GR5291 Final Project Report-Data Mining of COVID-19 Vaccine ADR Reports

Group members:

Roujia Dong (rd2962), Yiru Qian (yq2307), Yinru Shen (ys3478),

Huiyan Wang (hw2805), Yarong Wang (yw3598)

Abstract

Objectives:

(i) to implement three mainstream signal detection methods;**(ii)** to implement clustering and stratified PRR to deal with confounding;**(iii)** to compare drug-symptom pairs that are highlighted by PRR, for both unstratified and stratified data

Methods: We applied PRR, ROR and BCPNN data mining methods to data from the US Vaccine Adverse Event Reporting System (VAERS), and aggregated the results. We stratified PRR by age and sex using 2 clustering methods. To study the effects of stratification, we compared the unstratified PRR and stratified PRR.

Results: The signal detection results from these methods were somehow similar, but different in the ranking details. Therefore, an aggregated list of top signals based on distance metric optimization were generated. The final result of the aggregated top-ranked vaccine-event pairs showed that the majority of the detected signals were from Moderna COVID-19 vaccine, and half of the symptoms occurred at the vaccination site, which are basically similar to other vaccines that have already been widely used. Some of the detected signals are found to be mainly due to human errors or logistical issues, which indicates the need to improve the vaccine administration process in the future. Data quality issues are also implied by signals as well. Our findings may help guide improvements in future reporting data quality and data ingestion process.

Stratification not only changed the number of vaccine-event pairs that were highlighted, but it also changed which pairs were highlighted. There were 637 vaccine-event pairs that were highlighted by the unstratified PRR; 222 that were highlighted by the stratified PRR; and 190 that were highlighted by both. In addition, applying different clustering methods produces similar results with regards to the most significant drug-symptom combinations.

1. Objective

Covid-19 vaccinations are being populated around the world to fight against virus that had tortured human beings for almost two years. With the limited time for vaccination research and development, current vaccination is imperfect and people sometimes suffer from adverse events after injecting the vaccine. Thus, our goal is to make analysis through adverse reports to investigate the association between vaccination and adverse events, and thus provide insightful suggestions for clinications. In our project, we analyze 2 types of Covid-19 vaccines (Moderna/Pfizer) with their adverse effects using data mining techniques such as signal detection methods, which is the computational methods combining statistics, computer science, epidemiology, medicine, cheminformatics, and biology that can translate data into meaningful knowledge.

The aggregated results are generated through exploration of three objectives:
(i) to implement three mainstream signal detection methods
(ii) to implement stratified PRR to deal with confounding
(iii) to compare vaccine-event pairs that are highlighted by PRR, for both crude and stratified.

2. Material and Methods

2.1. Data Source and Structure[1]

Data is collected from VAERS (Vaccine Adverse Event Reporting System). It is the repository for adverse events that are reported after vaccinations by health care providers, vaccine recipients and other interested parties, and by manufacturers that are required by regulation. VAERS is sponsored by Centers for Disease Control and Prevention (CDC), and the Food and Drug Administration (FDA), agencies of the U.S. Department of Health and Human Services.

VAERSDATA.CSV contains detailed description of the data including vaccinator's basic information such as age in years, sex, symptoms and whether they have adverse effects such as death, life-threatening illness, disability and etc. VAERSVAX.CSV includes the remaining information of the vaccine such as vaccine name, manufacturer, number of doses administered and etc. VAERSSYMPTOMS.CSV includes adverse event coded terms utilizing the MedDRA dictionary, which is a system converting medical conditions described in the data into standardized code and each vaccinator has 5 columns for 5 different symptoms.

2.2. Data Preparation

The dataset used in this project covers the adverse events reported by September 23rd, 2021. Based on project objective and the fact that Moderna and Pfizer are the most commonly vaccinated COVID-19 vaccines in the United States and they are made using similar technology (mRNA), we are only exploring COVID-19 vaccine from Pfizer and Moderna, either first or second dose, no booster. Meanwhile, only records where SEX is either female or male are retained. As for SYMPTOM, we assume that "Vaccination site" is the same as "Injection site", and replace all "Injection site" with "Vaccination site". All symptoms of the same VAERS_ID are integrated into one record (row).

The three datasets (i.e. VAERSDATA.CSV, VAERSVAX.CSV and VAERSSYMPTOMS.CSV) are inner joined by VAERS_ID. In order to avoid the impact of the interaction of vaccines from different manufacturers on the accuracy of our analysis, VAERS_IDs that got two different COVID-19 vaccines for first and second doses are detected, and those VAERS_IDs' records are

removed. Finally, five variables: 'VAERS_ID', 'CAGE_YR', 'SEX', 'VAX_MANU' and 'SYMPTOM' are selected, and all NA values are dropped since each remaining variable must have a real value for further analysis. All variables used in the following analysis are shown in Table 1.

Table 1. Summary of fields used

Variable name	Type	Description of Contents
VAERS_ID	Num (7)	VAERS identification number
CAGE_YR	Num (xxx)	Age in years
SEX	Char (1)	Sex (either M or F)
VAX_MANU	Char (40)	Vaccine manufacturer (either MODERNA or PFIZER\BIONTECH)
SYMPTOM	Char(1000)	Adverse event MedDRA term

2.3. Analysis Methods

2.3.1. Disproportionality analysis measures[2]

Disproportionality analysis measures are used to figure out combinations of drug exposures and adverse drug reactions (ADRs), which often occur disproportionately, when compared to other drug-event combinations. By using data from any source, signal detection finds out and/or identifies signals, which follows a methodology taking into consideration the nature of data and the characteristics such as patient exposure, target population, and time on market. Also, the type of medicinal product needs to be considered as well like vaccines and products in biological medicine. Disproportionality measures can generally be divided into two categories: frequentist and Bayesian. The most popular frequentist methods are the Proportional Reporting Rate (PRR) and Reporting Odds Ratio (ROR). And as for Bayesian approaches, the Bayesian Confidence Propagation Neural Network (BCPNN) is one of the most widely used techniques.

