Canadian Journal of Undergraduate Research

Predicting Vaccine Uptake Behaviour Using Machine Learning: An infodemiological Study in the United States

Xingzuo Zhou[1*], Yiang Li[1*],
Shuheng Yang[1], Xiyan Shi[2], and Xingyou Zhou[2]

[1]University College London
[2]University of Toronto
[*]Two authors contributed equally and are co-first authors

July 18, 2021

Abstract

This paper outlines a framework to predict the COVID-19 vaccination uptake rate. Through web search queries and clinical data, this study aims to use both statistical and Machine Learning methods to better predict the vaccine uptake rate in the United States. This paper also uses RMSE among all methods and scores across different classifiers of Machine Learning methods to compare prediction accuracy, which suggests the random-forest algorithm with ensemble data gives the most accurate predictions of the probability that unvaccinated people receive their first dose of the COVID-19 vaccine.

Keywords
Population Health, Infodemiology, Machine Learning, COVID-19, Vaccine

1 Introduction

Roughly a year after the outbreak of the COVID-19 pandemic in Wuhan, China, two vaccines, the Pfizer-BioNTech vaccine (BNT162b2) and the Moderna vaccine (mRNA-1273), were approved for emergency use in the United States (Mahase, 2020 [1]). Understanding and predicting the future progress of the vaccination, especially with more accurate predictions using Machine Learning, serves important purposes to the public health policy-makers and practitioners who could make more optimal vaccine distribution and implement further infodemiological interventions tackling vaccine hesitancy (Ashrafian and Darzi, 2018 [2]). Though the US Center for Disease Control (CDC) publishes daily statistics on vaccines administered across the United States, monitoring vaccinations beyond that clinical data can make more accurate predictions of future vaccinatioin rate and allow a better understanding of the infodemiological nuances behind the statistics by using non-clinical web data such as search queries (Hansen et al., 2016 [3]).

Traditionally, statistical methods were widely used to make vaccination predictions based on non-clinical data which refers to the statistics collected on an individual's demographic, socio-economic, institutional, and other non-medical information. By using 15 non-clinical socio-demographic factors, Bryden et al., (2019) [4] applied regression methods to predict uptake of vaccination and the result suggests the model has explained 30% of the overall variance. With the advance of big data, the domain of non-clinical data expands to include public-health-related internet interactions made by individuals on a daily basis, including tweets posted, comments, and search queries. In Cimke (2020) [5], the author utilised an idea of extracting the non-clinical online data, the relative search volume of queries words related to "vitamin", from the web and used Kruskal-Wallis analysis to determine the vaccination differences between seasons and years, including during the period of the COVID-19. They have concluded that the interest in that term increased statistically significantly since the pandemic controlling for seasonal changes. Building upon the non-clinical prediction

methods, an infodemiology study by Ksete, et al (2020) [6] used Multivariate Regression to explore the predictive power of not only non-clinical personal attitude towards scientific information but also the clinical experience of severe respiratory diseases on the Influenza vaccination rate. The result suggests the area under ROC for this mixed prediction method is 85%. Such exploration utilising both clinical and non-clinical variables in the statistical regression can be used for infodemiological intervention such as a computerised reminder system to promote vaccination rate (Szilagyi et al., 1992 [7]). A similar application using statistical method has been made on the COVID-19 vaccination. Shmueli (2021) [8] found that having the Health Belief Model (HBM) variables such as perceived benefits of COVID-19 vaccine and the perceived costs associated with infection alongside the Theory of Planned Behavior (TPB) variables of self-efficacy and levels of subjective norms as co-variates could achieve a statistically significant prediction to the variance of vaccine uptake in Israel with the accuracy of 78%.

