The 2nd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/4/20220561

Predicting COVID-19 Induced Mortality Utilizing Vaccination Status Data

Xiaoxuan Chen[1,3,†] and Zhiqi Tang[2,4,†]

[1] Weinberg College of Art and Science, Northwestern University, 1918 Sheridan Road, Evanston, Illinois, 60201, United States
[2] Department of Art and Science, University of Toronto, 35 St. George Street, Toronto, M5S 1A4 ON, Canada
[3] hebechen2025.2@u.northwestern.edu

[4] zhiqi.tang@mail.utoronto.ca
[†]These authors contributed equally.

Abstract. COVID-19 is a prevalent pandemic that has caused of millions of mortalities. While medical and statistical research have proven the effectiveness of COVID vaccines to reduce mortality rate in specific areas. This study aims to explore the quantitative effect of vaccine in preventing in the global level. Mortality and vaccine data was acquired from Kaggle, which summarized related data from various sources. Multiple linear regression, quasi-Poisson regression, LASSO regression, and random forest are applied to the model to analyze and predict the effect of vaccine on mortality. The adjusted R square value of these four models is 0.6670, 0.6863, 0.4901, and 0.7152 respectively. The residual boxplots of all four models show that LASSO model sometimes make extremely inaccurate predictions, and the random forest model is the most accurate model, which matches the comparison result of the adjusted R square values. Therefore, the random forest model can be potentially used to predict future COVID-related deaths based on vaccine-related data.

Keywords: quasi-poisson regression, LASSO regression, random forest, COVID-19, vaccines

1. Introduction

COVID-19, causing respiratory problems, is a prevalent pandemic initiated from Wuhan, China at the end of 2019. Due to globalization and global travelling, the spread of COVID-19 became rapid during 2020 and developed into a global pandemic [1]. Symptoms of COVID-19 include fever, headache, and fatigue, and Mortality data around the world shows that COVID-19 is a highly transmissible disease that can cause death after infection [2]. To control the pandemic, many studies around the virus have been done. Research found out that the respiratory problem occurred in airway was caused by a novel coronavirus that was genetically like the SARS virus. After the publication of genome sequence of the virus on 11 January 2020, the development of vaccine targeting the genome of virus initiated. Vaccination is a useful method to control the impact caused from COVID-19 on not only personal level, but also national and global level [3].

Many medical studies have proven the effectiveness of COVID-19 vaccinations against the infection of COVID-19 and COVID-related deaths, suggesting the help of vaccination provided to personal health level. For example, a study published in The New England Journal of Medicine proved the effectiveness

DOI: 10.54254/2753-8818/4/20220561

of first and second dose against the infection of coronavirus. In addition, the uptake of the second dose should be encouraged because of its effectiveness [4]. Many clinical trials of vaccines done by researchers also prove the worthiness of their products. BNT162b2 vaccines have been proven to have a protection efficiency of 95% towards people who are older than 16 years old [5]. In addition to medical research, there are also research that based on statistical analysis to strengthen the fact about the effectiveness of vaccination. Research from the Journal of Public Health validated the positive impact of COVID-19 vaccine based on mortality data of Europe and Israel by using the method of non-linear Poisson regression. The result shows that the efficacy of COVID-19 vaccine in preventing death is 72%, indicating a very high efficiency of vaccine [6].

Based on research, multiple different regression models have been utilized to study the relationship between COVID-19 vaccination and COVID-related mortality. Much research conducted the model based on Poisson related models. Besides the non-linear Poisson model mentioned earlier, other Poisson-related models such as negative-binomial and quasi-Poisson model could also be choices. A study published on the Journal of Environmental Research explored the COVID data from Brazil based on negative-binomial and quasi-Poisson model [7]. These models might be preferred to use because of the over dispersed nature of COVID-19 mortality data [8]. Because of the large usage of Poisson-related models, Poisson would also be a starting point of this project to serve as a baseline for comparison. In addition to Poisson-related models, this project also aims to use multiple regression methods to construct the model that can predict the mortality rate based on vaccination status in the most accurate way.

