et al. (2022) and Ward et al. (2020) focused on France and used data from the EpiCov survey and self-collected data, respectively. Similarly, Gomes et al. (2022) conducted a study in Portugal using a community-based survey called the COVID-19 Barometer: Social Opinion. These three studies generally concluded that the COVID-19 vaccine uptake was positively associated with age, educational attainment and income. According to Bajos et al. (2022), the least educated, those with the lowest incomes, and racial minority groups were less likely to accept the vaccine, and these differences were maintained or increased over time. Additionally, people's lack of trust in the government and scientists to manage the health crisis remained the primary reason for refusing to vaccinate. Ward et al.'s (2020) pre-vaccine study also found that individuals feeling close to a Far-Right party would refuse the vaccine when it became available. The primary reason any individual would refuse the vaccine was that it would not be safe. Gomes et al. (2022) also concluded that higher odds of hesitancy were associated with low confidence in Portugal's health services response to COVID-19 and non-COVID-19 and perceived the measures implemented by the government as inadequate.

In cross-country analysis, Bergmann et al. (2022) and Pronkina and Rees (2022) used the 2021 summer SHARE Corona survey data (administered across 27 European countries). They confirmed the results of Bajos et al. (2022), Ward et al. (2020) and Gomes et al. (2022) by finding that the probability of being vaccinated increased with age, income, and educational attainment. Furthermore, Bergmann et al. (2022) concluded that prior illnesses were associated with a higher willingness to vaccinate. Interestingly, there was no clear and significant effect of subjective health and no strong effects with mental health issues were found. Pronkina and Rees (2022) argued that people who express trust in others are more likely to be vaccinated, while risk aversion and frequency of praying (a proxy for religiosity) were negatively correlated with the probability of being vaccinated against COVID-19. Furthermore, Europeans aged 50 and older did not base their decision to vaccinate against COVID-19 on case counts or excess mortality during the pandemic.

Turning to the American context, Corcoran et al. (2021), Czeisler et al. (2021), El-Mohandes et al. (2021), and Gatwood et al. (2021) found that Americans who express conservative political or religious beliefs are, on average, more vaccine-hesitant than those who do not. However, the relationship between political beliefs and COVID-19 vaccination hesitancy appears to be considerably more nuanced in Europe than it is in the United States (Ward et al., 2020; Lindholt et al., 2021; Raciborski et al., 2021; Bíró-Nagy & Szászi, 2022; Wollebæk et al., 2022). COVID-19 vaccine hesitancy is especially prevalent among individuals who express distrust in the government and scientists (Kerr et al., 2021; Latkin et al., 2021; Lindholt et al., 2021; Rozek et al., 2021; Bajos et al., 2022).

2.2 Factors associated with vaccination uptake: Evidence from machine learning

In terms of previous machine learning studies, Lincoln et al. (2022) used Random Forest to probe for the optimum prediction accuracy for vaccine hesitancy and to find an economical model based on a selection of common global predictors. They used SHapley Additive exPlanations (SHAP) and permutation feature importance to estimate the importance of each variable in their model across their sample of five advanced countries (UK, USA, Australia, Germany and Hong Kong). The authors found that by using only twelve variables (the combined most important variables from permutation feature importance and SHAP), they could achieve an 82% accuracy in predicting vaccine hesitancy, with the most crucial factors being vaccination conspiracy beliefs and a lack of confidence in governments, companies, and organisations in handling the pandemic (i.e., pandemic conspiracy beliefs).

Previous studies have successfully used XGBoost-based predictive models to predict influenza vaccine uptake. Shaham et al. (2020) used primary data from 250,000 Israelis collected between 2007 and 2017 to predict whether a patient would get vaccinated in the future. Their XGBoost-based predictive model achieved an ROC-AUC[6] score of 0.91 with accuracy and recall rates of 90% on the test set. Prediction relied mainly on the patient's individual and household vaccination status in the past, age, number of encounters with the healthcare system, number of prescribed medications, and indicators of chronic illnesses. Using the XGBoost regressor, Cheong et al. (2021) used sociodemographic data to predict vaccine uptake across counties in the United States (US). Their model predicted COVID-19 vaccination uptake across US counties with 62% accuracy. The results from their permutation analysis and SHAP revealed the most important factors to drive their predictive model were geographic location (longitude, latitude), education level (per cent of adults with less than a high school diploma, per cent of adults with a bachelor's or higher), and online access (households with broadband internet).

Also focusing on the US, Osman and Sabit (2022) use state-level vaccination rates to identify the most important features that predict which states will meet the vaccination threshold of 70%. Relying on a Chi-square Automatic Interaction Detector (CHAID), a decision tree algorithm, the authors include several variables that may influence the state-specific vaccination rate. They categorise the variables into four groups: economic indicators, COVID-19-related indicators, Google mobility data, and COVID-19-related policy measures. After using three different model specifications, they discovered that workplace travel, the political affiliation of the governor, and the vaccine mandate in schools were the top three features of achieving the vaccination threshold.

In the abovementioned studies on machine learning applications, the outcome variables were binary variables, for example – a person's decision to be vaccinated or whether a certain vaccination threshold would be reached. These studies determined the most important factors for reaching success (yes) during COVID-19. Our study differs from the previous literature in that we benefit from hindsight. Therefore, we investigate the most important factors contributing to reaching herd immunity (at different levels of 70% or 90%) and how these factors change when higher herd immunity levels are to be reached. Our outcome variable is the percentage of the population vaccinated as a percentage of a country's population (thus, the measure used to determine herd immunity). It is a continuous variable representing a high level of variance and is not restricted to only a yes or no answer. Furthermore, our study includes a wide-reaching dataset including variables related to COVID-

[6] Area Under the Curve of the Receiver Operating Characteristic curve.

19 regulations, vaccination policies, country characteristics, and, very importantly, subjective measures of well-being. We are the first study to include subjective well-being measures to highlight the importance of moods and emotions when higher vaccination thresholds must be attained.

3. Data and Variables

3.1 Construction of Datasets

The timeframe under consideration is from 1 December 2020 to 16 September 2022. This period includes the first vaccine rollout and ends when new COVID-19 tests reach almost zero in all countries. Consequently, the main data source related to COVID-19, the *COVID-19 Government Response Tracker* dataset (Mathieu et al., 2021), was discontinued on 31 December 2022. We consider the data to find a retrospective view of those factors that mattered most for higher vaccination rates.

We use a merged dataset, including the Google COVID-19 Open Data[7] and our three constructed time-series datasets derived from tweets[8]. The three Twitter datasets reflect i) happiness levels and emotions of countries, ii) happiness levels and emotions towards vaccines and iii) happiness levels and emotions towards government institutions. The construction and validation of the Twitter datasets are explained in Supplementary Information A.

This section briefly explains the Twitter data with a more detailed explanation available in Supplementary Information A. Tweets are extracted in real-time based on a geographic bounding box corresponding to the country in question. Next, we use sentiment and emotion analysis to score the tweets. We aggregate the scores and derive indices for happiness and each of the eight emotions. For the Twitter datasets related to the government and COVID-19 vaccines, we used specific keywords to identify those tweets directly related to the topic.

To derive the dataset related to the COVID-19 vaccines, we extracted tweets using the keywords: *vaccinate, vacc, vaccine, Sputnik V, Sputnik, Sinopharm, Astrazeneca, Pfizer (if NEAR) vaccine, Pfizer-BioNTech, Johnson & Johnson,* and *Moderna*.

For the dataset related to governments, we extracted tweets using the keywords: *government, parliament, ministry, minister, senator, MPs, legislator, political, politics, prime minister.*

[7] Available from https://health.google.com/covid-19/open-data/explorer
[8] Available from https://gnh.today/

After extraction, we analysed the text of the tweets to determine the noise captured in the tweets. Subsequently, we found that the noise was minimal in both instances.

The Google COVID-19 Open dataset is rich and includes variables related to COVID-19 cases, deaths, vaccinations, demographic, economic, geographical, climate, health, health infrastructure and health care.