In the following sections, we will have a more detailed introduction on these methods.

(1) Proportional Reporting Ratio (PRR) [3]

PRR is a simple way to get a measure of how common an adverse event for a particular drug is compared to how common the event is in the overall database. Using this method, we could measure the strength of the statistical association between a risk factor and a condition. For

example, we could measure the strength of the statistical association between specific vaccines and adverse events.

PRR allows the comparison of frequencies of reporting, in order to determine whether there is a disproportionate reporting of some specific adverse event with specific vaccines compared with others. When using this method, it is common to use the contingency table listed below:

Table 2. Contingency table of PRR.

	Event of Interest	All Other Events	TOTAL
Product of Interest	a	b	**a+b**
All Other Products	c	d	**c+d**
TOTAL	**a+c**	**b+d**	**a+b+c+d**

b is all reports for all other adverse events for product
c is all reports for all other products for adverse event
d is all reports for all other products for all other adverse event
c+d is all reports for all other products

According to above information, we could calculate PRR as:

$$PRR = \frac{a/(a+b)}{c/(c+d)}$$

If a/(a+b) is greater than c/(c+d), the event is disproportionately reported for the product. It means, when PRR is higher than 1, it indicates a higher probability that the event happened compared to other situations. The Corresponding lower and upper bounds of the PRR's confidence interval is:

$$PRR(-) = e^{ln(PRR)-1.96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}}$$

$$PRR(+) = e^{ln(PRR)+1.96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}}$$

The advantage of PRR is that it allows for the magnitude of the effect to be assessed, while its disadvantage is that it does not estimate relative risk.

(2) Reporting Odds Ratio (ROR)[4]

ROR stands for the Reporting Odds Ratio, which is the comparison between the odds of a certain event occurring with a medicinal product and the odds of the same event occurring with all other medicinal products. When the lower limit of the 95% confidence interval is larger than one, a signal is considered. For instance, if the ROR equals two, then it means that the odds of reporting

this certain event with a medicinal product are two times higher than the odds of reporting among all the others.

Using the same contingency table as PRR in Table 2, the ROR can be calculated with the formula as below:

$$ROR = \frac{ad}{bc}$$

The Corresponding lower and upper bounds of the ROR's confidence interval is:

$$ROR(-) = e^{ln(ROR)-1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

$$ROR(+) = e^{ln(ROR)+1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

Through estimating relative risk, ROR may have an advantage over PRR theoretically. But the advantage has been called to doubt, because risk comparison cannot be allowed by disproportionality measures. ROR and PRR only provide a general signal strength indication.

The reporting of an event's baseline risk is reflected by this reference risk. However, a question being asked is if the reference risk can often accurately predict an event baseline risk when a patient is receiving any type of drugs, because it is considered a reference group of drugs that could include drugs at risk for a certain event.

(3) Bayesian Confidence Propagation Neural Network (BCPNN)[5]

The basic BCPNN model is a feedforward neural network that learns knowledge and evaluates an action or event using the principles of Bayes' law. The independence assumptions are the same as in naïve Bayes formalism. In order to allow the use of the standard equation for propagating activity between neurons, transformation to log space was necessary. BCPNN has been used for machine learning classification and data mining, for example for discovery of adverse drug reactions.

In this project, drugs and ADR events are used as the input and output of the neural network, respectively. Again, using the same contingency table in Table 2, the information criterion (IC), which is used as the weight of the network and measures the association between the vaccine and symptom, is derived using a calculation based on information theory as below:

$$IC = log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

In general, the standard BCPNN is used to detect suspicious ADR in the following situations: If the lower 95% confidence limit of IC is greater than zero, then it indicates an ADR signal. If 0 < lower 95% CI $\leqslant$ 1.5, then it indicates a weak signal; if 1.5 < lower 95% CI $\leqslant$ 3.0; then it indicates a medium-intensity signal; if lower 95% CI > 3.0; then it indicates a strong signal.

The advantage with this procedure is that it is fast and due to the linear transform those units which only participate in background noise will be set to zero or close to zero. However, the disadvantage is that it may in some cases miss patterns which overlap, that is patterns who share units. This can, however, probably be dealt with using a somewhat more thorough exploration of the stimulus space around the patterns found.

2.3.2 Rank Aggregation for ensembled signal detection[6][7]

The disproportionality analysis measures introduced in section 3.1are three mainstream signal detection methods that have already been used in real-world pharmacovigilance monitoring by regulatory agencies globally. Those methods are somehow similar, but result in different ranking details. Therefore, a new signal detection method that can integrate the vaccine safety signals detected from any signal detection methods is developed to obtain an integration of the results of these approaches.

