Proposal of Predicting Adverse Events after COVID-19 Vaccination

Type of this project: Research Project
Group 4

September 29, 2021

1 Information of Members

1.1 Student ID & Name:

20793419 GUO, Yuchen
20797037 LEI, Lijun
20799126 YUAN, Fangxu
20806228 LONG, Yuepeng
20812186 LIN, Lirong

1.2 Declaration

DECLARATION

We hereby declare that all the work done in this Project is of our independent effort. We also certify that this project is done solely within the course but not other scopes(e.g., other courses and research projects), and we have never submitted the idea and product of this Project for academic or employment credits.

All members in Group 4	2021.9
Group	Date

2 Project Description

2.1 Introduction

Nowadays, the COVID-19 virus is still rampant around the world and the most effective way to combat it is through vaccination. As a result, the effectiveness and adverse events of the vaccine have become a global concern. Currently, there are three main types of COVID-19 vaccines licensed or in Phase III large-scale clinical trials in the United States.

1) mRNA-based two-dose vaccine series. These include Pfizer-BioNTec (BNT162b2), and Moderna (mRNA-1273). (The US Food and Drug Administration (FDA) issued an Emergency Use Authorization (EUA) for both in December 2020).

2) Viral vector-based vaccines. Includes the two-dose AstraZeneca vaccine, and the Johnson Johnson (J & J/Janssen) COVID-19 one-dose series for which the FDA issued an EUA in February 2021.

3) Protein subunit vaccines that are not being promoted at this time (e.g. the CHO cell vaccine that has been approved in China).

With the exception of the protein subunit vaccine, which has a small amount of data, the other two vaccines have gradually generated many reports of post-vaccination local and systemic adverse reactions after widespread implementation and have been submitted to the Vaccine Adverse Event Reporting System (VAERS)[1], which has become the source of data for our project. As of September 10th in 2021, over 150000 cases were reported. Although only a small percentage, some cases of death or disability occurred.
While the official challenges and experiments continue, we aim to monitor and even evaluate the adverse events of these vaccines through the data that come to light, to identify in advance those vaccinees who are likely to have serious adverse events so that they do not miss the best time for treatment, and to try to predict the vaccinees who develop adverse events with a greater need for hospitalisation so that the optimal and least costly option is available when the wards are tight. This will also enhance our knowledge of the safety issues of the new crown vaccine and give the public a more dimensional reference for vaccination.

2.2 Methods

This project proposes to use the data obtained from VAERS to:

1) Predict when adverse events will occur in the body of the vaccine recipient following types of vaccine, and identify triggers and people at risk (or give advice on which vaccine is more appropriate for the prospective vaccinees).

2) Determine the severity of symptoms and predict the vaccinees most likely to require hospitalisation.

Therefore, in order to achieve the project objectives, the methods of our project are broadly as follows.

2.2.1 Data Processing

a. Dataset
The dataset to be used in this project is from VAERS, a national early warning system established in 1990 to detect possible safety problems in licensed vaccines in the United States and jointly administered by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA). VAERS provides an open source annual dataset of CSV files for download, and gives each reporter ID ("VAERS_ID"), type of vaccination and various "SYMPTOMS" generated after vaccination in the dataset. The project selected adverse reaction records related to COVID-19 vaccination in the 2021/01/01-2021/09/10 time period of this dataset file for analysis.

b. Labeled data
Due to the large, sparse and dimensional nature of the data in the dataset, we propose to create a dictionary for these 'SYMPTOMS' in this project, which will be used to mark whether the data occurred or not(e.g. 1: occurred, 0: did not occur, or T: occurred, F: did not occur, etc.), forming an array of vaccinees adverse events symptoms and storing them in a list.

2.2.2 Methods for Hospitalization Prediction

To solve the problem of high-dimensional sparse features, we propose to use the common dimensionality reduction methods (such as Principal Component Analysis (PCA), Naive Bayes, Linear Discriminant Analysis (LDA), Locally linear embedding (LLE), Laplacian Eigenmaps) to reduce the dimensionality of original data and use the processed data to compute the risk of hospitalization.Specifically, we plan to use the SVM linear classifier to calculate the risk of hospitalization and use a plain Bayes classifier and Laplace smoothing for the Bernoulli distribution model to handle the zero probability problem.In addition, significant symptoms are selected based on posterior probabilities.[2, 6]

2.2.3 Methods for Onset Time Rrediction

a. Regularized regression
Lasso and Ridge. The optional regularization parameters are selected by 10-fold cross-validation. Variable importance is evaluated according to the same logic as above.

b. Random forest[3]
The number of trees depends on the available data, and the number of predictors sampled per node changes accordingly. The importance of the variables is assessed by the average decrease in accuracy.

c. Artificial neural network
The predicted values are compared with the actual values and the weights are adjusted so that the error sum of squares is minimized. And the activation function is used afterwards. The use of neural networks enables better prediction of the time to occur of adverse events and identification of key predictors.

d. Cluster
Firstly, an unsupervised pre-classification (K-Means)[4, 5] is done according to the disease to get the hypothetical five clusters A,B,C,D,E. The cluster class centers of the obtained clusters are the mean values of the disease features, but we assume that patients with similar diseases have similar onset times. After that, the previous cluster class centers of disease features are removed and replaced with the shortest time or mean time within each cluster to obtain five time clusters A,B,C,D,E. Finally, a threshold is set to calculate the MSE error of onset time and five time cluster class centers for each data, and the data are reclassified according to the error, while if there exists data whose time error is not satisfied for each threshold, it is used as a new cluster class center.

2.3 Expected Results

The data of VAERS from January 1, 2021 to September 10, 2021 recorded a total of XXXX cases of adverse reactions provided by XXXX people.
In this paper, we aim to obtain the best prediction of disease severity (mild, severe requiring hospitalization) by comparing different machine learning dimensionality reduction algorithms such as PCA. Afterwards, we intend to select the optimal data effect as input for subsequent training by various machine learning classification algorithms (linear regression, random forest, etc.) to obtain certain symptoms of the most frequent adverse events, the type of vaccine most likely to have adverse events after vaccination and its manufacturer, and the specific probability and timing of adverse events after vaccination of the vaccinees.

References

[1] Vaccine adverse event reporting system. [online]. https://vaers.hhs.gov/data.html.

[2] Afnan M. Alhassan and Wan Mohd Nazmee Wan Zainon. Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. *IEEE Access*, 9:87310–87317, 2021.

[3] V. Jackins, S. Vimal, Madasamy Kaliappan, and Mi Young Lee. Ai-based smart prediction of clinical disease using random forest classifier and naive bayes. *J. Supercomput.*, 77(5):5198–5219, 2021.

[4] Rony Chowdhury Ripan, Iqbal H. Sarker, Syed Md. Minhaz Hossain, Md Musfique Anwar, Raza Nowrozy, Mohammed Moshiul Hoque, and Md. Hasan Furhad. A data-driven heart disease prediction model through k-means clustering-based anomaly detection. *SN Comput. Sci.*, 2(2):112, 2021.

[5] Aman Singh, Jaydip Chandrakant Mehta, Divya Anand, Pinku Nath, Babita Pandey, and Aditya Khamparia. An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and improved ensemble learning. *Expert Syst. J. Knowl. Eng.*, 38(1), 2021.

[6] Zhuoyuan Zheng, Yunpeng Cai, Yujie Yang, and Ye Li. Sparse weighted naive bayes classifier for efficient classification of categorical data. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 691–696. IEEE, 2018.