Factors Contributing to Activeness in Vaccination Among Adults in the United States: Results from the 2019 and 2020 NHIS Data

Hanwen Zhang, Wenchu Pan, Weijia Jin, and Hao Wu

Abstract

Objective To study on factors contributing to activeness in vaccination and to provide recommendations to increase the acceptance and uptake of vaccines in the United States, and find the best combination of missing data imputation and feature selection to obtain the best prediction on the factors contributing to vaccination activeness among American adults aged from 18 to 99. **Method** Four methods for missing data imputation (impute using mean, draw, random forest and Multivariate Imputation via Chained Equations (MICE) and three methods for feature selection (Singular Value Decomposition, Random Forest, and LASSO) were proposed. Accuracy was evaluated by pairwise comparison across combinations of one imputation for missing data and one method for feature selection. XGBoost, LASSO, Ridge and random forest were used for prediction with weighted mean squared error (MSE) and area under the curve (AUC) evaluating the precision of comparison between 2019 and 2020 data and comparison between flu and pneumonia vaccines. **Results** 1. Respectively, the five combined methods each had mean MSE across four prediction models as 0.0903585725, 0.090954905, 0.0860341575, 0.08382767, and 0.0884698075, where 2-3 gave the smallest error under this criteria. 2. p-values of the difference between 2019 and 2020 data are all $p < 0.0001$. 3. The value of AUC of ROC respectively regarding LASSO, Ridge and Random Forest, were 0.8000824, 0.8012672, 0.785161. **Conclusion** 1. The combination of imputing draw and LASSO for feature selection gave the best estimation. 2. 2019 and 2020 data are significantly ($p < 0.0001$) different from each other and using 2019 vaccination data to predict 2020 vaccination data will get overestimated results. 3. Using flu vaccines to predict pneumonia vaccines will obtain great estimation.

Introduction

The success of vaccination efforts to curb pandemic requires a broad public uptake of immunization. However, despite being the most successful and the most effective public health action, vaccination is also perceived as unsafe and unnecessary so that some people will feel hesitated when they are informed to finish their vaccinations [1]. The issue clearly emerged during the 2009–2010 H1N1 pandemic [2], when vaccine uptake reported less than 50% of the expected coverage in target populations in many countries around the globe [3]. WHO SAGE working group defined this phenomenon as vaccine hesitancy [4]. According to their works, "vaccine hesitancy refers to a delay in acceptance or refusal of vaccination despite availability of vaccination services." Vaccine hesitancy, among the population who are at risk, may significantly abate the acceptance rate. (This issue came into public attention during the COVID-19 pandemic [5].) According to Sherman et al COVID-19 vaccination acceptability study (CoVAccS) in the United Kingdom, among 1500 UK adults, 64% of participants reported being very likely to be vaccinated against COVID-19, 27% were unsure, and 9% reported being very unlikely to be vaccinated [6]. Meanwhile among healthcare personnels' intentions on taking vaccines against COVID-19 in Greece, 803 of 1571 (51.1%) stated their intention to get vaccinated while 768 (48.9%) stated their intention to decline vaccination [7]. Therefore, to elevate the clinical impact along the vaccination process, there is a demand for practitioners to understand the factors associated with activeness to receive a vaccine, although it is considered as a complicated process since people's behavioral decision-making on vaccination are usually complex in nature, and can vary according to

context [8], including types of vaccines, age, educational background, clinical characteristics, and previous vaccination [8-9].

Methods

Data Overview

The data we used in this study is the National Health Interview Survey (NHIS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS) to monitor the population's health status in the United States [10]. It is a nationally representative household survey of the U.S. civilian noninstitutionalized population and conducted continuously throughout the year by interviews or telephone [10].