Recently, infodemiological studies also implemented Machine Learning algorithms on non-clinical data to better predict vaccination. Using unsupervised machine learning algorithms, Latent Dirichlet Allocation (LDA) and Contextual Random Walk Traps (CRWT), on the 1.99 million posts from the parenting blog websites over the 105 months, Tangherlini et al., (2016) [9] suggested the importance of non-clinical sources in shaping the narrative towards vaccination. Such importance is quantified in Carrieri et al., (2021) [10], where the authors used the supervised random-forest machine-learning algorithm on area-level indicators of institutional and socio-economic backgrounds to predict the vaccine hesitancy rate in Italian local authorities and help public health practitioners to run targeted awareness campaigns. The outcome suggests the non-clinical features of the waste recycling rate and the employment rate have the highest predictive powers in the random forest algorithm, having an area under ROC of 0.836. In addition to using those common algorithms, Gothai, et al (2021) [11] proposed using supervised machine learning of Holts Winter Model in the prediction that captures the seasonal variations across the year to improve accuracy. Apart from improving the algorithm for greater prediction accuracy, Hansen et al. (2016) [3] discovered a substantial improvement in prediction accuracy when using both clinical and non-clinical web data, that is, the ensemble data.

However, there is no such study that implements statistical and machine-learning prediction based on both clinical and web data on COVID-19 vaccination. To better set public health interventions on vaccination, understanding the speed of vaccination is crucial. Therefore, our research question focuses on how to predict the COVID-19 vaccination rate more accurately using both clinical and web information?

In this paper, we used an infodemiological method, that is, combining clinical and web data, to analyze vaccination uptake rate. Specifically, we compared the accuracy of different predictions through RMSE among all models and score among machine learning classifiers including SVR, LASSO and the random-forest. Lastly, we compared predictions of vaccination uptake rate using those models in terms of accuracy.

2 Materials & Methods

2.1 Sources of Data

We treated the daily new cases of COVID infection published by the United States Center for Disease Control (CDC) as the clinical data while the relative interest searched from Google Trends are treated as the non-clinical web data. We qualitatively assessed the social media posts in the CoAID social media data (Cui and Lee, 2020 [12]), which were first processed by using 'quanteda' and 'glmnet' packages in R. Similar to Cimke's extraction [5] method, we selected 66 vaccination-related words with high frequencies. Since some of them did not show enough data in Google Trends, we ultimately chose 18 words for each of the attitudes of positive, negative, and neutral. These words were then searched for the relative search volume on Google Trends from 21/12/2020 to 20/5/2021 in the United States. As Google Trend allows up to 5 words per search, a reference word 'Joker' was used during the search so that we could standardise the index of relative search volume across each search. Then we added all standardised indices together by three categories as listed in Table 1.

For the outcome variable, as Hansen [3] proposed, vaccination-to-expectation ratio is a more accurate measurement compared to simple daily vaccination. It is defined as follows,

$$\frac{\text{daily first-dose vaccination}}{\text{number of unvaccinated people expected to be vaccinated in hundreds}},$$

where people expected to be vaccinated are defined as those are not vaccinated at all.

Category	Related Search Word	
Negative (anti-vaccine)	vaccine hurts	vaccine disadvantage
	vaccine fever	no vaccine
	vaccine pain	vaccine variant
	vaccine headache	vaccine restriction
	vaccine side effect	vaccine reaction
	vaccine death	vaccine adverse
	get covid after vaccine	vaccine risk
	vaccine allergy	vaccine uncertainty
	vaccine cost	vaccine blood clot
Neutral (Uncertain)	vaccine update	vaccine reliability
	vaccine ingredient	vaccine passport
	vaccine type	vaccine tracker
	vaccine manufacturer	vaccine developer
	vaccine effectiveness	vaccine last
	vaccine safety	current vaccination
	vaccine function	pfizer
	vaccine quality	moderna
	vaccine rate	vaccine used
Positive (pro-vaccine)	covid vaccine available	vaccine cdc
	vaccine near me	vaccine authorized
	vaccine registration	vaccinate child
	vaccine advantage	vaccine doses
	vaccine appointment	vaccine card
	vaccine booking	second dose
	vaccine benefits	first dose
	vaccine location	vaccine certificate
	vaccine volunteering	vaccinated

Table 1: Web Data: Search Categories and related words

daily_ 0 in Table 2 represents this ratio.

variable	Obs	Mean	Std.Dev.	Min	Max
daily_0	151 (days)	.4420417	.2702492	.0036553	1.133301
negative	151	2.187026	.7559989	.5774648	4.821918
neutral	151	3.697505	1.177863	1.44898	6.704545
positive	151	45.64148	16.48795	17.4929	82.85833
date	151	22345	43.73404	22270	22420

Table 2: Summary of Raw Data

2.2 Statistical Approach

Naive

"Naive" refers to Hansen's [3] naive baseline $\hat{E(t)} = E(t-1)$, which simply assumes a constant vaccination-to-expectation ratio.

