Predicting the Adverse Events Following Receipt of mRNA-Based COVID-19 Vaccines

Xiaojuan Liu
Department of Epidemiology
Stanford University
Stanford, CA 94305
xjliu@stanford.edu

Yirong Yang
Department of Radiology
Stanford University
Stanford, CA 94305
yryangd@stanford.edu

1 Introduction

In December 2020, the US Food and Drug Administration (FDA) issued the Emergency Use Authorization (EUA) for two mRNA-based COVID-19 vaccines (BNT162b2 from Pfizer-BioNTech and the mRNA-1273 vaccine from Moderna) as 2-dose series.[1, 2, 3]

Following implementation of vaccination, local and systemic adverse reactions after receipt of the vaccines began to be reported.[4, 5] As of April 22, 2021, reports of 150,395 (0.07%) adverse events after receipt of vaccine had been submitted to the Vaccine Adverse Event Reporting System (VAERS). Although being rare, some uncommon allergic reactions can develop and lead to death or disability. For example, from December 14 to 23, 2020, 1,893,360 people in US received their first dose of vaccine and 21 of them reported suffering from anaphylaxis.[4]

Continued monitoring and assessing adverse events of these vaccines outside of trial settings could improve our understanding on the safety issues and contribute to the decision-making in terms of the implementation and administration of vaccination. It is also crucial for optimal outcomes of patients to identify patients at risk of severe adverse events in a safe medical environment. Our goal is to take use of the VAERS data to:

- Predict the onset time of adverse event and recognize the key predictors to inform the medication preparation after vaccination and identify the high risk population.
- Identify the most predictive symptoms for the patients' need of hospitalization and treatment after vaccination for patients' self-check and severity prediction.

For the first purpose, we used the baseline characteristics of patients, e.g. sex, age, medication history, etc. as inputs, and use different algorithms, e.g. linear regression, Lasso, Ridge, random forest, etc. to predict the onset time.

For the second purpose, we used the symptoms of patients as inputs, after algorithms dealing with sparse features, we gives the probability of a patient's needs of hospitalization.

2 Related Work

There are many works regarding the method for predicting vaccine outcomes and vaccine-associated adverse effects. For example, Gonzalez-Dia et al. [6] has provided a general procedure for predicting vaccine-induced immunity and reactogenicity using machine learning methods and described four basic steps including data processing, feature selection, choosing algorithm and testing.

Ahamad et al.[7] has conducted the identification and classification of post-vaccination reactogenicity of COVID-19 vaccination, using the same data source. They used decision tree and random forest, support vector machine and gradient boosting machine as classifiers to find the significant features leading to the hospitalization and death of patients. However, they pre-processed their data to solve

the sparse symptom feature problem, by which only 86 most frequently appeared symptoms were selected and combined. This method cannot perform well when encountered with high dimensional sparse features.

Several methods can be used to deal with the high dimensional sparse features along with severely imbalanced label. For example, we can use sparse PCA [8, 9] to extract the principle components of the sparse symptom feature, then apply a weighted logistic regression classifier to obtain the probability of hospitalization.Naive Bayes can be another promising algorithm for sparse features. A sparse naive Bayes algorithm [10] has also been proposed to solve the problem of selecting significant features. And we can further modify this algorithm with Laplace smoothing in our application.

3 Dataset and Features

VAERS is a national reporting system designed to detect early safety problems for licensed vaccines. Healthcare providers, vaccine manufacturers, and the public can submit reports to the system. VAERS provide open-source annually dataset to download and the reliability and validity of the data has been verified elsewhere.[5]

The total vaccination data in US was extracted from Our World in Data website from Dec 20, 2020 to April 22, 2021. VAERS 2020 and 2021 datasets (up to April 22, 2021) were used to select adverse events data associated with COVID-19 vaccinations only. Since one person may report multiple events, a unique 'VAERS ID' was used to identify each person.

The specific symptoms associated with each event in VAERS were encoded as MedDRA Terms from the standard MedDRA code book.[11] There are 5487 types of symptom (MedDRA term) in total, however, for each patient, they only have 5 symptoms reported averagely. The feature is of high dimension and very sparse. We created a dictionary of these symptom terms and assigned a array composed of 0 (not occur) and 1 (occurred) to each person indicating whether a specific symptom occurred to this person, and use a list of index to store the reported symptoms of each patient.

The baseline variables age, sex, vaccine manufacture, current illness, disability status, medication usage, allergic history, pre-existing conditions were of interest and encoded as continuous variables (standardized) or binary variables (0 or 1) correspondingly. Particularly, the pre-existing conditions were recorded as narrative text in VAERS, so we split each text and identified 17 most common conditions and transited them into 17 individual features (0,1 binary variable) for each person.