3.2 Data cleaning and validation

After merging the datasets outlined in Section 3.1, we had a total of 145 variables.

As a first instance, we set out to identify missing data. If the data was randomly missing with less than 3% overall missingness, we imputed the data by either using the mean or the previous data point as appropriate. Secondly, we dropped variables from our dataset with high missingness levels. For example, international support (67% missingness), emergency investment in health care (68% missingness) and mobility regulations (74% missingness), which reflects the strong regulations implemented during the first lockdowns in countries, such as access to retail and recreation, grocery stores, pharmacies and parks, were dropped. Thirdly, we removed highly correlated data so that only one of the variables remained in the dataset, for example, cumulative confirmed cases and cumulative tested cases; this eases the interpretation of the results.

Once the data was cleaned, we were left with 69 variables (including our outcome variable), which we classified into five categories (refer to section 3.4). Subsequently, these variables were used in the supervised algorithms (refer to section 4.1) to train the models. We have 6530 observations, which means we have 653 (just short of two years) observations per country in our sample.

In our study, the data comprising 69 variables are split randomly into a training and testing dataset with an 80:20 split on all data, with the evaluation done on the unseen testing data.

3.3 Target/outcome variable

Our primary variable of interest is the country-level vaccination rate. We calculate vaccination rates as the percentage of the vaccinated population as a percentage of the total population in the respective countries. This is in line with studies such as Randolph and Barreiro (2020), Bartsch et al. (2020) and Goldblatt et al. (2022).

Table 1. Maximum vaccination rates on 16 September 2022.

Country	Percentage of the population vaccinated on 16 September 2022
Australia	85.35
Belgium	76.19
Germany	76,43
Spain	86.58
France	80.07
Great Britain	76.15
Italy	79.46
The Netherlands	69.19
New Zealand	85.67
South Africa	32.64

Source: Authors' own calculations

In our sample, nine out of the ten countries met the lower threshold of 70% (see Table 1); South Africa lagged behind, reaching a mere 32.6%. Therefore, our 70% threshold model was reachable for the countries in the developed world but not for our developing country, South Africa (likely to be the same in other developing and underdeveloped countries). However, none of the countries in our sample achieved the higher 90% threshold, with Spain coming closest with 87%.

3.4 Predictor variables/features

As discussed in Section 3.2, our models include 68 features (independent variables) (apart from our outcome variable) to determine those factors most important for the vaccination thresholds. We remind the reader that two variables, international support and emergency investment in health care, were not included as predictors in our models due to their high levels of missingness, 67% and 74%, respectively.

We acknowledge that these variables could have ranked among the most important variables and potentially have been included in the top ten. Therefore, when we report the results of our models, their absence should be kept in mind.

We categorise the variables into five groups: demographic, geographical, economic, COVID-19-related indicators and COVID-19-related policy measures. The COVID-19-related and policy data are high-frequency daily data, while the demographic, geographical and economic data are more stable over time. Table 2 gives an abbreviated list of the variables included in the models. For a full list, see Supplementary Information B.

Table 2. An example of variables used.

Variable	Description	Scale	Coding	Source
Vaccination policy	Policies for vaccine delivery for different groups	Ordinal scale	0 - No availability 1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability	Mathieu et al. (2021)
Average temperature	Average temperature in the country	Celsius		World Bank (2023a)
Population density	People per square kilometre of land area			World Bank (2023b)
Restrictions on gatherings	Record limits on gatherings	Ordinal	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data	Mathieu et al. (2021)
GNH	Happiness	Ordinal	Score per hour ranges from 0 to 10, with higher values indicating higher happiness. To generate daily data, the mean GNH per day is calculated.	Greyling et al. (2019)

4. Methodology

The methodology first explains XGBoost (our model of choice). Next, we discuss the metrics used to evaluate the performance of the models.

4.1 eXtreme Gradient Boosting (XGBoost)

To determine the most important factors in achieving our vaccination thresholds of 70% and 90%, we use XGBoost.

XGBoost is a highly efficient and scalable machine learning algorithm implementing gradient boosting for decision trees. XGBoost is based on the gradient boosting framework, where models are built sequentially, and each new model corrects the errors of the previous one. This process continues until a strong predictive model is formed. It is designed for speed and performance and uses optimisation techniques that support parallel and distributed computing, which makes it highly scalable to large datasets. Furthermore, it includes regularisation (L1 and L2) to prevent overfitting.

XGBoost has demonstrated greater accuracy than other methods. For example, Abdurrahim et al. (2020), comparing the accuracy of different predictive modelling algorithms, shows that XGBoost shows the highest accuracy score compared to other methods such as logistic regression, naive Bayes classifier, Decision Trees, and Random Forest. However, our study uses Random Forest and Decision Trees to test the robustness of our results.

Multiple combinations of the parameters of the XGBoost model were tested. A tree depth of seven delivered optimal results. Our XGBoost model is defined in equation (1) as:

$$F_M(x) = F_0 + v\beta_1 T_1(x) + v\beta_2 T_2(x) + \cdots + v\beta_M T_M(x) \quad (1)$$

Where M is the number of iterations. The gradient boosting model is a weighted $(B_1 \ldots \beta_M)$ linear combination of simple models $(T_1 \ldots T_M)$. $F_M(x)$ is the vaccination threshold as described in section 3.3.

We first used the default settings of the XGBoost algorithm on the training data and refined the parameters to find the best fit. We started by refining the depth of the trees and tested depths between three and ten, finding seven, which resulted in the lowest root mean square error (RMSE) (see section 4.2). We set the number of iterations to 100, with a termination clause added to stop the algorithm if the RMSE does not decrease after five iterations. After completing the refining stage, the model reached the lowest RMSE after 16 iterations, ensuring we selected the most effective parameters for our analysis.

4.2 Evaluation

Model evaluation uses metrics to analyse the model's performance and, thus, how well the model generalises future predictions. Machine learning metrics include Accuracy, Precision, Recall and F1 score if the outcome variable is binary. However, as we have a continuous outcome variable, we make use of the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Root Mean Square Error (RMSE).

5. Results

In this section, we first discuss the results after training the models and the fit of the models to the test data. Second, we discuss the results of the models to answer our research questions.

5.1 Results of training and testing the fit of the models

Figure 1 shows the RMSE over iterations for XGBoost. The RMSE decreases over the number of iterations to reach a minimum at 16 and remains constant up to 20 iterations.

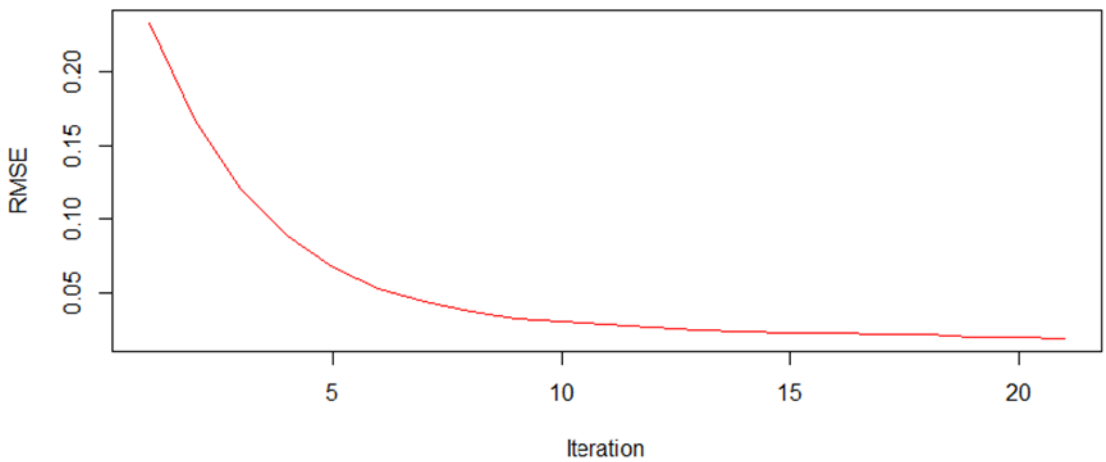

Figure 1. RMSE over iterations for XGBoost.