In 2019, CDC released a revised version of NHIS survey design. In the 1997-2018 NHIS, questions from the family questionnaire were asked about the family as a whole and about each member of the family, while in the 2019-2020 redesigned survey, family-level content is collected in the sample adult and/or sample child questionnaire [10]. Therefore, several data files are included in the currently released NHIS data, including the annual Sample Adult, annual Sample Child, Imputed Income for the Sample Adult and Sample Child, and Paradata. Only for the 2020 survey, it includes two new data sets: 1. Sample Adult Longitudinal for analyses of data from 2019 and 2020 for the same individuals, and 2. Sample Adult Partial for combining data from multiple years that include 2019 and 2020 [11]. For this analysis, we limited the studied data to the 2019 and 2020 survey, which includes 534 and 618 variables, respectively.

For each year's survey, we used the Sample Adult Interview component.Within this component, one adult (age $\geq$ 18) is randomly selected from each household to complete survey questions as total on their health conditions, health-related behaviors, and healthcare service utilization. We took every observation into our analysis to make it more generalizable among the US adult population. Before data manipulation, in the 2019 and 2020 survey, respectively, there are 31997 and 31568 respondents from the age range 18 to 99.

Statistical Analysis

The analysis part was implemented using R[12] programming language and Python[13]. Before the analysis part, data was merged across the years 2019 and 2020 which included various sets of vaccine choice outcomes: flu, pneumonia, and shingles. Highly missing and extremely correlated observations were also excluded from the analysis in consideration of the estimation bias. The statistical analysis consisted of two parts: missing data imputation and feature selection. The first part aimed to obtain a more comprehensive data set and eliminate sampling bias, and the latter part was designed to investigate the factors that would affect people's choices on vaccination. In this study, we proposed four methods to impute missingness and three methods to select features, and we were objective to find the best combination of two methods to obtain the best prediction. For example, if method 1 was used for missing data imputation and method 2 was used for feature selection, the combination would be named as 1-2. In total, there were five combined combinations before running the prediction due to the accessibility of combining methods (**TABLE 1**).

Table 1: Weighted MSE of Prediction Models using 19-20 Data.

Index	Imputation	Feature Selection
1-1	Impute using mean	SVD
2-1	Impute using draw	SVD
2-2	Impute using draw	Random Forest
2-3	Impute using draw	LASSO
4-1	MICE	SVD

For the missing data imputation, we chose to impute using mean, impute using draw since there exists a great amount of categorical data, impute using random forest - it was extremely slow, but was recommended in

prior studies[14], and Multivariate Imputation via Chained Equations (MICE), which pre-assumed missing at random (MAR), and similar with random forest, it was slow but recommended[15]. Within the feature selection, we proposed basic Singular Value Decomposition (SVD) dimension reduction, random forest predictor on imputed data and Lasso.

Status of three vaccinations were merged together as our first prediction label. Prediction was conducted with XGBoost, LASSO, Ridge and random forest on each method to evaluate their performances and obtain a best data preprocess method for the next steps. Afterwards, each model was conducted on the chosen method to compare the status between 2019 and 2020. When running the model, 2019 data was treated as the training data, while 2020 data was recorded as testing data and weighted mean squared error (MSE) were applied to evaluate the precision of each prediction model. Lastly, we want to dig into the prediction of training on one vaccination while predicting on another. Hence, we choose to use the merged data with flu vaccines and pneumonia vaccines as training and predicting labels, respectively, area under the curve (AUC) was utilized to evaluate the precision.

Results

The five combined methods 1-1, 2-1, 2-2, 2-3 and 4-1 were firstly modeled by XGBoost, LASSO, Ridge, and Random Forest to measure and compare their prediction results. Shown in **TABLE 2**, Respectively, the five combined methods each had mean MSE across four prediction models as 0.0903585725, 0.090954905, 0.0860341575, 0.08382767, and 0.0884698075, where 2-3 gave the smallest error under this criterion, although the difference between method appeared not to be huge Similar situations also exist among models, which both performed the best on 2-3 method. For instance, on Random Forest model, method showed a lowest mse with 0.0817, and a highest mse on method 2-1, which is 0.086. Therefore, we came to the conclusion that method 2-3, which was a combination of Imputation draw and Lasso feature selection, performed best on the prediction. Therefore, we decided to use 2-3 as our data preprocess method in the upcoming steps.