Based on distance metric optimization, the algorithm generates an aggregated list of top-ranked signals that would be simultaneously as close as possible to the lists produced by each measure. Rank aggregation is applied to create an ensembled safety signal list using the Spearman's footrule distance and genetic algorithm. The idea can be formalized as an optimization problem:

$$\delta^* = \arg\min \sum_M d(\delta, L_M)$$

where L_M is the ordered list produced by disproportionality analysis measure M, the minimization is carried over all possible ordered lists δ of size $k = |L_M|$, and d is an appropriate distance function that can measure the distance between rankings (i.e. Spearman's footrule distance). The Spearman's footrule distance here is modified by including certain weights to reflect variation of

the scores under each measure M for different ranked positions. The weighted Spearman's footrule distance between L_M and any ordered list of k algorithms is given by:

$$d(\delta, L_M) = \sum_{t \in L_M \cup \delta} |M(r^{\delta}(t)) - M(r^{L_M}(t))| \times |r^{\delta}(t) - r^{L_M}(t)|$$

The goal is to find a δ^* that minimizes the total distance between δ and L_M.

This method can potentially enhance the weak signals from individual methods while reducing the number of false-positive signals that are accidentally discovered.

2.3.3 Clustering Method[8]

In order to deal with confounding with respect to age and gender, we applied 2 clustering methods, one is manually dividing the people into 4 groups:

- Female and Age <40,
- Female and Age >=40,
- Male and Age<40,
- Male and Age >=40.

Within each group, the number of people is 137025, 296881, 54264, 108849 respectively.

In addition, we plan to implement the K-means algorithm method which aims for identifying k numbers of centroids and then gathers every data point to the nearest cluster, while at the same time keeping the centroids as small as possible. In order to determine the number of clusters K, we implement the elbow method as below. The Elbow method stands for determining the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, the "elbow" is a good indication that the underlying model fits best at that point. At this point, we choose K=15 as the number of clusters for this dataset. Hence, by using kmean function in R, we divide the data into 15 clusters with each cluster's conclusion in the Result part.

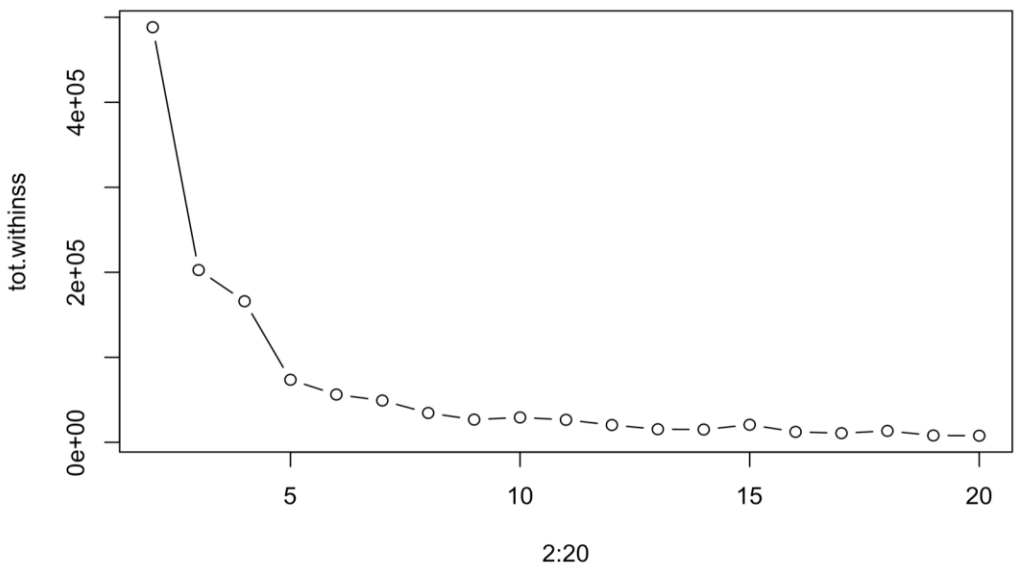


Figure 1. Elbow method.

2.3.4. Mantel-Haenszel Adjustments[9]

In order to obtain a pooled PRR from the stratification groups, we applied the Mantel-Haenszel algorithm. We first calculate the PRR for each group, where the groups are generated by the clustering methods. Then we calculate the stratified PRR using the following formula:

If there are K groups, then for group k (k = 1, 2, ..., K), let a_k, b_k, c_k, d_k be the four cell counts in group k. More specifically,

- a is the number of cases with the specific event and the specific drug.
- b is the number of cases with the specific event but not the specific drug.
- c is the number of cases with the specific drug but not the specific event.
- d is the total number of cases with neither the specific event nor the specific drug.

And then let

- M1k = ak + bk
- M0k= ck + dk
- N1k = ak + ck
- N0k = bk + dk
- T = M0k + M1k = N0k + N1k = ak + bk + ck + dk

In this way we can calculate the stratified PRR = $(\Sigma_K a_k N0_k/T_k) / (\Sigma_K b_k N1_k/T_k)$

3. Results

Main Findings

3.1. Signal Detection

The signal detection analysis methods (PRR, ROR, BCPNN) are implemented in the "PhViD" package with the processed dataset using statistical software R. There are 14255 pairs of vaccines and the adverse events in the cleaned dataset, after eliminating those whose marginal count is too small that is less than 100, 1942 different vaccine-adverse event pairs are left for the following signal detection.

The PRR method detects 637 vaccine-adverse event pairs that are significant signals, and details of the top 20 pairs ranked by the lower bound of the 95% two-sided confidence interval for log(PRR) is shown in Table 3. A relatively higher PRR indicates stronger association between the vaccine and the symptom. The result shows that "Product dose omission issue" ranks the top and it also has the highest PRR, indicating that this adverse event is more commonly reported for individuals taking Moderna, relative to Pfizer.