Linear Regression

Linear Regression simply estimates $\hat{E(t)} = \mu + \beta_1 Positive + \beta_2 Neutral + \beta_3 Negative$. Positive, Neutral, and Negative are variables representing the calculated interest for the three attitudes. To find the coefficients, we use Ordinary Least Square (OLS) estimation. This model assumes linear specification.

Auto-Regressive Model

To predict with higher accuracy, Auto-Regressive Model can be used. Specifically, we chose to use use 7 lags, that is, AR(7) model, as there is statistically significant partial correlation among those periods. AR(7) estimates $\hat{E}(t) = \mu + \sum_{i=7}^{7} \beta_i E(t-i)$, where there are 7 auto-regressive terms involved, and 7 βs control the weight that each past observation has on the prediction. AR(7) assumes linear specification, and other common restrictions such as exo-geneity.

2.3 Machine Learning Approach

To implement machine learning, we used STATA module *r_ ml_ STATA* [13] to run Python packages *pandas*, *NumPy*, *scikit-learn*. For all classifications, we used 10-fold cross-validation.

LASSO Regression

The LASSO regression, stands for Least Absolute Shrinkage and Selection Operator, is very similar to Linear Regression but has better accuracy of the predictions. It avoids over-fitting of the irrelevant features by selecting a reduced set of the known co-variates to be used in a model. In this study, we combined clinical and web data, and made predictions using LASSO.

Support Vector Machine classification

A support vector machine classification is a supervised machine learning model which splits the data-set into two categories. We applied it on the continuous variable here, i.e using SVR. Support Vector Regression (SVR) generates the regression similar to the linear regression model. The difference is, in contrast to the ordinary least squares, we find the coefficients of support vector regression by minimising the norm of the coefficient vector. It means we minimise

$$\sum_t V(E(t) - \mu - \beta_1 \hat{E_c}(t) - \beta_2 \hat{E_w}(t)) + \frac{\lambda}{2}(\mu^2 + \beta_1^2 + \beta_2^2)$$

In the formula, we defined $V(r) = |r| - \epsilon$, unless $|r| < \epsilon$, V(r)=0 . Both ϵ and λ are hyper parameters which allow us to define the tolerance level of error in our model.

Boost classification

This is a ensemble method which reduces error by introducing a strong classifier from a number of weak classifiers. This is done by building a initial model to the training data and then building models to minimise the variance. Each model is based on the former small models. The final prediction is a weighted sum of these sequenced model, known as weak classifiers.

Random forest classification

Random forecast classification is another ensemble method called bagging algorithm, together with the featured randomness. It reduces the variance of individual classification trees by randomly selecting from the data-set. Averaging these uncorrelated predictors produces the final prediction in this algorithm.

3 Results

Table 3 shows the RMSE results of predicting vaccination uptake using clinical, web data, and

the combination of the two (with the methods presented in Section 2). "Naive" refers to naive baseline $\hat{E(t)} = E(t-1)$ as proposed by Hansen, Lioma and Mølbak (2016 [3]).

Table 4 shows the results of predictions with different approaches. Compared to all other approaches, random-forest prediction is relatively accurate. Accuracy can be increased further by appending more actual data, namely, training. Notably, all predictions are lower than actual data.

Table 5 shows the scalars (in other word, parameters) of our best performance – SVR, boost and random-forest ensemble model. Maximun of accuracy attained is 72.74% through random-forest classification.(Score=0.7274).

Figure 1-3 illustrate 10-fold Cross-Validation with ensemble data. The optimal index for SVR, boost and random-forest classification is 96, 114, and 77 respectively. Figure 4-6 illustrate in-sample predictions of three classifications with ensemble data; and figure 7-9 show the out-sample predictions.