The Random Forest model took much longer to train compared to the effectiveness of the training of the XGBoost model. After 50 iterations, it seemed the model converged, but upon further inspection, the results improved with minute increments with each additional iteration. Figure 2 shows the MSE decreases; after 50 iterations, the MSE is relatively small. The MSE becomes smaller with each iteration but does not converge to a specific value.

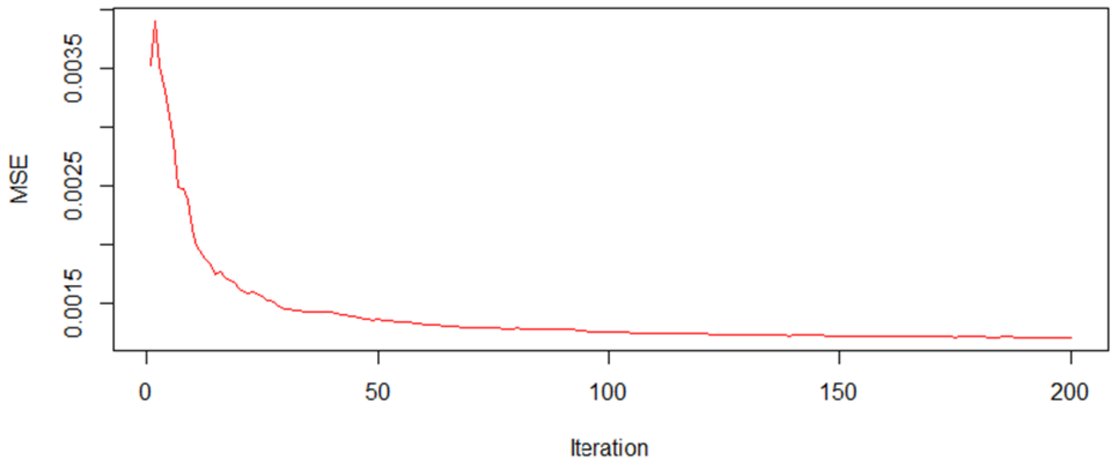

Figure 2. MSE over iterations for Random Forest.

Table 3 gives the fit statistics for the three models, explaining how well our models predict the outcome variable of our test dataset. We discuss the fit measures to predict a 90% threshold since this provides us with the largest possible test dataset (the fit measures are also available for the 70% level). We notice that all measures of fit reveal very small errors, indicating good-fitting models. Across all three of the fit statistics, the XGBoost performs the best with the lowest values. For the XGBoost, the MSE is 0.0014, the MAE is 0.0227, and the RMSE is 0.0375.

Table 3. Evaluation metrics across models.

Model	MSE	MAE	RMSE
XGBoost	0.001412552	0.022707714	0.0375839
Random Forest	0.001861686	0.029981258	0.043147264
Decision Tree	0.01222601	0.07180425	0.11057130

Source: Author's own calculations

Though the fit statistics indicate that the XGBoost model performed best when considering all models, a visual representation of all three models is also provided. In Figure 3, the true value of the dependent variable is represented in red, while the predictions for the three models are represented in blue, XGBoost, green, Random Forest, and magenta, the Decision Tree. Figure 3 supports the results in Table 3, as the XGBoost predictions (blue line) are consistently closer to the true value (red line) than those using the other two models. This aligns with our expectations that the XGBoost model outperforms the other models.

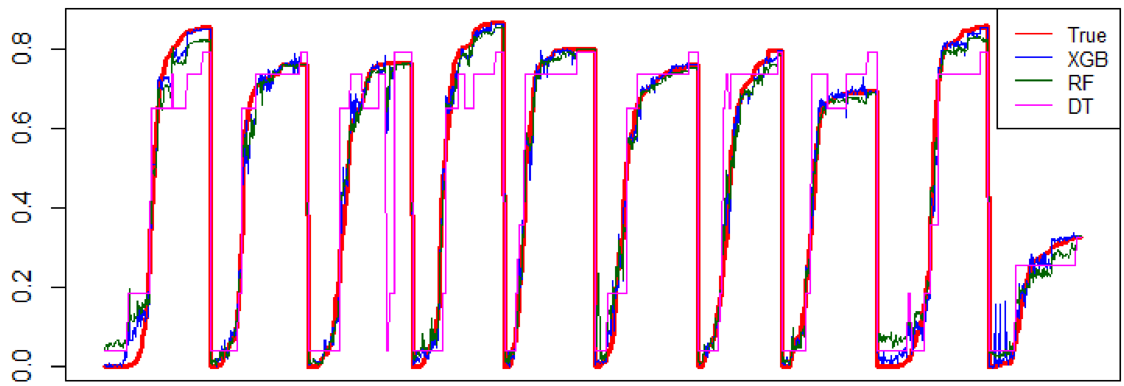

Figure 3. True value with all model predictions.

As mentioned previously, the XGBoost model performs better and uses less computational power. Therefore, in discussing the application of the model to answer our research question, we interpret the XGBoost results[9].

5.2 Results of the XGBoost model on variable importance

Table 4 shows the results from our XGBoost model on ranking the importance of variables to reach a 70% and 90% vaccination threshold rate (see Supplementary Information C for the Random Forest and Decision Tree).

Table 4. Results on the order of the importance of the variables predicting vaccination thresholds of 70 and 90%, respectively.

70% threshold	90% threshold
Vaccination policy	Vaccination policy
Population aged between 10-19	International travel controls
International travel controls	Percentage of population in rural areas
Percentage of population in rural areas	**Happiness**
Average temperature	Average temperature
Workplace closing	Population density
Restrictions on gatherings	Human Development Index
Life expectancy	Facial coverings
Happiness	Workplace closing
Pollution mortality rate	Restrictions on gatherings

[9] The reader should note that although the XGBoost outperforms the other models and is computationally less expensive, the Random Forest and Decision Tree Models have the benefit that they are easier to understand and visualise.

Considering the results from reaching the 70 and 90% thresholds, we notice recurring factors among the five most important factors. The factors are related to the vaccination policies, the COVID-19 policies to limit the spread of the virus, and country characteristics such as the percentage of the population residing in rural areas and the average temperature in the countries. This implies that regardless of the vaccination threshold goal, the vaccination policy, policies related to international travel controls, the percentage of the population in rural areas, and the average temperature are important to achieve maximum vaccination rates.

It's worth highlighting the significant role that subjective well-being measures play in attaining vaccination goals. To gain a 70% vaccination (all countries met this threshold except SA), happiness was among the top ten important factors at number nine (Figure 4). However, to reach the vaccination threshold of 90% or more, we notice that people's happiness has become increasingly important and has reached fourth place (Figure 5). Therefore, regardless of the threshold level, happiness plays an important role, and the higher the vaccination threshold governments want to achieve, the more important it becomes.