Unlike most research which aims to explore the vaccination efficiency by using models which include many detailed variables, such as climate and variants of viruses, and focusing on data in specific countries or area, this project aims to explore the general effect of vaccinations on the number of COVID-19 related deaths by analyzing data published by different countries. Doing so, a more direct of the impact of COVID-19 vaccination can be obtained. Therefore, the objective is to regress the data of number of deaths by using more direct variables such as vaccination rate and population to find out how quantitatively vaccinations can help to slow down the death rate based on data from multiple countries. This quantitative evaluation is important not only because it can provide an overall report of how vaccination help to control the pandemic, but also imply the effectiveness of vaccinations and potential areas for COVID-19 vaccinations to improve. Moreover, this model may also serve as a reference to predict death rate based on vaccination status in different countries.

2. Methods

2.1. Dataset

The dataset used to achieve the goal was called COVID Vaccination vs. Mortality, which was acquired from Kaggle [9]. It mainly describes the number of daily new deaths and people who got vaccinations and full vaccinations for different countries at different dates. There are no missing data in this dataset. This dataset combines information from three other datasets. These three datasets are COVID-19 World Vaccination Progress from Kaggle, COVID-19 cases and deaths data from WHO, and 2021 World Population dataset from Kaggle.

The data set contains 9 variables and 32911 observations in total. The first two variables are country and the corresponding ISO code. There are 196 countries in total. Countries have their different number of dates which when they collect data, which is the third variable. The fourth variable, total vaccination, describes the total number of dose of vaccines each country used at a particular date. The fifth and the sixth variable, describe the number of people who got at least one shot of vaccine and people who got full vaccination respectively. The seventh variable is the 2021 population for different countries, and the final variable describes the daily new death for different countries at each date.

2.2. Dataset Transformation

Since the factor of time is included in the original dataset, running analysis will be difficult since observations will not be independent of each other (e.g., observations made in Jan. very similar to those

DOI: 10.54254/2753-8818/4/20220561

made in Feb). To counter this difficulty, we transformed the original dataset into cumulative rate data. Through such transformation the factor of time will be removed from the data, and it will be more appropriate to apply general regression methods on rate data rather than count data.

Another issue that needed to be accounted for is that some countries have very little observations. For instance, Madagascar had less than 10 observations in 2021, but in comparation some countries like Denmark and Germany had a total of 365 observations in 2021 (1 observation per day, that is). Having observations from countries like Madagascar in our dataset means we will likely have some extreme observations. If one such country had its observations made in the earlier part of the year, our calculation of its death rate will yield a relatively low value, compared to its "true" value. To prevent this, we simply removed observations from such countries by keeping only countries that has more than 200 observations. This certainly have the potential of making our new dataset biased as it seems developed countries tend to have more complete data on vaccination and covid-induced mortality, while developing countries generally have less observations (and thus less accurate annual rate data). In Table 1, we listed the new variables and their corresponding units, descriptions, and calculations. Concluding our dataset transformation, our new dataset had 61 observations of 8 variables, each row representing a country and contains information on its vaccination, mortality, and population.

Table 1. New variables created from original variables and their corresponding units, descriptions, and calculations. Note that 3 variables, "country", "time" and "population" remained unchanged.

New Variable	Description	Calculation
Death rate (count/year)	The rate of death caused by covid-19	Annual cumulative death / 1
Vaccination use rate (count/year)	The rate at which doses of vaccine are being used	Annual cumulative vaccines usage / 1
Vaccination rate (count/year)	The rate at which people who got at least one shot of vaccine increases	(Number of vaccinated people max. - # vaccinated people min.) / 1
Initial # vaccinated people (count)	# People who got at least one shot of vaccine at the start of 2021	# Vaccinated people min.
Fully Vaccination rate (count/year)	The rate at which people who got full vaccine shots increases	(Number of fully vaccinated people max. - # fully vaccinated people min.) / 1
Initial # fully vaccinated people (count)	# People who got full vaccine shots at the start of 2021	# Fully vaccinated people min.