Table 3. Top 20 vaccine-event pairs detected by PRR.

	drug code	event effect	count	expected count	LB95(log(PRR))	PRR	drug margin	event margin
1	MODERNA	Product dose omission issue	1072	627.47467	4.2734452	191.569151	941212	1076
2	MODERNA	Body temperature	1950	1153.48038	3.5346070	49.781429	941212	1978
3	PFIZER\BIONTECH	Product preparation issue	1153	501.88151	3.1735985	31.627754	672788	1204
4	PFIZER\BIONTECH	Product preparation error	268	115.04925	3.1440721	46.865583	672788	276
5	MODERNA	Maternal exposure during pregnancy	553	330.06567	2.8647309	30.406929	941212	566
6	MODERNA	Off label use	173	101.46895	2.8519355	123.662176	941212	174
7	MODERNA	Vaccination site pruritus	15990	10192.96440	1.9851080	7.676169	941212	17479
8	MODERNA	Vaccination site induration	4449	2826.55177	1.9757479	7.990429	941212	4847
9	MODERNA	Vaccination site rash	9188	5862.45616	1.9575490	7.592690	941212	10053
10	MODERNA	Vaccination site warmth	11860	7586.84518	1.9372188	7.371869	941212	13010
11	PFIZER\BIONTECH	Adenovirus test	124	59.19200	1.7712957	9.637367	672788	142
12	MODERNA	Vaccination site erythema	22035	14361.93875	1.7635073	6.074371	941212	24628
13	MODERNA	Intentional product use issue	100	61.23126	1.7618313	14.296205	941212	105
14	PFIZER\BIONTECH	Human metapneumovirus test	88	42.10136	1.6657683	9.469969	672788	101
15	MODERNA	Vaccination site hypersensitivity	111	69.39543	1.5768686	9.917992	941212	119
16	MODERNA	Vaccination site urticaria	1171	770.34762	1.5492975	5.580285	941212	1321
17	MODERNA	Vaccination site inflammation	770	504.42898	1.5436663	5.793725	941212	865
18	MODERNA	Pregnancy test	452	295.07638	1.5067772	5.983227	941212	506
19	MODERNA	Vaccination site cellulitis	265	171.44754	1.4933548	6.531887	941212	294
20	MODERNA	Vaccination site reaction	3510	2348.94792	1.4854593	4.843599	941212	4028

The ROR method detects the same number of vaccine-event pairs as ROR, which is 637 significant signals, and details of the top 20 pairs ranked by the lower bound of the 95% confidence interval for log(ROR) is shown in Table 4. The result listed in Table 4 is very similar to that of Table 3 using PRR, and a relatively higher ROR indicates stronger association between the vaccine and the symptom.

Table 4. Top 20 vaccine-event pairs detected by ROR.

	drug code	event effect	count	expected count	LB95(log(ROR))	ROR	drug margin	event margin
1	MODERNA	Product dose omission issue	1072	627.47467	4.2745689	191.786449	941212	1076
2	MODERNA	Body temperature	1950	1153.48038	3.5366130	49.882704	941212	1978
3	PFIZER\BIONTECH	Product preparation issue	1153	501.88151	3.1752246	31.680333	672788	1204
4	PFIZER\BIONTECH	Product preparation error	268	115.04925	3.1444481	46.883861	672788	276
5	MODERNA	Maternal exposure during pregnancy	553	330.06567	2.8652814	30.424217	941212	566
6	MODERNA	Off label use	173	101.46895	2.8521128	123.684726	941212	174
7	MODERNA	Vaccination site pruritus	15990	10192.96440	1.9998419	7.791549	941212	17479
8	MODERNA	Vaccination site induration	4449	2826.55177	1.9797986	8.023629	941212	4847
9	MODERNA	Vaccination site rash	9188	5862.45616	1.9659315	7.657681	941212	10053
10	MODERNA	Vaccination site warmth	11860	7586.84518	1.9480266	7.453184	941212	13010
11	MODERNA	Vaccination site erythema	22035	14361.93875	1.7830934	6.196017	941212	24628
12	PFIZER\BIONTECH	Adenovirus test	124	59.19200	1.7714411	9.638959	672788	142
13	MODERNA	Intentional product use issue	100	61.23126	1.7619192	14.297618	941212	105
14	PFIZER\BIONTECH	Human metapneumovirus test	88	42.10136	1.6658685	9.471077	672788	101
15	MODERNA	Vaccination site hypersensitivity	111	69.39543	1.5769610	9.919044	941212	119
16	MODERNA	Vaccination site urticaria	1171	770.34762	1.5502618	5.585991	941212	1321
17	MODERNA	Vaccination site inflammation	770	504.42898	1.5442976	5.797650	941212	865
18	MODERNA	Pregnancy test	452	295.07638	1.5071426	5.985621	941212	506
19	MODERNA	Vaccination site cellulitis	265	171.44754	1.4935677	6.533445	941212	294
20	MODERNA	Vaccination site reaction	3510	2348.94792	1.4883191	4.857986	941212	4028

The BCPNN algorithm detects 429 vaccine-event pairs as significant signals, which is less than ROR and PRR. Details of the top 20 pairs ranked by the 0.025 quantile of log(IC) is shown in Table 5. The “Pfizer-Product preparation issue” has the highest 0.025 quantile of log(IC), but since this highest lower bound is 1.08 <1.5, the signals detected by BCPNN are not so strong.