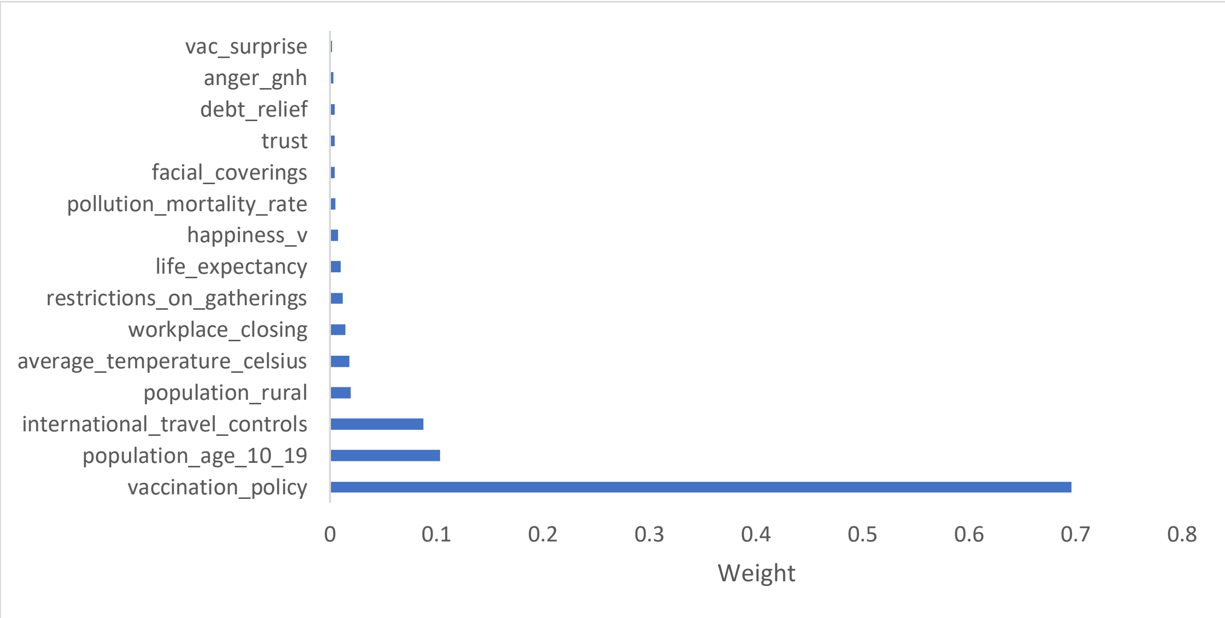

Figure 4. Ranked variable importance - 70% vaccination threshold.

If we only consider the lowest threshold of 70% vaccination (Figure 4), most factors are objective and similar to the ones mentioned before. However, the share of the younger population also seems to be relatively important. From our sample, we note that all except one country managed to reach the 70% threshold, and therefore, more attention should be paid to those factors from the 90% threshold models.

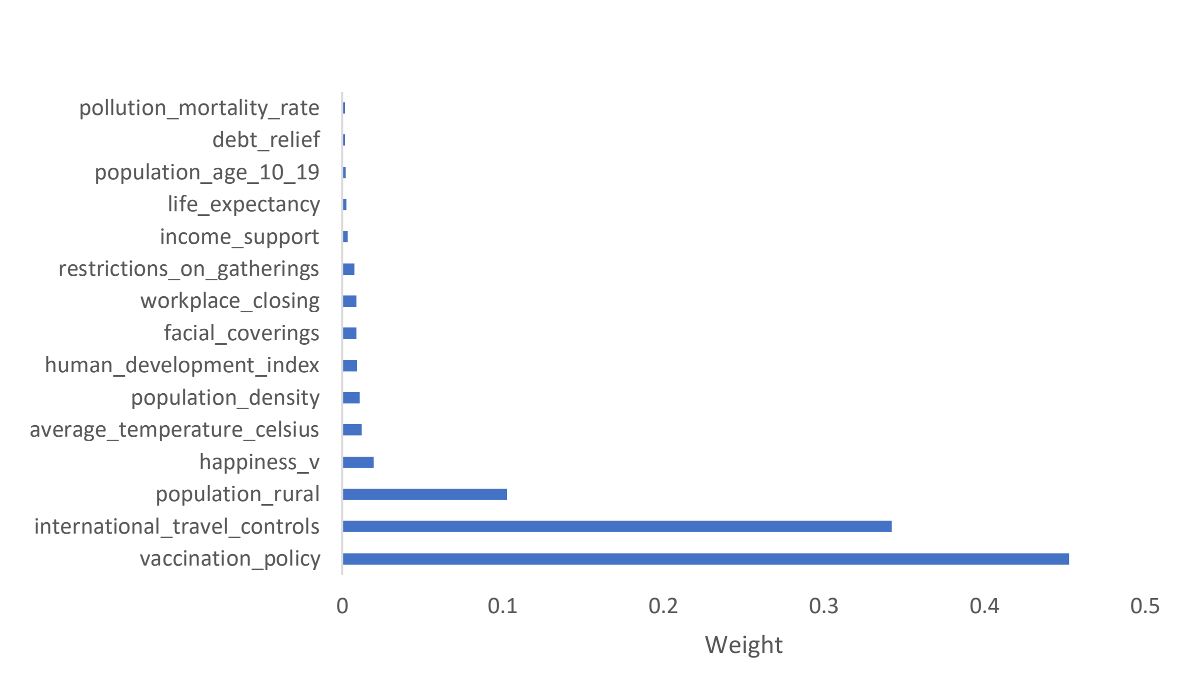

Figure 5. Ranked variable importance - 90% vaccination threshold.

6. Discussion on the application

We will focus our discussion on the top 5 factors and use information from previous studies (see Section 2) to allude to the relationship with vaccination thresholds. Since we know from Plans-Rubió (2022) that more than 90% of a country's population would need to be vaccinated, given the infectiousness of the pathogen, to achieve herd immunity, our discussion will focus on achieving this "golden standard". Subsequent discussions will highlight where factors have significantly changed in ranking and discuss how happiness and collective emotions can increase vaccination rates. As far as we know, this is the first study that shows the importance of subjective well-being measures.

As noted in Section 5.2, regardless of the vaccination threshold goal, governments should focus on their vaccination policy, international travel controls, the percentage of the population in rural areas and the average temperature to achieve their maximum vaccination rates (see Figures 4 and 5).

The vaccination policy implemented (groups that can access the COVID-19 vaccine) was shown by Greyling and Rossouw (2022) that when more groups of people can access the vaccine, for example, all age groups compared to fewer groups, it is positively related to attitude towards the vaccine. This means more people will be vaccinated when more people can access the COVID-19 vaccine.

Regarding international travel controls, we know that, for example, in New Zealand (one of the countries with the most stringent lockdowns and highest number of lockdowns), people were told to get vaccinated if they wanted their freedoms back. The then Prime Minister, Jacinda Ardern, clearly stated, "If you want summer […] get vaccinated." If you don't, "there will be everyday things you will miss out on". It wasn't until September 2022 that New Zealand fully opened their international borders, allowing visitors. Rossouw et al. (2021) found that international border controls acted as a dual shock, economic and social. Hospitality operators were impacted directly by the lack of international and domestic tourism and experienced a significant economic shock that negatively influenced their livelihoods. Furthermore, being unable to travel the world is a social shock, causing a decrease in happiness.

When it comes to the population percentage in rural areas, Barbieri et al. (2022) and Polašek et al. (2022) show that vaccine hesitancy is significantly higher in the rural than in the urban population. Additionally, De Boeck et al. (2020) and Oli et al. (2017) found that the complexity of the pipeline for vaccines from the regional depot to the facility level may create breaking points due to inadequate infrastructure and skills gap and that travelling to rural health facilities is more difficult than to urban health facilities. Rural populations, vulnerable and excluded people are among those for whom improved vaccination rates and access to care were urgently needed to prevent and treat COVID-19. Therefore, governments need to ensure that the rural populations receive targeted information related to the safety of vaccines and that the rural population's access to vaccines is not hampered by procurement and capacity issues.

This study is the first to show the importance of subjective well-being in achieving vaccination thresholds. Concerning the vaccination threshold of 90%, happiness ranks fourth (and ninth in the 70% model) and is therefore important for governments to address. Measuring happiness, a subjective measure that captures people's evaluative mood, is very important in any decision-making process. In an ideal world, people make rational choices. The rational choice theory states that when humans are presented with various options under the conditions of scarcity, they will choose the option that maximises their individual satisfaction. However, humans are not rational, and their emotions drive them; therefore, they make irrational decisions. Therefore, emotions and happiness levels also drive decision-making processes when deciding whether to vaccinate. Additionally, previous studies such as Kim et al. (2015) show that happier people make better health-related decisions since happier people are less inclined to engage in high-risk activities and take preventative action to mitigate risk. Also, happy people are not just self-centred or selfish; the literature suggests that happy individuals tend to be relatively more cooperative, prosocial, charitable, and "other-centred" (Kasser & Ryan, 1996; Williams & Shiaw, 1999).