		Clinical Data	Web Data				Ensemble			
	Naive	AR(7)	Linear Regression	SVR	boost	random-forest	AR(7)+LASSO	SVR	boost	random-forest
RMSE	.16328	*.09946*	*.08622*	.08622	.07647	.06551	.27732	**.08057**	**.04744**	**.04135**

Table 3: RMSE of predictions; Blue Highlighted: The lowest RMSE; Bold: Better than Clinical/Web Data; Italic: Better than Ensemble

Dates	Actual Data	AR(7) Output	LR Output	SVR Output	boost Output	random-forest Output
14/5/2021	.58029807	.4319468	.5629587	0.52139005	.47253132	.48134117
15/5/2021	.47418791	.3345546	.3060468	0.24165748	.2522806	.25984949
16/5/2021	.24403456	.2079464	.2257222	0.19238094	.11112721	.11764043
17/5/2021	.47782029	.2168641	.3324861	0.30286211	.27370921	.27816318
18/5/2021	.53782621	.4076946	.3471294	0.31646487	.36477288	.39335935
19/5/2021	.53458529	.3819385	.4125393	0.34810558	.39031478	.41188358
20/5/2021	.42125088	.381938	.2765025	0.29561842	.33157799	.35025324

Table 4: Actual data and predictions

Classifier	Key Scalars
Support Vector Machine	Optimal C= 11 Optimal Gamma = 0.1 Training Score (Accuracy) = 0.9248 Testing Score (Accuracy) = 0.1441 Best Index = 95 SE in Testing Accuracy = 0.7585
boost	Optimal Learning Rate = 0.2 Optimal Number of Estimators = 15 Training Score (Accuracy) = 0.9871 Testing Score (Accuracy) = 0.7209 Best Index = 114 SE in Testing Accuracy = 0.1538
random-forest	Optimal Maximum Depth = 8 Optimal Maximum Features = 8 Training Score (Accuracy) = 0.9750 Testing Score (Accuracy) = 0.7274 Best Index = 77 SE in Testing Accuracy = 0.1003

Table 5: Scalars (Parameters) of SVR, boost and random-forest, with ensemble data, 10-fold CV
Note: Scores = accuracy*100%