4 Method

4.1 Data Interpretation

Data was described by the incidence (rate) and the distribution of the adverse events following receipt of COVID-19 vaccinations in US. Number of adverse events were calculated by adding the people vaccinated on the same day in VAERS, and then this number was divided by the total vaccination on that day to get the rate data. Missing values of vaccination date in VAERS were imputed by the value of next record. Bar plot and line chart were used to describe the rate and distribution of the events.

4.2 Time of Onset Prediction

To predict the time of the onset of adverse event and identify the key predictors, we considered the first event record for each person. The interval (in days) were calculated as event onset date minus vaccination date (continuous). Predictors included all 27 (7+17) baseline variables. Data were split into training set and test set (8:2). Model was trained by a series of algorithms on training set (80% sample) as below and Prediction performance was evaluated by mean square error (MSE) on the test set (20% sample).

- Ordinary least square. Variable importance was assessed by the sign and magnitude of coefficients.
- Regularized regression (lasso and ridge). The optimal regularization parameter chosen by 10-fold cross validation. Variable importance was assessed by the same logic above.

- Random forest. The number of trees was set to be 500 and the number of predictors sampled for spliting at each node was set to be 8 (p/3). Variable importance was assessed by the mean decrease in accuracy.
- Neural network. For simplification, we used 2 hidden layers with 2 and 1 neurons in each of the layers and Sigmoid activation function. Variable importance was assessed by the weights of the first layer.

4.3 Needs of Hospitalization Prediction

We use machine learning approaches to classify the patients with needs for hospitalization for treatment after SARS-CoV-2 vaccination and find the significant symptoms.

In this task, the input features is a sparse encoded symptoms reported by patients. There are 5487 symptoms reported and encoded by MedDRA terms in total [11]. However, on average each patient only self-reported 5 symptoms resulting in a high dimensional and highly sparse input feature matrix.

In addition, since most patients with self-reported adverse effects have no needs for hospitalization, the output labels is highly imbalanced. Even if we predicted all the outcomes as no needs for hospitalization, the accuracy can still be up to 95%.

These two ill conditioned data structure should be taken into consideration in our method.

4.3.1 Dimensionality reduction with Sparse PCA

To solve the problem of high dimensional sparse feature, we first used sparse PCA [8, 9] to reduce the dimensionality of features from 5487 to 5. On the other hand, as the label is highly imbalanced and our focus was the significant symptoms that would lead to patients' needs for hospitalization, we only used the symptoms reported by patients with need for hospitalization to obtain the principle components.

Later, we used the transformation to obtain the low dimensional features of all training set samples, and logistic regression [12] with balanced class weights to compute the risk of hospitalization.

4.3.2 Sparse Feature Selection with Naive Bayes

We used naive Bayes classifier with sparse constraint [10] for Bernoulli distribution model and Laplace smoothing to tackle the zero probability problem. The risk of hospitalization is then computed. And the significant symptoms are selected according to the posterior probabilities.

5 Results

5.1 Data Interpretation

From Dec 20, 2020 to April 22, 2021, a total 118,746 persons reported adverse events in VAERS system producing 150,395 adverse events. Of all events, 71605 (47.6%) was reported on the first day of vaccination. The maximum duration between the event and vaccination date was 50 days.

Figure 1 shows the top 15 ranking symptoms of adverse event. The 5 most common symptoms are headache, pyrexia (aka fever), chills, fatigue, and pain respectively.

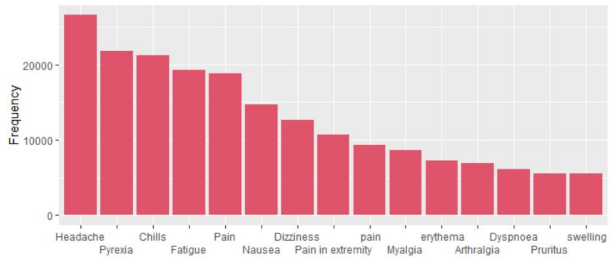

Figure 1: Frequency of the top 15 symptoms

Figure 2 shows the trend of adverse events overtime by the vaccine Manufacturer. Two peaks of adverse events showed up in late January and late April. There is no detectable difference between Moderna and Pfizer vaccination. Starting from May, most of the adverse events were associated with Janssen vaccination. Figure 3 shows rate of adverse event over time. The highest rate occurred at last December and early January, which was the time of the beginning of the vaccination implementation.