2.3. Data Analysis

Several data analysis methods were applied to our dataset, including multiple linear regression, quasi-Poisson regression, LASSO regression, and random forest. To fit our multiple linear regression model, we first built our model using all available variables (except "country" since it is not a variable of interest; a model that can predict covid mortality rate of any country is desired, and the dataset does not include all countries in the world) using the "lm ()" function. Model output showed that only population and vaccine use rate are not significant predictors. An automatic stepwise selection was then ran using the "step ()" function to generate a final model. We then attempted to apply quasi-Poisson regression to our dataset as the distribution of the response variable does not resemble that of a Poisson distribution, and thus the normal Poisson regression is not applicable to our data. Again, we built a full model with all potential variables, using the "glm ()" function (with argument "family = QuasiPoisson"). A manual backward selection was then performed based on the p-values of the variables, since quasi-Poisson regression models do not have AIC and thus, we cannot apply the "step ()" function to it. This project also applied the LASSO (Least Absolute Shrinkage and Selection Operator) regression on our data, using 0.8 data split and 10-fold repeated (5 times) cross-validation. The value of lambda was decided

DOI: 10.54254/2753-8818/4/20220561

using cross-validation and minimum MSE rule (chosen lambda == 148.4132). The random forest model was fitted using the "ranger ()" function from the R studio package of the same name.

3. Results

For the linear regression model, the full model we first built showed only 4 statistically significant variables with varying p-values. The most important predictor is the initial number of vaccinated people, with a p-value of 1.6e-05. The other significant variables are vaccination rate (p-value = 0.01678), the initial number of fully vaccinated people (p-value = 0.02149), and fully vaccination rate (p-value = 0.00548). The rest of the variables, that is population and vaccination use rate, had non-significant p-values (p-value = 0.23860 for population and p-value = 0.47256 for total vaccination). This full model reached an adjusted R squared of 0.8253 and a extremely low overall p-value of 2.2e-16.

The automatic stepwise selection using the function "step ()" indeed selected the 4 statistically significant predictors mentioned before, and the output of the final model showed significantly small p-values for all predictors included, with the variable initial # fully vaccinated people having the highest p-value (0.025) and the variable vaccination rate having the smallest p-value (3.26e-08). This final model reached an adjusted R squared of 0.8268 (slightly higher than the full model) and an overall p-value of 2.2e-16 (same as the full model). We tested whether all not the assumptions of linear regression were met and found that all these assumptions are roughly satisfied. The linearity assumption was checked using correlation plots (the plots showed rough linear trends between response and predictors). The normality assumption was examined using the histogram and normal QQ plot of studentized residuals (histogram showed approximately normal distribution; normal QQ plot showed rough straight line). The equal variance and independent errors assumptions were checked by the fitted values vs studentized residuals plots (the plots showed unequal but reasonable variance).

For the Poisson regression, as previously mentioned, we first checked the distribution of our response variable (death rate) using a histogram (Fig. 1) and found that it does not follow a Poisson distribution. A test of dispersion (test whether the mean and variance of the data are the same; if the variable follows Poisson distribution, then the mean and variance should be roughly the same) yielded a mean of 40464.54 and a variance of 7960893890, again suggesting over-dispersion and that the distribution of the response variable does not resemble that of a Poisson distribution. As afore mentioned in the method section, this means the assumption of Poisson regression is violated by our data and we cannot apply the approach to the dataset. We thus applied a similar method, the quasi-Poisson regression, which does not assume a Poisson distribution of the response variable and account for over-dispersion. Resulting final model had only 3 variables, including vaccination rate (p-value = 2.75e-12), fully vaccination rate (p-value = 2.66e-13) and initial number of fully vaccinated people (p-value = 0.00788). Again, like the multiple linear regression model, the p-value for initial number of fully vaccinated people is the highest among all.