Table 5. Top 20 vaccine-event pairs detected by BCPNN.

	drug code	event effect	count	expected count	Q_0.025(log(IC))	n11/E	drug margin	event margin	FDR	FNR	Se	Sp	postH0
1	PFIZER\BIONTECH	Product preparation issue	1153	501.88151	1.0806575	2.297355	672788	1204	1.504794e-90	0.4973624	0.001035861	1.0000000	1.504794e-90
2	PFIZER\BIONTECH	Product preparation error	268	115.04925	0.9653560	2.329437	672788	276	3.791355e-23	0.4971033	0.002071722	1.0000000	7.582709e-23
3	PFIZER\BIONTECH	Product storage error	2732	1489.80441	0.8020985	1.833798	672788	3574	2.527570e-23	0.4968440	0.003107583	1.0000000	5.655412e-126
4	PFIZER\BIONTECH	Adenovirus test	124	59.19200	0.6967210	2.094878	672788	142	4.637874e-10	0.4965843	0.004143444	1.0000000	1.855150e-09
5	PFIZER\BIONTECH	Electrocardiogram ST segment elevation	340	185.07923	0.6665011	1.837051	672788	444	3.710299e-10	0.4963244	0.005179305	1.0000000	2.760360e-17
6	MODERNA	Body temperature	1950	1153.48038	0.6660498	1.690536	941212	1978	3.091916e-10	0.4960643	0.006215166	1.0000000	5.423708e-61
7	PFIZER\BIONTECH	Product temperature excursion issue	197	102.12705	0.6642729	1.928970	672788	245	2.658972e-10	0.4958038	0.007251027	1.0000000	6.130785e-12
8	MODERNA	Product dose omission issue	1072	627.47467	0.6484184	1.708436	941212	1076	2.326601e-10	0.4955431	0.008286889	1.0000000	1.680807e-35
9	PFIZER\BIONTECH	Human metapneumovirus test	88	42.10136	0.6199178	2.090194	672788	101	4.964981e-08	0.4952822	0.009322749	1.0000000	4.449870e-07
10	MODERNA	Vaccination site pruritus	15990	10192.96440	0.6186215	1.568729	941212	17479	4.468483e-08	0.4950209	0.010358610	1.0000000	0.000000e+00
11	MODERNA	Vaccination site warmth	11860	7586.84518	0.6085403	1.563232	941212	13010	4.062257e-08	0.4947594	0.011394471	1.0000000	9.262226e-273
12	MODERNA	Vaccination site rash	9188	5862.45616	0.6072767	1.567261	941212	10053	3.723735e-08	0.4944976	0.012430332	1.0000000	9.562052e-214
13	PFIZER\BIONTECH	Incorrect product formulation administered	327	185.91292	0.6018189	1.758888	672788	446	3.437294e-08	0.4942356	0.013466193	1.0000000	7.136121e-15
14	MODERNA	Vaccination site induration	4449	2826.55177	0.5952921	1.574003	941212	4847	3.191773e-08	0.4939733	0.014502054	1.0000000	3.643193e-106
15	PFIZER\BIONTECH	Respiratory viral panel	235	130.47252	0.5948725	1.801146	672788	313	2.979041e-08	0.4937107	0.015537916	1.0000000	7.825542e-12
16	MODERNA	Vaccination site erythema	22035	14361.93875	0.5913534	1.534264	941212	24628	2.792851e-08	0.4934478	0.016573777	1.0000000	0.000000e+00
17	PFIZER\BIONTECH	SARS-CoV-2 test positive	7628	4954.62092	0.5808288	1.539573	672788	11886	2.628565e-08	0.4931846	0.017609638	1.0000000	1.239811e-190
18	PFIZER\BIONTECH	Enterovirus test negative	102	51.68879	0.5773501	1.973348	672788	124	4.435266e-08	0.4929212	0.018645498	1.0000000	3.514917e-07
19	PFIZER\BIONTECH	Mycoplasma test negative	86	42.51820	0.5715165	2.022663	672788	102	1.200286e-07	0.4926575	0.019681358	1.0000000	1.482196e-06
20	MODERNA	Maternal exposure during pregnancy	553	330.06567	0.5712262	1.675424	941212	566	1.140272e-07	0.4923936	0.020717219	1.0000000	4.515688e-18

3.2. Rank Aggregation for ensembled signal detection

The three methods detect a total of 429 significant signals in common, and only these signals are retained for subsequent analysis. After rearranging the detected signals in each method, the rank aggregation is implemented in the "RankAggreg" package with the above data using statistical software R. Figure 2 and Figure 3 show the final sample distribution of objective function scores and the rank aggregation by data, genetic algorithm and mean of the optimal list, respectively.

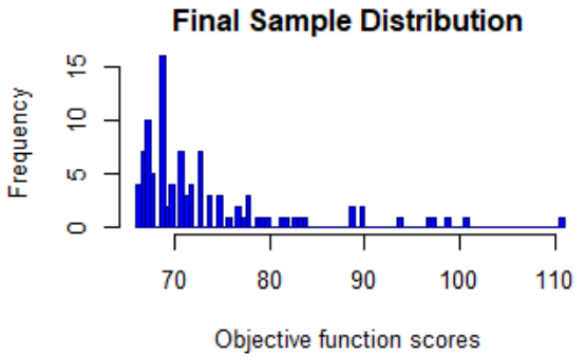


Figure 2. Final sample distribution of objective function scores.