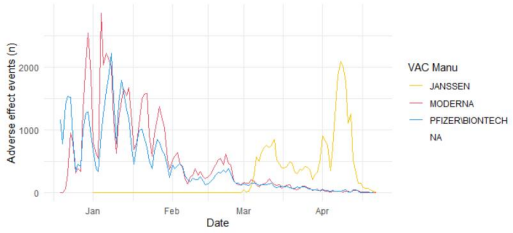

Figure 2: Trends of adverse event

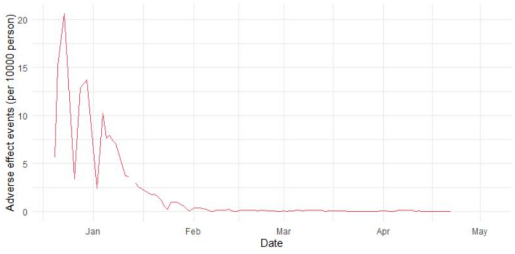

Figure 3: Rate of adverse event

5.2 Onset time Prediction

Table 1 shows the performance of different algorithms in predicting the time of the event occurrence. The linear regression (with or without penalization) produced an error of around 5 days and implied the best predictors for shorter duration of event onset were: **anxiety, depression, allergic history, cancer and diabetes**. By contrast, the predictors for longer duration of event onset were: **female sex, thyroid disorder, other medication usage, kidney disease, and anemia**. By applying the random forest and Neural network, the error reduced to less than 3 days. The random forest further showed that **age** and **other medications** were of great importance in prediction and the neural network also implied the importance of **dementia** and **kidney disease** in prediction.
Interestingly, there is evidence indicating that the Moderna was predictive of shorter onset time while Pfizer was more predictive of longer onset time. This distinction may be due to either the vaccine mechanism or the fact that waiting time between the first and the second dose of Pfizer is shorter than that of Moderna, allowing for more time for Moderna to develop adverse events.

Table 1: Predicting event time by different algorithms

	Training MSE	Test MSE	Best predictors for shorter duration	Best predictors for longer duration
OLS	5.188198	5.298035	Prevalent CVD Female sex Arthritis Hypertension Allergic history	Disability Hyperlipidemia Moderna manufacturer Obesity Depression
Lasso	5.189760	5.297949	Prevalent CVD Female sex Arthritis Asthma Allergic history	Disability Hyperlipidemia Moderna Obesity Depression
Ridge	5.215689	5.297750	Prevalent CVD Pfizer manufacturer Arthritis Asthma Allergic history	Disability Hyperlipidemia Moderna manufacturer Obesity Depression
Random forest	2.709732	2.900811	Age, Allergic history, Other medication use, Disability, Asthma	
Neural Network	2.513042	2.557245	Disability, Dementia, Kidney disease Hyperlipidemia, Allergic history	

5.3 Needs of Hospitalization Prediction

The results of each method are shown in Table 2. The ROC curves of each method are shown in Figure 4 and Figure 5. Both the two methods have good performance in ruling in the patients with needs for hospitalization (high specificity). However, both of them have low sensitivity, which means high rate of false negative.

The top 20 significant symptoms leading to hospitalization obtained by sparse PCA and logitstic regression method were: **Death, Pain, Chills, Malaise, Nausea, Vomiting, Pyrexia, Fatigue, Asthenia, Dizziness, Diarrhoea, Abdominal pain, Pain iscisnessn extremity, Myalgia, Tachycardia, Arthralgia, Dehydration, Loss of consciousness, Hyperhidrosis, Decreased appetite**.

And the top 20 significant symptoms leading to hospitalization obtained by sparse naive Bayes method were: **Chills, Dyspnoea, Injection site pain, Pain, Injection site swelling, Nausea, Rash, Arthralgia, Injection site erythema, Myalgia, Dizziness, Fatigue, Pain in extremity, Pyrexia, Headache, Pruritus, Vomiting, Paraesthesia, Injection site pruritus, Asthenia**.

There are 11 overlapped symptoms of these two methods: **Pain in extremity, Nausea, Pyrexia, Chills, Fatigue, Pain, Arthralgia, Dizziness, Vomiting, Myalgia, Asthenia**.

The sparse naive Bayes method successfully extracted the frequently occurred symptoms but failed to specified the symptoms that will lead to hospitalization. On the contrary, as the sparse PCA method obtained the principle components from the samples in needs for hospitalization, it successfully extracted some of the severe adverse side effect that would lead to hospitalization.