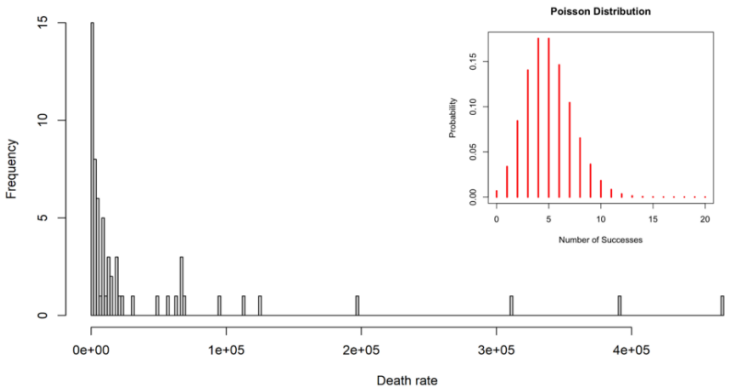

Figure 1. The histogram of the response variable, death rate. The smaller graph on the top right shows the shape of a Poisson distribution, for comparison.

DOI: 10.54254/2753-8818/4/20220561

4. Model Evaluation

We wish to evaluate and compare the models on two things: their overall efficiency (the ability to explain the variation in the dataset with less predictors), and their accuracy (the predictions made should deviate from the actual values as little as possible). However, for the purpose of this project, that is to generate a model that is potentially helpful in predicting future covid-19 mortality based on current or estimated future vaccination conditions, we tend to value accuracy over efficiency, that is, if a model can provide more accurate and precise predictions using more predictors, we still consider it as a better model, compared to a model that uses less variables but gives worse predictions.

The last 2 models, that is the LASSO regression model, and the random forest model are both more complicated and difficult to represent or interpret, compared to the other 2 models, as they involved the application of algorithms and does not show every step of the model-fitting or predicting process. They also wouldn't generate simple and common parameters like overall p-values, adjusted R squared and AIC values, which means more general parameters or comparison approaches are required for us to find the better model. The methods we selected for this job are: 1. Cross-validation (a method that evaluates the models on their performance on different subsets of training data and then calculates their average error) that generates RMSE (root mean square error), R squared (the amount of variation explained by the model), and MAE (mean absolute error), and 2. Visual comparison of the models' residuals (i.e., a rather simple comparison that visualizes the precision of the models and sees whether a model tends to make more inaccurate predictions) using residual boxplots. We first evaluated the models using these two approaches with 2021 data (the data we sued to fit and generate these models). For the cross-validation, we mainly used "train ()" function from the "caret" R package. Seeds were set first prior to the application of this method (set.seed(1) before applying train() function, and seed = c(1,2,3,4,5,6,7,8,9,10,11) for the trainControl() function), to make the process reproducible. 10 – fold (the number of different subsets that the given data set is to be split into) cross-validation generated the following results (Table 2).

Based on the values of RMSE, R squared and MAE from the 10 – fold cross-validation tests, the random forest model clearly out-performed the rest of the models in terms of accuracy. We then compared the models again, visually, using residual boxplots, shown below in Figure. 2.

Table 2. The comparison between the number of variables, RMSE, R squared and MAE of the models, from cross-validation tests done using 2021 data.

Models	# Variables	RMSE	R^2	MAE
Multiple linear	4	67279.55	0.667036	37902.5
Quasi-Poisson	3	52323791	0.68634	21371939
LASSO	4	78906.73	0.490166	46560.3
Random forest	/	38380.75	0.715235	19421.11