Furthermore, Sarracino et al. (2024) showed that happiness and trust are positively correlated, meaning that as trust increases, so does happiness. Trust in others also promotes cooperation and solidarity with positive spillovers on compliance and well-being (Bargain & Aminjonov, 2020). The takeaway from trust and happiness is quite straightforward: the lower your vaccination rates, the more important people's levels of happiness and trust become. Happiness and trust are connected to compliance and doing something "for the greater good". Therefore, the more you want people to engage in a specific activity, such as getting vaccinated, the more important emotions and happiness levels become.

Average temperature ranks fifth in importance in our threshold models. Jansson and Yamamoto (2022) studied five states in the US to determine the relationship between average temperature, the level of humidity and COVID-19 infection rates. The authors found that a higher-than-average temperature was consistently associated with a decreased relative risk of infection. Given that Fieselmann et al. (2022) found that one of the main reasons people do not get vaccinated is a perceived lower risk of infection, we can deduce that higher-than-average temperatures could lead to countries not meeting their maximum number of vaccine dosage uptake as a proportion of the population size of a country. Apart from the above, we know from studies conducted by Streefland et al. (1999a and b) that in developing countries, parents who do not adhere to vaccination schedules often do so because they are unable to go due to climatic conditions such as the weather being too hot, or roads being flooded from significant rainfall, or a crop needs to be harvested before it withers in the heat. However, we note that the vaccine rollout was hampered in several European countries as well as the US as severe snowstorms and unusual cold fronts caused inoculation centres, including mega facilities capable of vaccinating up to 20,000 people a day, to close (The Guardian, 2021; CBC News, 2021; John Hopkins Healthcare, 2021).

A factor rated among the top 5 in our 70% threshold models that did not appear in the 90% threshold model is the population aged between 10-19.

As the percentage of the population between 10-19 decreases, the population rate increases since they were last to be vaccinated. Therefore, if only a small proportion were this age, more people would be allowed, according to vaccine policy, to get vaccinated, and the vaccination rate would increase.

For example, for all developed countries in the sample groups, people between 10 and 19 were 12% or less of the population – whereas in South Africa, it was almost 18%. This indicates many things – also, Western countries' populations are getting older – thus, there is a higher need to vaccinate the larger older population.

7. Conclusion

In this paper, we employed supervised machine learning using XGBoost to retrospectively evaluate the COVID-19 pandemic and determine the factors most important in increasing vaccine uptake. Therefore, we determined those factors associated with achieving herd immunity at the 70% vaccination threshold, estimated at the beginning of the COVID-19 pandemic and the 90% vaccination threshold, estimated later in the pandemic. By doing the aforementioned, we also determined those factors that differed between the 70% to 90% vaccination threshold, which were responsible for reaching the higher vaccination level. Throughout our analyses, we paid special attention to the role of subjective well-being measures in achieving vaccine thresholds since we know that negative emotions, such as fear of the side effects of vaccines, influence peoples' attitudes towards receiving the vaccine and that happier people make better health-related decisions.

We trained our models on the merged data set of 6530 observations and 69 variables using an eXtreme Gradient Boosting (XGBoost) algorithm and also used Random Forest and Decision Tree algorithms as robustness tests. After evaluating the models, we found that the XGBoost gave the best-fit metrics compared to the other two methods. We made several contributions to existing literature. First, ours was the first study to conduct a post-COVID-19 cross-country analysis of the most important variables to reach different herd immunity levels. Second, we were also the first study to include subjective measures of well-being in our estimations. Third, we were the first study to differentiate between the most important factors to reach different herd immunity levels. To address our research questions, we used various machine learning algorithms to train our models and determine which algorithm gives us the best fit, i.e., the most reliable predictions. Subsequently, our XGBoost model can be used as a benchmark for future research related to the most important factors for reaching herd immunity levels. Furthermore, this study offered some actionable insights for policymakers on increasing vaccination rates to curb pandemics' health and economic and political effects.

The XGBoost model revealed similar important factors in predicting the 70% and 90% vaccination thresholds to reach herd immunity levels. These included the vaccination policy implemented, international travel controls, the percentage of the population in rural areas and the average temperature. Of significance was happiness's role in attaining the 90% vaccine threshold. Whereas happiness had a lower importance level in achieving the 70% threshold, the importance of happiness in achieving the 90% vaccine threshold was clear. If governments want higher levels of compliance and vaccine uptake, subjective well-being measures such as mood and emotions must be prioritised. Addressing how people feel, in general, towards vaccines and governments is vitally important when policymakers want to push beyond the lower 70% vaccine threshold and achieve the "golden standard" of 90% fully vaccinated.

It would be negligent of us not to discuss our study's limitations. First, the sample of countries under investigation is mostly developed. It will be interesting to extend the sample to determine the policies, characteristics, and subjective well-being measures deemed necessary to increase vaccination rates in developing countries and contrast those with the factors applicable to developed nations.

Second, although we know that lack of international support and cooperation played a significant role in procuring and disseminating vaccines in developing countries, we could not add variables reflecting international support or emergency investment in health care to our models due to high missingness. We acknowledge that these variables could have ranked among the most important variables and potentially have been included in the top five. The missingness of the observations of these variables is further proof of the failures of countries to prepare for pandemics and give international support. The missingness on international support was 67%, implying that international support was given infrequently. When we added the amounts from the developed countries to our sample, they were still minimal. Furthermore, countries did not frequently invest in emergency health care. Of the observations in our dataset on this variable, 74% were missing. Note that these numbers are for developed countries; therefore, it is easy to imagine what the variable would reveal for developing countries. When we added these amounts, it was very little compared to the amounts spent on, for example, vaccines.

References

Abdurrahim, Y., Ali, A. D., Sena, K., & Huseyin, U. (2020). Comparison of deep learning and traditional machine learning techniques for classification of pap smear images. *arXiv*, 2009.06366v1.

Andrade, C., Gillen, M., Molina, J. A., & Wilmarth, M. J. (2022). The Social and Economic Impact of Covid-19 on Family Functioning and Wellbeing: Where do we go from here? *Journal of Family and Economic Issues,43*(2),205-212.

Anik, L., Aknin, L.B, Norton, M.I., &. Dunn, E.W. (2009). Feeling good about giving: The benefits (and costs) of self-interested charitable behavior. Harvard Business School Marketing Unit Working No. 10-012.

Bajos, N., Spire, A., Silberzan, L., Sireyjol, A., Jusot, F., Meyer, L., et al. (2022). When Mistrust in the Government and Scientists Reinforce Social Inequalities in Vaccination against Covid-19. *Frontiers in Public Health, 10*, 908152.

Baldwin, R. (2020). Keeping the lights on: Economic medicine for a medical shock. VoxEU.Org. 2020.

Barbieri, V., Wiedermann, C.J., Lombardo, S., Ausserhofer, D., Plagg, B., Piccoliori, G., et al. (2022). Vaccine Hesitancy in the Second Year of the Coronavirus Pandemic in South Tyrol, Italy: A Representative Cross-Sectional Survey. *Vaccines, 10*, 1584.

Bargain, O., & Aminjonov, U. (2020). Trust and compliance to public health policies in times of covid-19*. Journal of Public Economics, 192*, 104316.

Bartsch, S. M., Ferguson, M. C., McKinnell, J. A., O'Shea, K. J., Wedlock, P. T., Siegmund, S.S., et al. (2020). The Potential Health Care Costs and Resource Use Associated With COVID-19 In The United States. *Health Affairs*, *39*(6),927-935.

Bergmann, M., Bethmann, A., Hannemann, T.-V., & Schumacher, A. T. (2022). Who are the Unvaccinated? Determinants of SARS-CoV-2 Vaccinations among Older Adults Across Europe. *Easy Social Sciences, Mixed 1*, 1-11.

Bíró-Nagy, A., & Szászi, A. J. (2022). The Roots of COVID-19 Vaccine Hesitancy: Evidence from Hungary. *Journal of Behavioral Medicine*, 1–16.