Afterwards, we built models using the data from 2019 to make predictions in 2020. (For the convenience of prediction, predicting data should have the same features with training data. Hence, no feature selection method was conducted on 2020's data, instead, we directly selected the same features with 2019's data, so as the pneumonia vaccine data in the next step). As shown in **TABLE 3**, each model made a favorable predict on 2020's data. XGBoost performed the best with weighted MSE of 0.0815. An error of 0.02 may come from the difference between the situation of 2019 and 2020, which was an outbreak of pandemic. The p value of t test between predicted value and actual value in 2020 is far less than 0.01 in all of our models, suggesting a significant decrease in the real data, compared with our predicted value.

Lastly, in order to explore the consistency between the status of two types of vaccination, we conducted predictions on the pneumonia vaccination data by training flu vaccination data. The value of AUC of ROC curve (receiver operating characteristic curve), respectively regarding LASSO, Ridge and Random Forest, were 0.8000824, 0.8012672, 0.785161. In XGBoost model, error rate was used instead because it was a binary classification. The precision of the prediction on pneumonia vaccination applying machine learning models training on flu vaccination strongly proved that the status of the two vaccination were similar.This part of the results is shown in **TABLE 4**.

Table 2: Weighted MSE of Prediction Models using 19-20 Data.

Methods	Data	XGBoost	LASSO	Ridge	Random Forest	Mean MSE
1-1	19-20	0.08418563	0.09608238	0.09586431	0.08530197	0.0903585725
2-1	19-20	0.08486711	0.09652833	0.09637586	0.08604832	0.090954905
2-2	19-20	0.08160598	0.08962064	0.08952294	0.08338707	0.0860341575
2-3	19-20	0.08036669	0.08673767	0.08650456	0.08170176	0.08382767
4-1	19-20	0.08455289	0.08839483	0.09501444	0.08591707	0.0884698075

Table 3: Results Comparing between 2019 and 2020 data

Tests	XGBoost	LASSO	Ridge	Random Forest
MSE	0.08147308	0.0862166	0.0866432	0.08146679
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 4: Results Comparing between Flu and pneumonia vaccines

Tests	XGBoost (MSE)	LASSO	Ridge	Random Forest
AUC	0.09601987	0.8000824	0.8012672	0.785161

Conclusion

In this study, comparing the different combination methods of imputation and feature selection, we found that method 2-3, the combination of imputing draw and LASSO for feature selection, gave the best estimation when using weighted MSE as criteria.among random forest, XGBoost, lasso and Ridge. The performance of XGboost, which is a nonparametric model, is better than the other three parametric models for prediction. On the other hand, we have also observed that if the results produced by the 2019 data are used to predict the 2020 data, the predicted results are significantly greater than the true value, namely, using the vaccination data from 2019 to predict 2020 vaccination will get overestimated results. This difference is guessed to be related to the COVID-19 pandemic and will be mentioned in detail in the discussion section. Although there was a difference between annual data, while using flu vaccine data to construct a model to predict pneumonia vaccine, according to the value of AUC of ROC curve (receiver operating characteristic curve), the prediction performs great.