DOI: 10.54254/2753-8818/4/20220561

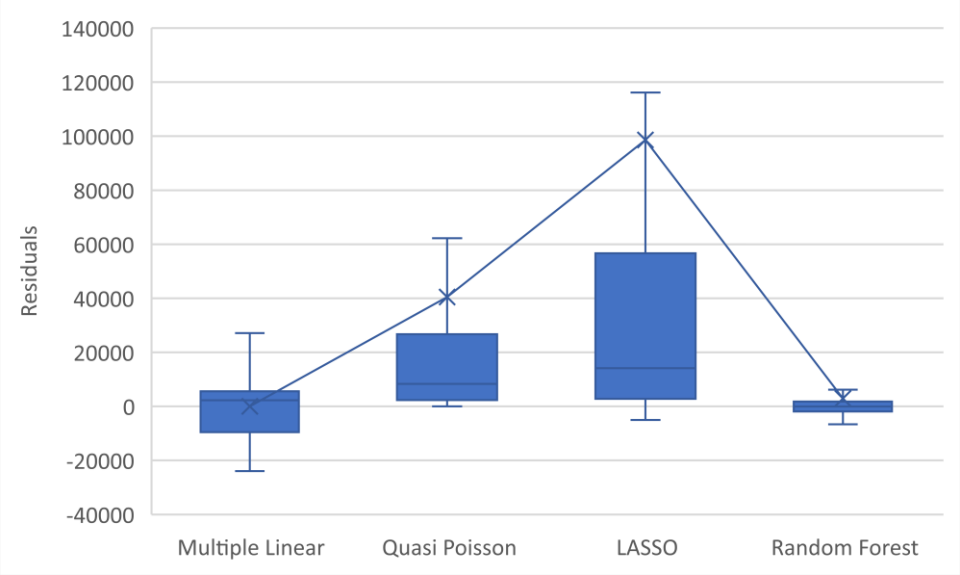

Figure 2. Residual boxplots of the 4 models. Outlier Points were hidden for the ease of viewing.

These boxplots demonstrated that the LASSO regression model will sometimes make extremely inaccurate predictions, compared to the rest of the models. The plots also confirmed that the random forest model is indeed more accurate and precise, agreeing with the conclusion of the cross-validation tests. To further test the predictive power of the models, we also utilized the 2022 data from the original dataset. We performed the same transformation to these observations and acquired a dataset with 61 observations (equal number of observations to the 2021 dataset we used in model building). We again applied 10-fold cross-validation test, using the same seeds for the "trainControl ()" and "train ()" functions. Models were fitted using their parameters decided by data analysis we did by 2021 data. Resulting RMSE, R squared, and MAE (Table 3) again demonstrated that the random forest model had significant better performance, as it has notably lower RMSE and MAE, and explained more than 80% of variation in the dataset.

Table 3. The comparison between the number of variables, RMSE, R squared and MAE of the models, from cross-validation tests done using 2022 data.

Model	# Variables	RMSE	R^2	MAE
Multiple linear	4	95794.32	0.6094378	53967.44
Quasi-Poisson	3	94165534	0.7502265	38456540
LASSO	4	114011.7	0.5161049	66456.2
Random Forest	/	52301.19	0.8045589	31157.63

Residual boxplots (Fig. 3) are again used to make a visual comparison, though in this case it seems that the multiple linear regression model is also doing well, though based on overall comparison, the random forest model would be the better model.