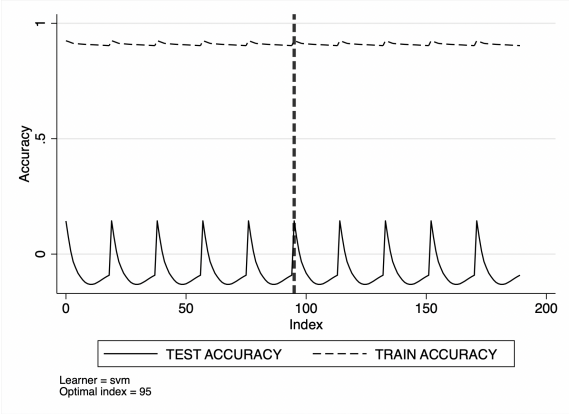

Figure 1: 10-fold Cross-Validation
Classifier = Support Vector Machine

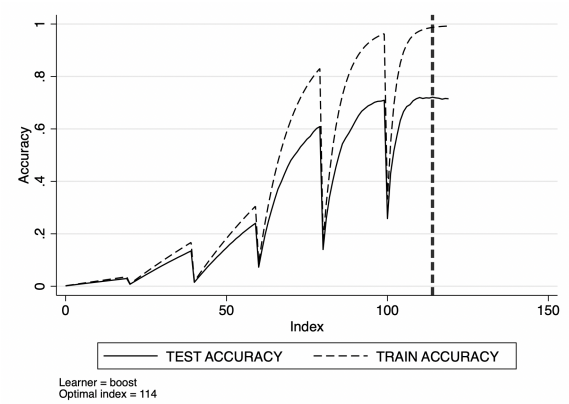

Figure 2: 10-fold Cross-Validation
Classifier = boost

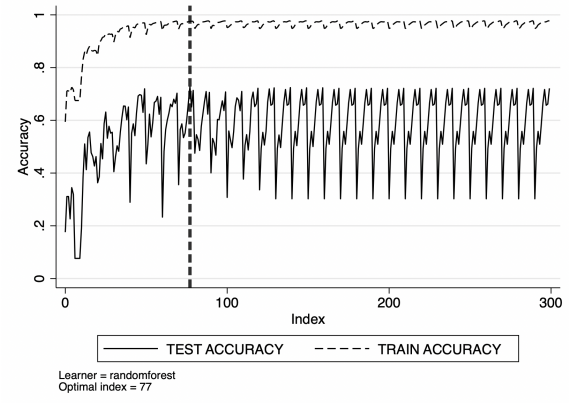

Figure 3: 10-fold Cross-Validation
Classifier = random-forest

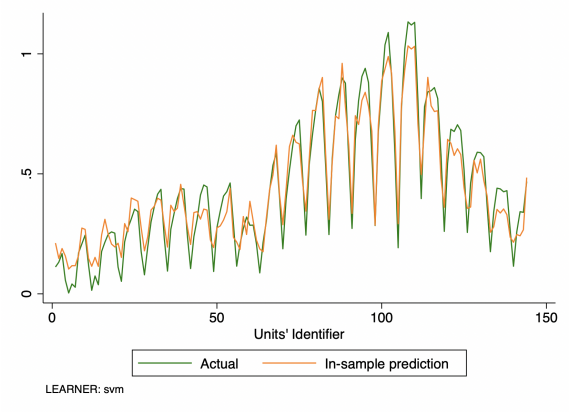

Figure 4: In-Sample Prediction
Classifier = Support Vector Machine

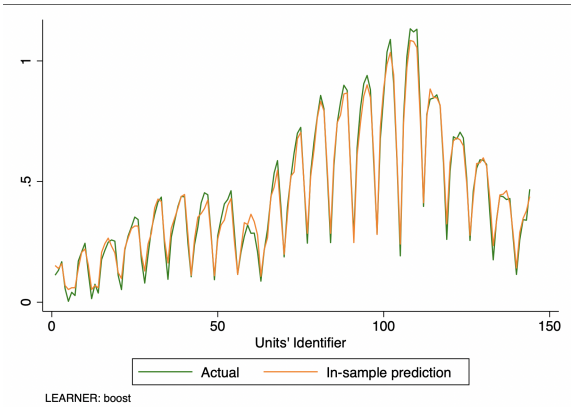

Figure 5: In-Sample Prediction
Classifier = boost

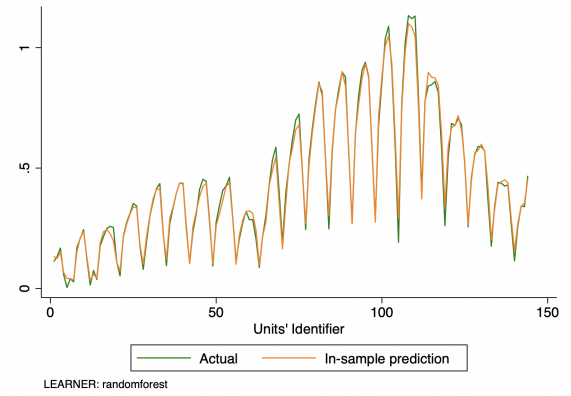

Figure 6: In-Sample Prediction
Classifier = random-forest

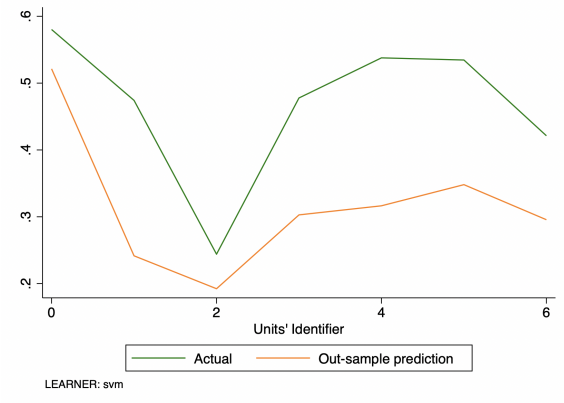

Figure 7: Out-Sample Prediction
Classifier = Support Vector Machine

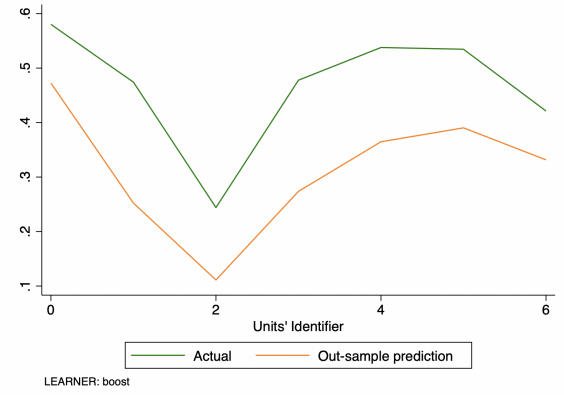

Figure 8: Out-Sample Prediction
Classifier = boost

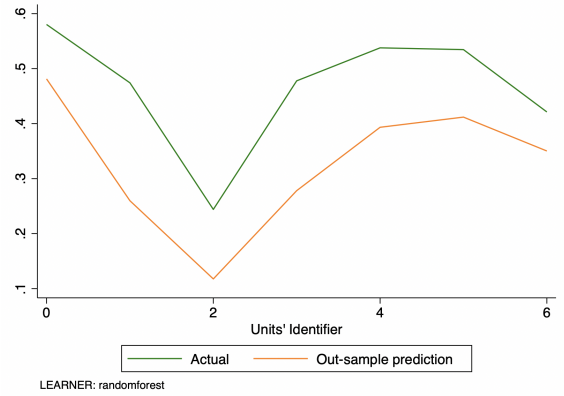

Figure 9: Out-Sample Prediction
Classifier = random-forest

4 Discussion

By Hansen's [3] definition, across all statistical and Machine Learning approaches, a model with the smallest RMSE has the best prediction. RMSE, the root mean square error, gives the differences between predictions and actual data.

$$RMSE = \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}}$$

With STATA modules "regress", "dsregress", "r_ml_STATA" and Python packages **pandas**, **NumPy** and **scikit-learn**, we implement AR(7), Lasso, SVR, boost and random-forest on our data. Using Hansen's criteria, Ensemble random-forest has the best prediction due to the smallest RMSE. As observed, random-forest with ensemble data also has the largest prediction accuracy (72.74%) and smallest standard error (0.1003) of test accuracy among all machine learning classifiers. This indicates that the random-forest with ensemble data performs the best.

Compared to machine learning with Clinical/Web Data, Machine Learning with ensemble data have smaller RMSE, indicating improvements in the prediction.