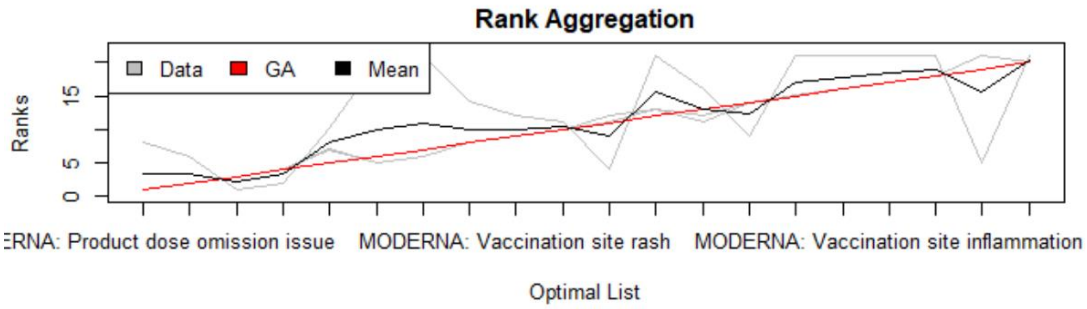


Figure 3. Rank aggregation of optimal list.

In order to keep the problem scale tractable, we only optimize and produce a top-20 optimal list shown in Figure 4:

```
 [1] "MODERNA: Product dose omission issue"
 [2] "MODERNA: Body temperature"
 [3] "PFIZER\\BIONTECH: Product preparation issue"
 [4] "PFIZER\\BIONTECH: Product preparation error"
 [5] "MODERNA: Maternal exposure during pregnancy"
 [6] "MODERNA: Off label use"
 [7] "MODERNA: Vaccination site pruritus"
 [8] "MODERNA: Vaccination site induration"
 [9] "MODERNA: Vaccination site rash"
[10] "PFIZER\\BIONTECH: Adenovirus test"
[11] "MODERNA: Vaccination site warmth"
[12] "MODERNA: Vaccination site erythema"
[13] "MODERNA: Intentional product use issue"
[14] "PFIZER\\BIONTECH: Human metapneumovirus test"
[15] "MODERNA: Vaccination site hypersensitivity"
[16] "MODERNA: Vaccination site urticaria"
[17] "MODERNA: Vaccination site inflammation"
[18] "MODERNA: Pregnancy test"
[19] "MODERNA: Vaccination site cellulitis"
[20] "MODERNA: Vaccination site reaction"
```

Figure 4. Aggregated top 20 vaccine-adverse event pairs.

The final result of the aggregated top-ranked vaccine-adverse event pairs shows that in general, no serious adverse events are detected. The majority of the top detected signals are from Moderna COVID-19 vaccine compared with Pfizer, and half of the symptoms occurred at the vaccination site (and they are all from Moderna), which are basically similar to other vaccines that have already been widely used. We also find that some of the detected signals might be mainly attributed to human errors or logistical issues, such as "MODERNA: Product dose omission issue", "PFIZER\\BIONTECH: Product preparation issue", "PFIZER\\BIONTECH: Product preparation error", etc., which indicates that the vaccine administration process may need to be improved in

the future. Data quality issues are also implied by signals such as "MODERNA: Body temperature", "MODERNA: Pregnancy test", etc. Those MedDRA terms are defined so broadly that it would be hard to interpret. These findings may help guide improvements in future reporting data quality and data ingestion process.

3.3. Stratified PRR

Although PRR, ROR and BCNPP provide meaningful analysis, the reduction of drug-symptom analysis to 2 dimensions may result in the loss of important information. Which means, these 2-D approaches are not properly equipped to deal with confounding, which is important to association analysis. A confounder is an extraneous variable which can mediate the association between specific drug-symptom combinations, leading to the discovery of spurious association and therefore wrong conclusions.

After applying the manual stratification, the results show that there are in total 637 vaccine-event pairs that were highlighted by the unstratified PRR; 222 that were highlighted by the stratified PRR; and 190 that were highlighted by both. The following table shows the top 5 stratified PRR and the corresponding unstratified PRR, we can find that they are all greater than 1 but stratified PRR and unstratified PRR differ in values slightly.

Table 6. Top stratified PRR and the corresponding unstratified PRR.

Vaccine name & symptom	Stratified PRR	Unstratified PRR
MODERNA Skin warm	3.1413546	2.702224
MODERNA Rash erythematous	2.3036337	2.251401
MODERNA Erythema	2.1379809	2.254016
PFIZER\\BIONTECH Echocardiogram normal	1.8372345	2.712597
PFIZER\\BIONTECH Presyncope	1.7731324	2.23916

For example, “MODERNA Skin warm” has stratified PRR 3.14, which is greater than the unstratified PRR (2.7). After further investigating the PRR within each cluster, we found that the PRR in cluster 1 and cluster 2 in general is greater than that in cluster 3 and cluster 4. In other

words, the occurrence of “skin warm” after receiving Moderna may be more frequent in Female groups than in Male groups.

Another example is the “PFIZER\\BIONTECH Presyncope”. In this case, the stratified PRR is smaller than the unstratified PRR. In cluster 1,2,3 the PRR is significant, but in cluster 4 (Male, age >40) this drug-symptom combination has PRR=1.2, indicating only slight importance.

Table 7. PRR within each cluster.

	Cluster 1 (F, age<40)	**Cluster 2 (F, age >=40)**	**Cluster 3 (M, age<40)**	**Cluster 4 (M, age >=40)**
MODERNA Skin warm	3.378388	3.295303	1.827508	2.04066
PFIZER\\BIONTECH Presyncope	2.257942	1.751989	1.637779	1.17283

3.3.1 Sensitivity Analysis

Considering the influence of clustering on stratified PRR, we applied k-means clustering and compared the results with the previous clustering method. The k-means clustering produces 15 clusters and the following figure demonstrates the clustering results, where the different colors represent different groups.