Table 2: Prediction sensitivity and specificity of hospitalization needs

	Sparse PCA and logistic regression	Sparse naive Bayes
Optimal probability threshold	0.46	0.03
AUC	0.7625	0.5327
Training set sensitivity	0.5595	0.1581
Training set specificity	0.8699	0.9753
Validation set sensitivity	0.5570	0.4638
Validation set specificity	0.8651	0.9289

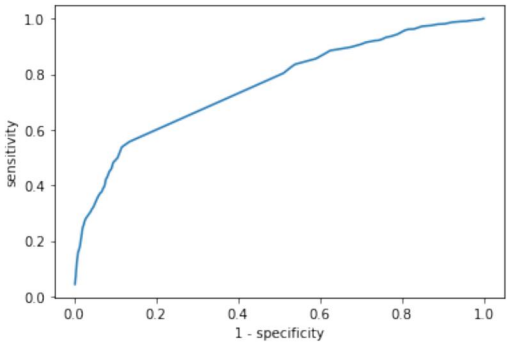

Figure 4: ROC curve: sparse PCA and logistic regression method

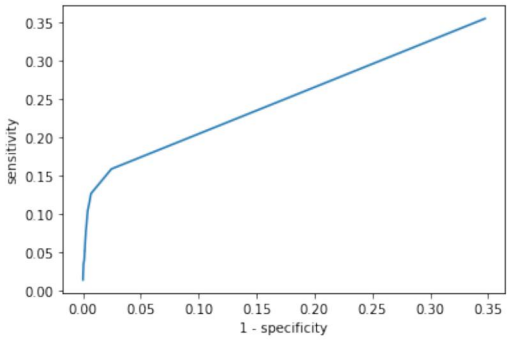

Figure 5: ROC curve: sparse naive Bayes method

6 Conclusion and future work

In this study, the neural network outperformed other algorithms in predicting the event onset time and this may due to the hidden layer exploited the interactions between predictors and improve the prediction performance.

Machine learning with high dimensional sparse features with highly imbalanced labels is always a challenging problem. In this study, sparse PCA outperforms naive Bayes. In future work, we will keep exploring machine learning method that can handle high dimensional sparse feature effectively.

7 Member contributions

Xiaojuan worked on goal 1. Yirong worked on goal 2. We both worked on the data processing, results interpreting and manuscript drafting.

Github link `https://github.com/XiaojuanLiu/CS229-final-project.git`

References

[1] C Buddy Creech, Shannon C Walker, and Robert J Samuels. Sars-cov-2 vaccines. *Jama*, 325(13):1318–1320, 2021.

[2] Sara E Oliver, Julia W Gargano, Mona Marin, Megan Wallace, Kathryn G Curran, Mary Chamberland, Nancy McClung, Doug Campos-Outcalt, Rebecca L Morgan, Sarah Mbaeyi, et al. The advisory committee on immunization practices' interim recommendation for use of pfizer-biontech covid-19 vaccine—united states, december 2020. *Morbidity and Mortality Weekly Report*, 69(50):1922, 2020.

[3] Sara E Oliver. The advisory committee on immunization practices' interim recommendation for use of moderna covid-19 vaccine—united states, december 2020. *MMWR. Morbidity and mortality weekly report*, 69, 2020.

[4] Tom Shimabukuro and Narayan Nair. Allergic reactions including anaphylaxis after receipt of the first dose of pfizer-biontech covid-19 vaccine. *JAMA*, 325(8):780–781, 2021.

[5] CDC COVID and Response Team. Allergic reactions including anaphylaxis after receipt of the first dose of moderna covid-19 vaccine—united states, december 21, 2020–january 10, 2021. *Morbidity and Mortality Weekly Report*, 70(4):125, 2021.

[6] Patrícia Gonzalez-Dias, Eva K Lee, Sara Sorgi, Diógenes S de Lima, Alysson H Urbanski, Eduardo Lv Silveira, and Helder I Nakaya. Methods for predicting vaccine immunogenicity and reactogenicity. *Human vaccines & immunotherapeutics*, 16(2):269–276, 2020.

[7] Md Martuza Ahamad, Sakifa Aktar, Md Jamal Uddin, Md Rashed-Al-Mahfuz, AKM Azad, Shahadat Uddin, Salem A Alyami, Iqbal H Sarker, Pietro Liò, Julian MW Quinn, et al. Adverse effects of covid-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. *medRxiv*, 2021.

[8] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009.

[9] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373. JMLR Workshop and Conference Proceedings, 2010.

[10] Armin Askari, Alexandre d'Aspremont, and Laurent El Ghaoui. Naive feature selection: Sparsity in naive bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1813–1822. PMLR, 2020.

[11] Patricia Mozzicato. Meddra. *Pharmaceutical Medicine*, 23(2):65–75, 2009.

[12] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.