Discussion

Prediction on vaccination activeness has been a public health issue that is worth attention. In this study,we proposed a method by comparing different models to model this phenomenon. This study, however, still has room for improvement, including but not limited to the recall bias that comes with the questionnaire and the weakening of reference caused by non-random missing values. For example, the lack of background information mitigated the power of interpretation of the result. According to previous studies, concerns about safety due to COVID-19 were raised under the current pandemic, which caused patients' delay and avoidance of medical care[5]. Our result did show the vaccination activeness was relatively lower than that of the year of 2019, but given that there was no related data about COVID-19, a well-rounded result could not be concluded. Efficiency and accuracy can also be improved in future study regarding the algorithm, such that we used algorithms with higher accuracy but higher time complexity to process missing values. Besides, our imputation method is not optimal. Imputing draws does not correctly reflect correlation between variables, thus underestimates the variance, so multiple imputation or Gibbs Sampling should be used if more background knowledge related to the data was provided.

Contribution

- Wenchu Pan: Data prepossession and model fitting for Random Forest;
- Weijia Jin: Model fitting for XGBoost;
- Hao Wu: Model fitting for LASSO AND Ridge;
- Hanwen Zhang: Presentation slides editing and report writing.

Reference

[1] Dubé, E., Vivion, M., & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert review of vaccines*, 14(1), 99–117.

[2] Poland G. A. (2010). The 2009-2010 influenza pandemic: effects on pandemic and seasonal vaccine uptake and lessons learned for seasonal vaccination campaigns. *Vaccine*, 28 Suppl 4, D3–D13.

[3] Schmid, P., Rauber, D., Betsch, C., Lidolt, G., & Denker, M. L. (2017). Barriers of Influenza Vaccination Intention and Behavior - A Systematic Review of Influenza Vaccine Hesitancy, 2005 - 2016. *PloS one*, 12(1), e0170550.

[4] MacDonald, N. E., & SAGE Working Group on Vaccine Hesitancy (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34), 4161–4164.

[5] Czeisler MÉ, Marynak K, Clarke KE, et al. Delay or Avoidance of Medical Care Because of COVID-19–Related Concerns — United States, June 2020. *MMWR Morb Mortal Wkly Rep 2020*;69:1250–1257.

[6] Sherman, S. M., Smith, L. E., Sim, J., Amlôt, R., Cutts, M., Dasch, H., Rubin, G. J., & Sevdalis, N. (2021). COVID-19 vaccination intention in the UK: results from the COVID-19 vaccination acceptability study (CoVAccS), a nationally representative cross-sectional survey. *Human vaccines and immunotherapeutics*, 17(6), 1612–1621.

[7] Maltezou, H. C., Pavli, A., Dedoukou, X., Georgakopoulou, T., Raftopoulos, V., Drositis, I., Bolikas, E., Ledda, C., Adamis, G., Spyrou, A., Karantoni, E., Gamaletsou, M. N., Koukou, D. M., Lourida, A., Moussas, N., Petrakis, V., Panagopoulos, P., Hatzigeorgiou, D., Theodoridou, M., Lazanas, M., ... Sipsas, N. V. (2021). Determinants of intention to get vaccinated against COVID-19 among healthcare personnel in hospitals in Greece. *Infection, disease and health*, 26(3), 189–197.

[8] Osterholm, M. T., Kelley, N. S., Sommer, A., & Belongia, E. A. (2012). Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *The Lancet. Infectious diseases*, 12(1), 36–44.

[9] Lin, Y., Huang, L., Nie, S., Liu, Z., Yu, H., Yan, W., & Xu, Y. (2011). Knowledge, attitudes and practices (KAP) related to the pandemic (H1N1) 2009 among Chinese general population: a telephone survey. *BMC infectious diseases*, 11, 128.

[10] Centers for Disease Control and Prevention. About the National Health Interview Survey. 2020. Available at www.cdc.gov/nchs/nhis/about_nhis.htm Accessed December 8, 2021

[11] Centers for Disease Control and Prevention. 2020 NHIS. 2021. Available at https://www.cdc.gov/nchs/nhis/2020nhis.htm. Accessed December 8, 2021

[12] R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

[13] Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

[14] Nguyen, C. , Wang, Y. and Nguyen, H. (2013) Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, 6, 551-560. doi: 10.4236/jbise.2013.65070.

[15] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03