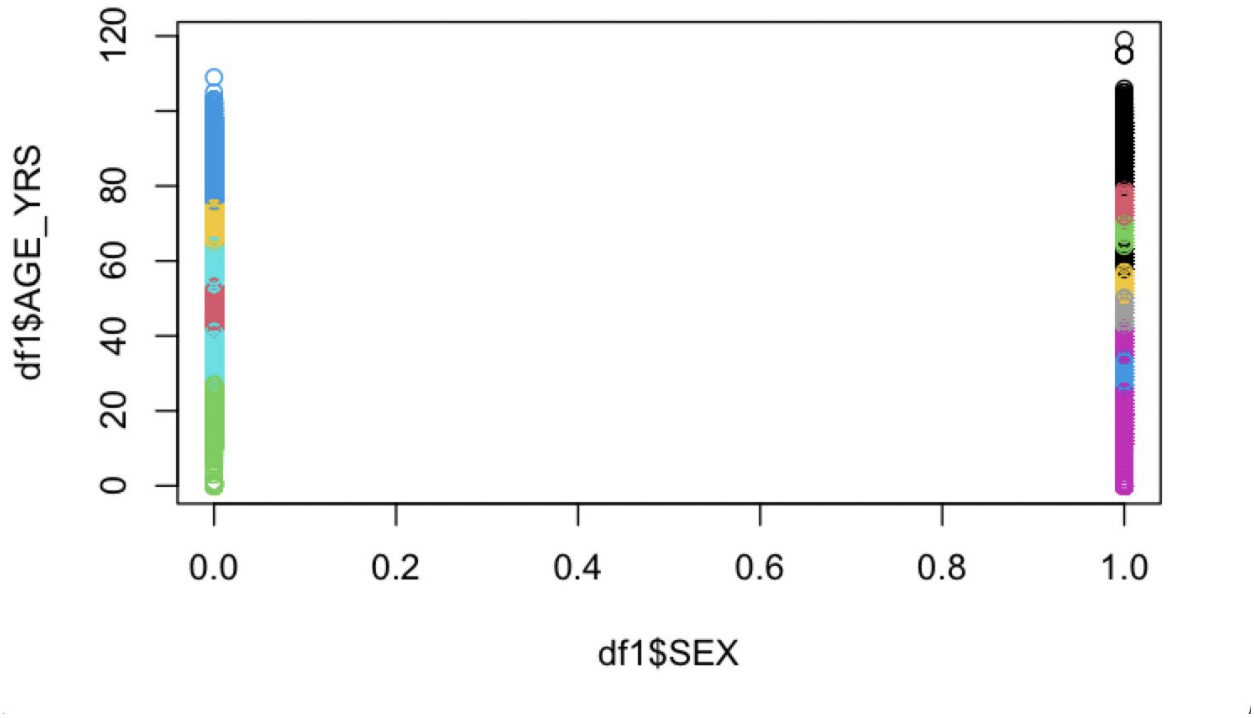


Figure 5. Clustering results.

After calculating the stratified PRR with the new clustering method, we find that there are 144 highlighted stratified PRR; and 118 drug-symptom combinations highlighted by both. Compared with the manual clustering, the top 3 drug- combinations are the same, indicating that the stratified PRR may be robust to the choice of clustering methods.

Table 8. Most significant drug-symptom combination.

Vaccine name & symptom	Stratified PRR	Unstratified PRR
MODERNA Skin warm	2.7824825	2.702224
MODERNA Rash erythematous	2.2058648	2.251401
MODERNA Erythema	2.0103638	2.254016
PFIZER\\BIONTECH White blood cell count normal	1.9617161	1.730932
PFIZER\\BIONTECH Haematocrit normal	1.953753	1.994831

4. Discussion

4.1. Significance of Results

A total of 429 vaccine-adverse event pairs are detected as significant signals by PRR, ROR and BCPNN methods in common. It seems that stratification can eliminate some drug-symptom combinations that occurred due to confounding, while most drug-symptom combinations highlighted by stratified PRR are also highlighted by unstratified PRR.

4.2. Limitation of Analysis, data, etc.

4.2.1. Limitation of Analysis

The text mining used to improve the quality of VAERS data is not implemented in depth ahead of our analysis. In this project we only considered age and gender as the confounders, while there are more complex confounders such as symptom-symptom effects that need to be considered. In addition, we can also explore the associations between symptoms using unsupervised learning methods such as network analysis.

4.2.2. Limitation of Data

The origin of the data in VAERS comes from the surveillance system and shows unsubstantiated reports of health events after vaccination. The data are limited to under-reporting, reporting bias and lack of incidence rates in unvaccinated groups. When reporting the data, there is no cause-and-effect relationship and it includes more than one vaccine with more than one reported adverse events. At some point, VAERS data needs more information from the reporter, vaccine provider, etc. In addition, only the first report received is included in the dataset when there are multiple reports of a single case. There is no guarantee that the data provided is the most accurate or current. When the report meets more than one criterion for classification, it is considered as “serious”. The reports do not allow incidence rate calculation due to uncertain extent of under-reporting and lack of information on the number of people being vaccinated.