DOI: 10.54254/2753-8818/4/20220561

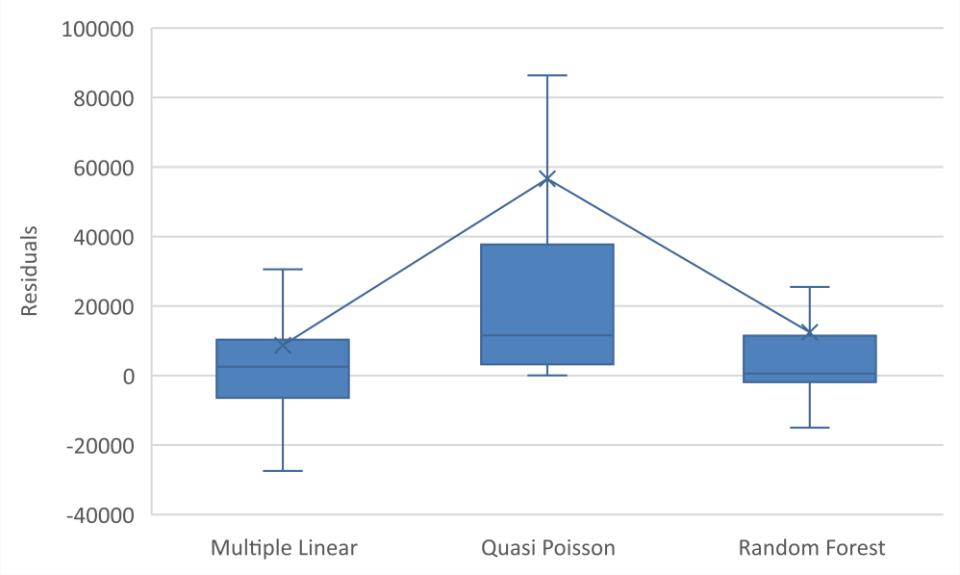

Figure 3. Residual boxplots of 3 of the models we built. Again, the outlier points were hidden. Plot for the LASSO model is not present as its extremely large residuals will alter the scale of the graph and weaken the comparation.

5. Discussion

Although the random forest model clearly outperformed the rest of the models in terms of precision and accuracy, the difficulties in presenting and interpreting this model made it less ideal if one wishes to better understand the correlation between vaccination status and mortality rate. Nonetheless, based on the criteria of the project, that is to generate a model with the most reliable predictive power, the random forest model is indeed the best choice, based on the analyses ran during this project. Also, as previously mentioned, the multiple linear models also showed relatively good performance, and is potentially useful for interpretation and indicating vaccine efficiency. For further research, it might be helpful to include more potential factors that may affect covid mortality to further increase the precision of the model. It is also intriguing to compare the results from this research to information regarding measures taken to contain covid spread.

6. Conclusion

In summary, the analysis provided sufficient evidence that there is significant correlation between vaccination status and covid-induced mortality, as all models built that can output p-values showed significantly low p-values. In terms of making future predictions using vaccination status data, our evaluation of the models built demonstrated that the random forest model tends to generate more accurate and precise predictions, despite of its disadvantage of being a “black box” model that is difficult to interpret and present. The project also found that the “vaccination rate” and “initial # fully vaccinated people” are the 2 most “popular” variables that are constantly selected by the different models, but further analysis would be required to determine whether one can draw real-world inferences from this (i.e., whether vaccination rate and the number of fully vaccinated people are mor important for reducing covid-mortality). Nonetheless, the research indeed confirmed the correlation between vaccination status and covid mortality, rebuking the anti-vaccination claims and provided a potentially applicable way of predicting future mortality using current or estimated vaccination status.

References

[1] S. Platto, T. Xue, and E. Carafoli. COVID19: an announced pandemic. *Cell D*eath Dis 11, 799 (2020).

[2] CDC. Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-

testing/symptoms.html.

[3] T. Le, Z. Andreadakis, and A. Kumar. The COVID-19 vaccine development landscape. *Nature Reviews*. 19: 305-306 (2020).

[4] J. L. Bernal, N. Andrews, and C. Gower.Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. The New England Journal of Medicine. 385:585-594 (2021).

[5] F. P. Polack, S. J. Thomas, and N. Kitchin. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. The New England Journal of Medicine. 383:2603-2615 (2020).

[6] K. Jabłońskaa, S. Aballéab, and M. Toumic. The real-life impact of vaccination on COVID-19 mortality in Europe and Israel. Public Health. 198: 230-237 (2021).

[7] S. Ibarra-Espinosa, E. Dias de Freitas, and K. Ropkins. Negative-Binomial and quasi-poisson regressions between COVID-19, mobility and environment in S˜ao Paulo, Brazi. Environmental Research. 204(2022): 112369.

[8] J. M. Ver Hoef, and P. L. Boveng. Quasi-Poisson VS. Negative Binomial Regression: How Should We Model Overdispersed Count Data? Ecology, 88: 2766-2772 (2007).

[9] Kaggle. COVID Vaccination V.S. Mortality. https://www.kaggle.com/datasets/sinakaraji/covid-vaccination-vs-death.