All prediction methods suggest lower vaccination-to-expectation ratios, that is, smaller values of $\frac{\text{daily first-dose vaccination}}{\text{number of people expected to be vaccinated in hundreds}}$. As the denominator is number of people expected to be vaccinated in hundreds, the percentage of this ratio can also be interpreted as probability that unvaccinated people receiving their first dose of COVID vaccine. With random-forest classification and ensemble data, we obtained the highest accuracy of 72.74%. This result of accuracy is the percentage format of "score", which involves statistical inference. As shown in Table 3 and Figure 9, our prediction provides a relatively accurate trend in vaccination update rate.

The last column in Table 4 is our most accurate prediction of the "vaccination-to-expectation ratio". For example, with random-forest classification, we predict that 0.35 out of 100 unvaccinated people will receive their first dose of COVID-19 vaccine on May 20, 2021. In other words, those unvaccinated people are anticipated to receive their first dose COVID-19 vaccine with a probability of 0.35% on May 20, 2021.

Indeed, reverse causality may potentially exist when using cases to predict vaccination

uptake rate. Briefly, less infection provides people for the sense of safety, hence the vaccination goes down; as vaccination goes down, infection potentially rises. Fortunately, our model solves this issue. In reality, protection of COVID vaccines take at least 2-4 weeks, including registration, and second dose; that is, vaccination influences new COVID cases two weeks after the first dose. As vaccination does not affect new cases rate in seven days, the exogeneity assumption between our variables are satisfied. Furthermore, the new cases rate are all related due to the spread of virus, so our relevance assumption also holds. In this way, we use instrumental variable regression, with seven-day data, to eliminate potential reverse causality issue. More details please refer to Appendix I. Instrumental variable regression also applies to the "web data". As the searches are more volatile and unpredictable, one-day data of web data is determined to use. As searches are serial correlated, this one day data covers other days as well.