Bloom, D. E., Cadarette, D., & Ferranna, M. (2021). The Societal Value of Vaccination in the Age of COVID-19. *American Journal of Public Health, 111*(6), 1049-1054.

CBC News. (2021). Canada's Pfizer vaccine shipment delayed by winter weather in the US Available from https://www.cbc.ca/news/politics/pfizer-delays-winter-weather-1.5915661 Accessed on 27 July 2023.

Cheong, Q., Quon, S., Concepcion, K., & Kong, J. D. (2021). Predictive Modeling of Vaccination Uptake in US Counties: A Machine Learning–Based Approach. *Journal of Medical Internet Research*, *23*(11).

Corcoran, K. E., Scheitle, C. P., & DiGregorio, B. D. (2021). Christian Nationalism and COVID-19 Vaccine Hesitancy and Uptake. *Vaccine, 39*(45), 6614– 6621.

Czeisler, M. E., Rajaratnam, S. W. M., Howard, M. E., & Czeisler, C. A. (2021). COVID-19 Vaccine Intentions in the United States—December 2020 to March 2021. Working paper, MedRxiv.

De Boeck, K., Decouttere, C., & Vandaele, N. (2020). Vaccine distribution chains in low- and middle-income countries: a literature review. *Omega, 97*,102097.

El-Mohandes, A., White, T. M., Wyka, K., Rauh, L., Rabin, K., Kimball, S. H., et al. (2021). COVID-19 Vaccine Acceptance among Adults in Four Major US Metropolitan Areas and Nationwide. *Scientific Reports, 11*(1), 21844.

Fetzer, T. R., Witte, M., Hencel, L., Jachimowicz, J., Haushofer, J., Ivchenko, A., et al. (2020). Global Behaviors and Perceptions at the Onset of the COVID-19 Pandemic. National Bureau of Economic Research Working Paper No, 27082.

Fieselmann, J., Annac, K., Erdsiek, F. Yilmaz-Aslan, Y., & Brzoska, P. (2022). What are the reasons for refusing a COVID-19 vaccine? A qualitative analysis of social media in Germany. *BMC Public Health, 22*, 846.

Gatwood, J., McKnight, M., Fiscus, M., Hohmeier, K. C., & Chisholm-Burns, M. (2021). Factors Influencing Likelihood of COVID-19 Vaccination: A Survey of Tennessee Adults. *American Journal of Health-System Pharmacy, 78*(10), 879–889.

Goldblatt, D., Fiore-Gartland, A., Johnson, M., Hunt, A., Bengt, C., Zavadska, D., et al. (2022). Towards a population-based threshold of protection for COVID-19 vaccines. *Vaccine, 40*(2), 306-315.

Gomes, I. A., Soares, P., Rocha, J. V., Gama, A., Laires, P. A., Moniz, M., Pedro, A. R., Dias, S., Goes, A. R., Leite, A., & Nunes, C. (2022). Factors Associated with COVID-19 Vaccine Hesitancy after Implementation of a Mass Vaccination Campaign. *Vaccines, 10*(2), 281.

Greyling, T., & Rossouw, S. (2022). Positive attitudes towards COVID-19 vaccines: A cross-country analysis. *PLOS ONE, 17*(3), 0264994.

Greyling, T., Rossouw, S., & Afstereo. (2019). *Gross National Happiness.today*. Available from http://gnh.today Accessed on 20 July 2023.

Jansson, M. K., & Yamamoto, S. (2022). The effect of temperature, humidity, precipitation and cloud coverage on the risk of COVID-19 infection in temperate regions of the USA—A case-crossover study. *PLOS ONE, 17*(9), e0273511.

John Hopkins Healthcare. (2021). Winter Storm Slows US COVID Vaccine Rollout. Available from https://johnshopkinshealthcare.staywellsolutionsonline.com/RelatedItems/6,1650551452. Accessed 25 July 2023.

Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin, 22,* 280–287.

Kerr, J. R., Schneider, C. R., Recchia, G., Dryhurst, S., Sahlin, U., Dufouil, C., et al. (2021). Correlates of Intended COVID-19 Vaccine Acceptance Across Time and Countries: Results from a Series of Cross-Sectional Surveys. *BMJ Open, 11*(8), e048025.

Kim, E.S., Kubzansky, L.D., & Smith, J. (2015). Life satisfaction and use of preventive health care services. *Health Psychology*, *34*(7), 779– 782.

Latkin, C. A., Dayton, L., Yi, G., Konstantopoulos, A., & Boodram, B. (2021). Trust in a COVID-19 Vaccine in the US: A Social-Ecological Perspective. *Social Science and Medicine, 270*, 113684.

Lincoln, T. M., Schlier, B., Strakeljahn, F., Gaudiano, B. A., So, S. H., Kingston, J., et al. (2022). Taking a machine learning approach to optimise prediction of vaccine hesitancy in high income countries. *Scientific Reports*, *12*(1), 1-12.

Lindholt, M. F., Jørgensen, F., Bor, A., & Petersen, M. B. (2021). Public Acceptance of COVID-19 Vaccines: Cross-National Evidence on Levels and Individual-Level Predictors using Observational Data. *BMJ Open, 11*(6), e048172.

Lu, H., Nie, P., & Qian, L. (2020). Do Quarantine Experiences and Attitudes Towards COVID-19 Affect the Distribution of Psychological Outcomes in China? A Quantile Regression Analysis. Global Labor Organization Discussion Paper No, 512.

Ludvigson, S. C., Ma, S., & Ng, S. (2020). Covid19 and the Macroeconomic Effects of Costly Disasters. National Bureau of Economic Research Working Paper No, 26987.

Lyubomirsky, S., Sheldon, K.M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology, 9*(2), 111–131.

Mathieu, E., Ritchie, H., Roser, M., Hasell, J., Appel, C., & Giattino, C. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*, *5*(7), 947-953.

Mental Health and Wellbeing Commission. (2023). COVID-19 and safety in the home. COVID-19 Impact Insights Paper Number 4. Available from https://www.mhwc.govt.nz/assets/COVID-19-insights/Paper-4-COVID-and-safety-in-the-home/ENG_SafetyReport_Summary.pdf Accessed on 21 July 2023.

Nielsen, D. (2016). Tree boosting with XGBoost: why does XGBoost win "every" machine learning competition? Norwegian University of Science and Technology. Available from https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf Accessed on 10 July 2023.

Oli, A. N., Agu, R. U., Ihekwereme, C. P., & Esimone, C. O. (2017). An evaluation of the cold chain technology in South-East Nigeria using Immunogenicity study on the measles vaccines. *Pan African Medical Journal, 27*, 1–5.

Osman, S. M. I., & Sabit, A. (2022). Predictors of COVID-19 vaccination rate in USA: A machine learning approach. *Machine Learning with Applications, 10,* 100408.

Paul, E., Steptoe, A., & Fancourt, D. (2021). Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet Regional Health-Europe, 1*(100012), 1–10.

Plans-Rubió, P. (2022). Percentages of Vaccination Coverage Required to Establish Herd Immunity against SARS-CoV-2. *Vaccines, 10*(5), 736.

Polašek, O., Wazny, K., Adeloye, D., Song, P., Chan, K.Y., Bojude, D.A., et al. (2022). Research Priorities to Reduce the Impact of COVID-19 in Low- and Middle-Income Countries. *Journal of Global Health, 12*, 09003.

Pronkina, E., & Rees, D. I. (2022). Predicting COVID-19 Vaccine Uptake. Institute of Labor Economics (IZA) Discussion Paper No. 15625.

Raciborski, F., Samel-Kowalik, P., Gujski, M., Pinkas, J., Arcimowicz, M., & Jankowski, M. (2021). Factors Associated with a Lack of Willingness to Vaccinate Against COVID-19 in Poland: A 2021 Nationwide Cross-Sectional Survey. *Vaccines, 9*(9), 1000.

Randolph, H. E., & Barreiro, L. B. (2020). Herd Immunity: Understanding COVID-19. *Immunity*, 19;52(5), 737-741.

Rossouw, S., Greyling, T., Adhikari, T. (2021). The evolution of happiness pre and peri-COVID-19: A Markov Switching Dynamic Regression Model. *PLOS ONE, 16*(12), e0259579.

Rozek, L. S., Jones, P., Menon, A., Hicken, A., Apsley, S., & King, E. J. (2021). Understanding Vaccine Hesitancy in the Context of COVID-19: The Role of Trust and Confidence in a Seventeen-Country Survey. *International Journal of Public Health, 66*, 636255.

Sallam, M. (2021). COVID-19 vaccine hesitancy worldwide: a systematic review of vaccine acceptance rates. *Vaccines, 9*(160), 1–14.

Sarracino, F., Greyling, T., O'Connor, K., Peroni, C. & Rossouw, S. (2024). Trust predicts compliance with COVID-19 containment policies: Evidence from ten countries using big data. *Economics and Human Biology,* 101412. https://doi.org/10.1016/j.ehb.2024.101412

Shaham, A., Chodick, G., Shalev, V., & Yamin, D. (2020). Personal and social patterns predict influenza vaccination decisions. *BMC Public Health, 20*(222), 1–12.

Sheikh, A. B., Pal, S., Javed, N., & Shekhar, R. (2021). COVID-19 Vaccination in Developing Nations: Challenges and Opportunities for Innovation. *Infectious Disease Report, 14*(2),429–436.

Streefland, P. H., Chowdhury, A. M. R., & Ramos-Jimenez, P. (1999a). Patterns of Vaccination Acceptance. *Social Science and Medicine, 49*, 1705–16.

Streefland, P. H., Chowdhury, A. M. R., & Ramos-Jimenez, P. (1999b). Quality of vaccination services and social demand for vaccinations in Africa and Asia. *Bulletin of the World Health Organization, 77*,722–30.

The Guardian. (2021). Severe snowstorm forces Greece to halt Covid vaccination drive. Available from https://www.theguardian.com/world/2021/feb/16/severe-snowstorm-forces-greece-to-halt-covid-vaccination-drive Accessed on 20 July 2023.

United Nations, Department of Economic and Social Affairs, Population Division (2022). *World Population Prospects 2022, Online Edition*. Available from https://population.un.org/wpp/Download/Standard/MostUsed/ Accessed on 20 July 2023.

Ward, J. K., Alleaume, C., & Peretti-Watel, P. (2020). The French Public's Attitudes to a Future COVID-19 Vaccine: The Politicization of a Public Health Issue. *Social Science and Medicine, 265*, 113414.

Williams, S., & Shiaw, W. T. (1999). Mood and organisational citizenship behavior: The effects of positive affect on employee organisational citizenship behavior intentions. *Journal of Psychology, 133,* 656–668.

Wollebæk, D., Fladmoe, A., Steen-Johnsen, K., & Ihlen, Ø. (2022). Right-Wing Ideological Constraint and Vaccine Refusal: The Case of the COVID-19 Vaccine in Norway. *Scandinavian Political Studies, 45*(2), 253–278.

World Bank. (2018). The Human Capital Project. World Bank, Washington, DC. Available from https://data.worldbank.org/indicator/HD.HCI.OVRL.UB.MA?end=2020&start=2020&view=bar Accessed on 20 July 2023.

World Bank. (2023a). Climate Change Knowledge Portal. Available from https://climateknowledgeportal.worldbank.org/ Accessed on 20 July 2023.

World Bank. (2023b). Food and Agriculture Organization and World Bank population estimates. Available from https://data.worldbank.org/indicator/EN.POP.DNST?view=chart Accessed on 20 July 2023.

World Bank. (2023c). World Bank Open Data 2023. Available from https://data.worldbank.org/ Accessed on 20 July 2023.

World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision. Available from https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS Accessed on 20 July 2023.

World Health Organization (WHO). (2023). WHO Coronavirus (COVID-19) Dashboard. Available from https://covid19.who.int/ Accessed on 20 July 2023.

Supplementary Information A

To derive our time-series data, which captures sentiment and emotions, we construct variables using Big Data by extracting tweets from Twitter. In our analysis, we extracted two sets of tweets based on keywords, one related to COVID-19 vaccines and the other related to the government. The tweets containing these words amounted to 1,047,000 tweets. We extracted all tweets according to specific geographical areas (country).

For COVID-19 vaccines, we extract tweets using the keywords: *vaccinate, vacc, vaccine, Sputnik V, Sputnik, Sinopharm, Astrazeneca, Pfizer (if NEAR) vaccine, Pfizer-BioNTech, Johnson & Johnson,* and *Moderna*.

For the government, we extract tweets using the keywords: *government, parliament, ministry, minister, senator, MPs, legislator, political, politics, prime minister.*

The first step in our analysis is determining the tweets' language (we detected 64 different languages), and all non-English tweets were translated into English. After translation, we use NLP to extract the tweets' sentiment and underlying emotions. To test the robustness of coding the sentiment of the translated tweets, we use lexicons in the original language, if available, and repeat the process. We compare the coded sentiment of the translated and original text and find the results strongly correlated.

We make use of a suite of lexicons. Each differs slightly but primarily aims to determine the sentiment of unstructured text data. The two lexicons mostly used in our analysis are Sentiment140 and NRC (National Research Council of Canada Emotion Lexicon developed by Turney and Mohammad (2010)). The other lexicons are used for robustness purposes and are part of the Syuzhet package. The lexicons include Syuzhet, AFINN and Bing. The sentiment is determined by identifying the tweeter's attitude towards an event using variables such as context, tone, etc. It helps one form an entire opinion of the text. Depending on the lexicon used, the text (tweet) is coded. For example, if a tweet is positive, it is coded as 0; if neutral, 2; and if negative, 4.

We use the NRC lexicon to code the sentiment (as explained above) and analyse the underlying tweets' emotions. It distinguishes between eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy and disgust (the so-called Plutchik (1980) wheel of emotions). NRC codes words with different values, ranging from 0 (low) to 8 (the highest score in our data), expressing the intensity of an emotion or sentiment.

To construct the time-series data, we use the coding of the tweets and derive daily averages. In this manner, we derive a positive sentiment, a negative sentiment and eight emotion time series. We derive the sentiment time series using different lexicons as a robustness test and compare these results using correlation analyses. For example, we perform additional robustness tests to determine whether the sampling frequency significantly influences the results.

To test the robustness of the *frequency,* we construct the relevant index (time series) per day (the norm); we repeat the exercise but construct the time series per hour. We find similar trends in our hourly and daily time series, indicating that the timescale at which sampling occurs does not significantly influence the observed trend.

To test whether the *volume* of tweets affects the derived time-series data, we extract random samples of differing sizes from the daily text corpus of tweets. The time series based on these smaller samples (50 per cent and 80 per cent of the daily extracted tweets) are highly correlated to the original time series.

Supplementary Information B

Full list of variables

Variable	Description	Scale	Coding	Source
Vaccination policy	Policies for vaccine delivery for different groups	Ordinal	0 - No availability 1 - Availability for ONE of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 2 - Availability for TWO of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 3 - Availability for ALL of following: key workers/ clinically vulnerable groups (non-elderly) / elderly groups 4 - Availability for all three plus partial additional availability (select broad groups/ages) 5 - Universal availability	Mathieu et al. (2021)
Workplace closing	Record closing of workplaces	Ordinal	0 - no measures 1 - recommend closing (or recommend work from home), or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operations 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (e.g. grocery stores, doctors) Blank - no data	Mathieu et al. (2021)
Restrictions on gatherings	Record limits on gatherings	Ordinal	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data	Mathieu et al. (2021)
International travel controls	Restrictions on international travel	Ordinal	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions 4 - ban on all regions or total border closure Blank - no data	Mathieu et al. (2021)
Contact tracing	Record government policy on contact tracing after a positive diagnosis	Ordinal scale	0 - no contact tracing	Mathieu et al. (2021)
Testing policy	Record government policy on who has access to testing Note: this records policies about testing	Ordinal scale	0 - no testing policy 1 - only those who both (a) have symptoms AND (b) meet specific criteria (e.g. key workers, admitted to hospital, came into contact with a known case, returned from overseas)	Mathieu et al. (2021)