Additionally, another limitation of our study is to use Google API solely. For search queries alone, there are many other available services such as Microsoft Bing and the past studies have also revealed the potential for predictions via other API from discussion forums and video websites (Kwok et al., 2021 [16], Tang et al., 2021 [17]); hence, combining other data sources in the future may improve the accuracy further.

Despite those limitations, the Machine Learning algorithms we explored are versatile to many research aspects within and beyond public health predictions. It can be taken by national, municipal, or county public health practitioners to better understand the nuances behind an individual's decision making on whether to take or not take a vaccine and to predict the future rate of vaccination given the current trend. When adding more time-series non-clinical observations on individual behaviour apart from web queries such as tweet posts and most-liked comments, the predictions can be more accurate. Most importantly, the machine learning algorithm we presented can be used for other purposes when adopted for public health surveillance, empirically testing hypotheses, and causal inferences (Beam and Kohane, 2018 [14]). However, it is worth noting in future studies on the intrusion of privacy and its relative benefits to society, especially during an era that social norms evolved the traditional definition of privacy (Mooney and Pejaver, 2018 [15]).

5 Conclusions

The key findings are:

- With ensemble data, the classifier, random-forest, gives the most accurate predictions among our frameworks. It gives a score of 0.7274, indicating 72.74% test accuracy. Besides, it has the smallest RMSE comparing to other methods.
- From 14/05/2021 to 20/05/2021, the most accurate predictions of probabilities that unvaccinated people receive their first dose COVID vaccine are 0.48%, 0.26%, 0.12%, 0.28%, 0.39%, 0.41%, 0.35% respectively. It may be also converted into the number of people receiving first dose COVID vaccine per 100 unvaccinated people. For instance, on 20/05/2021, 0.35 per 100 (equivalently, 35 per 10000) unvaccinated people in the U.S. are anticipated to receive their first dose of COVID vaccine.

For future studies, we aim to add Holt-Winters to our framework. As coronavirus is potentially to be weaker in summer (Bhattacharjee, S.., 2020)[18]), Holt-Winters, which analyses seasonal changes, may increase the accuracy of predictions. It will be available if the timeframe of our data is more than 12 months, which is not possible given the vaccine is just approved less than 7 months prior to the time of writing.

Acknowledgements

We appreciate Dylan Kneale and Meg Wiggins for their helpful guidance.

References

[1] Mahase, E. (2020). Covid-19: Pfizer and BioNTech submit vaccine for US authorisation. BMJ (Clinical research ed.), 371, m4552. doi:10.1136/bmj.m4552

[2] Ashrafian, H., & Darzi, A. (2018). Transforming health policy through machine learning. PLoS Medicine, 15(11), e1002692.

[3] Dalum Hansen, N., Lioma, C., & Mølbak, K. (2016). Ensemble learned vaccination uptake prediction using web search queries. Paper presented at the Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.

[4] Bryden, G. M., Browne, M., Rockloff, M., & Unsworth, C. (2019). The privilege paradox: Geographic areas with highest socioeconomic advantage have the lowest rates of vaccination. Vaccine, 37(32), 4525-4532. doi:doi.org/10.1016/j.vaccine.2019.06.060

[5] Çimke, S., & Yıldırım Gürkan, D. (2021). Determination of interest in vitamin use during COVID-19 pandemic using Google Trends data: Infodemiology study. Nutrition, 85, 111138. doi: doi.org/10.1016/j.nut.2020.111138

[6] Keske, Ş., Mutters, N. T., Tsioutis, C., & Ergönül, Ö. (2020). Influenza vaccination among infection control teams: A EUCIC survey prior to COVID-19 pandemic. Vaccine, 38(52), 8357-8361. doi: doi.org/10.1016/j.vaccine.2020.11.003

[7] Szilagyi, P. G., Rodewald, L. E., Savageau, J., Yoos, L., & Doane, C. (1992). Improving Influenza Vaccination Rates in Children With Asthma: A Test of a Computerized Reminder System and an Analysis of Factors Predicting Vaccination Compliance. Pediatrics, 90(6), 871-875. Retrieved from https://pediatrics.aappublications.org/content/pediatrics/90/6/871.full.pdf

[8] Shmueli, L. (2021). Predicting intention to receive COVID-19 vaccine among the general population using the health belief model and the theory of planned behavior model. BMC public health, 21(1), 804. doi:10.1186/s12889-021-10816-7

[9] Tangherlini, T. R., Roychowdhury, V., Glenn, B., Crespi, C. M., Bandari, R., Wadia, A., . . . Bastani, R. (2016). "Mommy Blogs" and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites. JMIR Public Health Surveill, 2(2), e166. doi:10.2196/publichealth.6586

[10] Carrieri, V., Lagravinese, R., & Resce, G. (2021). Predicting vaccine hesitancy from area-level indicators: A machine learning approach. medRxiv, 2021.2003.2008.21253109. doi:10.1101/2021.03.08.21253109

[11] Gothai, E., Thamilselvan, R., Rajalaxmi, R. R., Sadana, R. M., Ragavi, A., & Sakthivel, R. (2021). Prediction of COVID-19 growth and trend using machine learning approach. Materials Today: Proceedings. doi: doi.org/10.1016/j.matpr.2021.04.051

[12] Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.

[13] Cerulli, G. (2021). Machine Learning using Stata/Python. arXiv preprint arXiv:2103.03122.

[14] Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. Jama, 319(13), 1317-1318. doi:10.1001/jama.2017.18391

[15] Mooney, S. J., & Pejaver, V. (2018). Big data in public health: terminology, machine learning, and privacy. Annual review of public health, 39, 95-112.

[16] Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. J Med Internet Res, 23(5), e26953. doi:10.2196/26953

[17] Tang, L., Fujimoto, K., Amith, M. T., Cunningham, R., Costantini, R. A., York, F., . . . Tao, C. (2021). "Down the Rabbit Hole" of Vaccine Misinformation on YouTube: Network Exposure Study. Journal of Medical Internet Research, 23(1), e23262.

[18] Bhattacharjee, S. (2020). Statistical investigation of relationship between spread of coronavirus disease (COVID-19) and environmental factors based on study of four mostly affected places of China and five mostly affected places of Italy. arXiv preprint arXiv:2003.11277.

Appendix I

For unbiased and consistent IV estimators, the order and rank conditions are required. The order condition states that the number of exogenous variables is no less than the number of endogenous variables. In our model, endogenous variables are the dates that vaccine becomes effective. The 7-day and other web-data variables are exogenous variables. In this case, the order condition is satisfied. Let λ_n represent the effective dates of different vaccines; θ_ms represent the first seven days after first dose vaccine. "n" is the number of types of vaccine, and "m" represents the first seven days. Through two-stage-least-square regression, we substitute the effective dates of vaccines by the first seven days.

$$\hat{\lambda_1} = \hat{\alpha_1}\theta_1 + ... + \hat{\alpha_7}\theta_7$$

$$\hat{\lambda_2} = \hat{\beta_1}\theta_1 + ... + \hat{\beta_7}\theta_7$$

$$...$$

$$\hat{\lambda_m} = \hat{\pi_1}\theta_1 + ... + \hat{\pi_7}\theta_7$$

Another vital condition is the rank condition. In brief, relevance required second-stage regressions are not perfectly multicollinear. As the effectiveness of vaccines are not constantly distributed, the rank condition also holds. Mathematically, it means that λs are not linearly dependent.

Appendix II

Machine Learning STATA code [13]:

```
r_ml_stata $y $x, mlmodel($learner)
  in_prediction("in_pred")
  cross_validation("CV")
  out_sample("data_test")
  out_prediction("out_pred") seed(10)
  save_graph_cv("graph_cv")
```

Data and codes are available at here for checking and replication purpose.