	for current infection (PCR tests), not testing for immunity (antibody test)		2 - testing of anyone showing Covid-19 symptoms 3 - open public testing (e.g. "drive through" testing available to asymptomatic people) Blank - no data	
Face coverings	Policies on the use of facial coverings outside the home	Ordinal	0 - No policy 1 - Recommended 2 - Required in some specified shared/public spaces outside the home with other people present or some situations when social distancing not possible 3 - Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible 4 - Required outside the home at all times regardless of location or presence of other people	Mathieu et al. (2021)
Income support	Record if the government provides direct cash payments to people who lose their jobs or cannot work. Note: only includes payments to firms if explicitly linked to payroll/salaries	Ordinal	0 - no income support 1 - government is replacing less than 50% of lost salary (or if a flat sum, it is less than 50% of median salary) 2 - government is replacing 50% or more of lost salary (or if a flat sum, it is greater than 50% of median salary) Blank - no data	Mathieu et al. (2021)
Debt relief	Record if the government is freezing financial obligations for households (e.g. stopping loan repayments, preventing services like water from stopping, or banning evictions)	Ordinal	0 - no debt/contract relief 1 - narrow relief, specific to one kind of contract 2 - broad debt/contract relief	Mathieu et al. (2021)
Public information campaigns		Ordinal	0 -No COVID-19 public information campaign 1 - public officials urging caution about COVID-19 2 - coordinated public information campaign (e.g. across traditional and social media) No data - blank	Mathieu et al. (2021)
Physicians per 1000		Continuous		Mathieu et al. (2021)
Nurses per 1000		Continuous		Mathieu et al. (2021)
Health expenditure (USD)		Continuous		World Bank (2023c)
Out-of-pocket health expenditure		Continuous		World Bank (2023c)
Population rural (Percentage of population in rural areas)	People living in rural areas are defined by national statistical offices. It is calculated as the difference between the total and urban populations.	Percentage		World Bank staff estimates based on the United Nations Population Division's

				World Urbanization Prospects (2018)
Population density	People per square kilometre of land area			United Nations (2022)
Infant mortality rate		Continuous		World Bank (2023c)
Population age 0-9		Continuous		World Bank (2023c)
Population age 10-19		Continuous		World Bank (2023c)
Population age 20-29		Continuous		World Bank (2023c)
Population age 30-39		Continuous		World Bank (2023c)
Population age 40-49		Continuous		World Bank (2023c)
Population age 50-59		Continuous		World Bank (2023c)
Population age 60-69		Continuous		World Bank (2023c)
Population age 70-79		Continuous		World Bank (2023c)
Population age 80 and older		Continuous		World Bank (2023c)
Life expectancy	The average number of years a newborn would live if age-specific mortality rates in the current year were to stay the same throughout its life.	Years		United Nations (2022)
Diabetes prevalence		Continuous		World Bank (2023c)
Comorbidity mortality rate		Continuous		Mathieu et al. (2021)
Smoking prevalence		Continuous		Mathieu et al. (2021)
Pollution mortality rate		Continuous		United Nations (2022)
Average temperature		Celsius		World Bank (2023a)
Human capital index				World Bank (2018)

Human development index				Mathieu et al. (2021)
GDP (USD)		Continuous		World Bank (2023c)
GDP per capita (US$)		Continuous		World Bank (2023c)
GNH	General happiness	Continuous		Greyling et al. (2019)
Sadness GNH	The emotion general sadness	Continuous		Greyling et al. (2019)
Trust GNH	The emotion general trust	Continuous		Greyling et al. (2019)
Anticipation GNH	The emotion general anticipation	Continuous		Greyling et al. (2019)
Fear GNH	The emotion general fear	Continuous		Greyling et al. (2019)
Surprise GNH	The emotion general surprise	Continuous		Greyling et al. (2019)
Joy GNH	The emotion general joy	Continuous		Greyling et al. (2019)
Anger GNH	The emotion general anger	Continuous		Greyling et al. (2019)
Disgust GNH	The emotion general disgust	Continuous		Greyling et al. (2019)
GNH Gov	Happiness towards government	Continuous		Greyling et al. (2019)
Trust Gov	The emotion trust towards government	Continuous		Greyling et al. (2019)
Joy Gov	The emotion joy towards government	Continuous		Greyling et al. (2019)
Surprise Gov	The emotion surprise towards government	Continuous		Greyling et al. (2019)
Sadness Gov	The emotion sadness towards government	Continuous		Greyling et al. (2019)
Anticipation Gov	The emotion anticipation towards government	Continuous		Greyling et al. (2019)
Disgust Gov	The emotion disgust towards government	Continuous		Greyling et al. (2019)
Fear Gov	The emotion fear towards government	Continuous		Greyling et al. (2019)
Anger Gov	The emotion anger towards government	Continuous		Greyling et al. (2019)
VADER pos Gov	Positive sentiment towards the government	Continuous		Greyling et al. (2019)
VADER neg Gov	Negative sentiment towards the government	Continuous		Greyling et al. (2019)
VADER sent Vac	Sentiment towards the vaccine	Continuous		Greyling et al. (2019)
VADER neg Vac	Negative sentiment towards the vaccine	Continuous		Greyling et al. (2019)
VADER pos Vac	Positive sentiment towards the vaccine	Continuous		Greyling et al. (2019)
GNH Vac	Happiness towards the vaccine	Continuous		Greyling et al. (2019)

Surprise Vac	The emotion surprise towards vaccines	Continuous		Greyling et al. (2019)
Anticipation Vac	The emotion anticipation towards the vaccine	Continuous		Greyling et al. (2019)
Disgust Vac	The emotion disgust towards the vaccine	Continuous		Greyling et al. (2019)
Sadness Vac	The emotion sadness towards the vaccine	Continuous		Greyling et al. (2019)
Fear Vac	The emotion fear towards the vaccine	Continuous		Greyling et al. (2019)
Anger Vac	The emotion anger towards the vaccine	Continuous		Greyling et al. (2019)
Trust Vac	The emotion trust towards the vaccine	Continuous		Greyling et al. (2019)
Joy Vac	The emotion joy towards the vaccine	Continuous		Greyling et al. (2019)

Supplementary Information C

Importance of factors, XGBoost, Random Forest and Decision Tree – 90 % threshold

XGBoost – 90% threshold	Random Forest – 90% threshold	Decision Tree – 90% threshold
Vaccination policy	Vaccination policy	Vaccination policy
International travel controls	Restrictions on gatherings	Testing policy
Percentage of population in rural areas	International travel controls	Public information campaigns
Happiness	Debt relief	Contact tracing
Average temperature	Facial coverings	Facial coverings
Population density	Testing policy	International travel controls
Human Development Index	Income support	Income support
Facial coverings	Contact tracing	Restrictions on gatherings
Workplace closing	Comorbidity mortality rate	Population aged between 20-29
Restrictions on gatherings	Average temperature	Population aged between 0-9
Income support	Infant mortality rate	Population aged between 10-19
Life expectancy	Workplace closing	Human Development Index
Pollution mortality rate	Population aged 80 and older	Percentage of population in rural areas
Out-of-pocket health expenditure	GDP (USD)	Infant mortality rate
Debt relief	Public information campaigns	Out-of-pocket health expenditure
Trust (GNH)	Diabetes prevalence	Population aged between 30-39
Human capital index	Out-of-pocket health expenditure	Population aged between 40-49
Diabetes prevalence	Population density	Health expenditure (USD)
Human Development Index	Smoking prevalence	GDP per capita (USD)
DDP per capita (USD)	Life expectancy	Human capital index
Sadness (GNH) – lack of happiness	Disgust (GNH) – lack of happiness	Population density
Sentiment towards vaccines	Human capital index	Smoking prevalence
Smoking prevalence	Anger (GNH) – lack of happiness	Anger (GNH) – lack of happiness