4.3. Conclusion

In our project, we extracted significant drug-symptom combinations by different methods, hoping to provide insightful suggestions to medical researchers. The signal detection results from the stated signal detection methods are somehow similar, but different in the ranking details. Therefore, an aggregated list of top signals based on distance metric optimization are generated. The final result of the aggregated top-ranked vaccine-event pairs showed that the majority of the detected signals were from Moderna COVID-19 vaccine, and half of the symptoms occurred at the vaccination site, which are basically similar to other vaccines that have already been widely used. Some of the detected signals are found to be mainly due to human errors or logistical issues, which indicates the need to improve the vaccine administration process in the future. Data quality issues are also implied by signals as well. Our findings may help guide improvements in future reporting data quality and data ingestion process.

Stratification not only changed the number of vaccine-event pairs that were highlighted, but it also changed which pairs were highlighted. There were 637 vaccine-event pairs that were highlighted by the unstratified PRR; 222 that were highlighted by the stratified PRR; and 190 that were highlighted by both. In addition, applying different clustering methods produces similar results with regards to the most significant drug-symptom combinations.

5. Appendix

5.1. Details of EDA Analysis

In order to analyze and investigate data sets and summarize their main characteristics, we try EDA first. It helps us to determine how best to manipulate our data to get the answers we need, making it easier for us to discover some meaningful outcomes.

5.1.1. Data overview

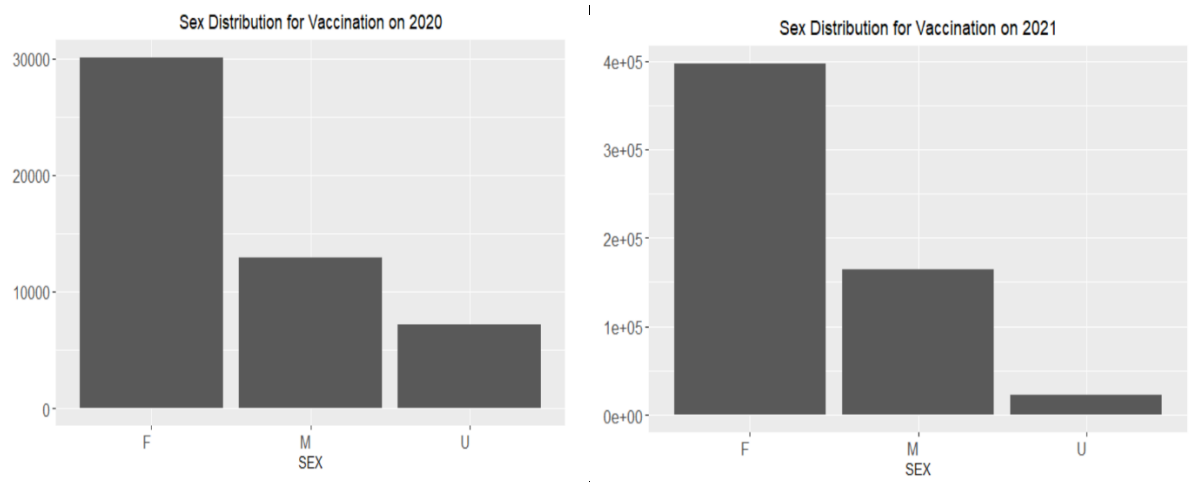
Some main data descriptions are listed below:

Header	Type (Max characters or format)	Description of Contents
VAERS_ID	Num(7)	VAERS identification number
AGE_YRS	Num(xxx.x)	Age in years
CAGE_YR	Num(xxx)	Calculated age of patient in years*
SEX	Char(1)	Sex
SYMPTOM_TEXT	Char(32,000)	Reported symptom text
VAX_TYPE	Char(15)	Administered vaccine type
VAX_MANU	Char(40)	Vaccine manufacturer
VAX_DOSE_SERIES	Char(3)	Number of doses administered

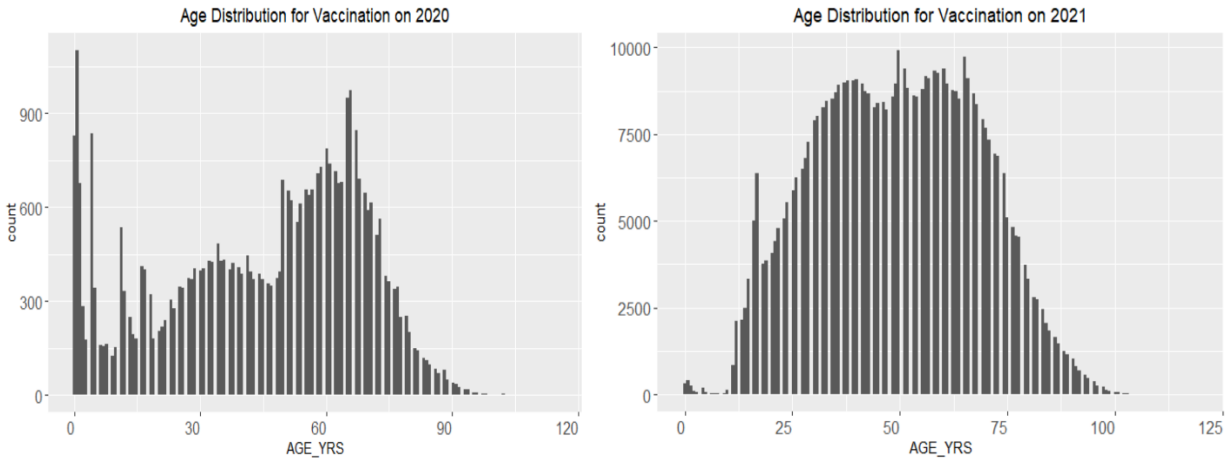
For VAERSDATA, the 2020 data has 50204 rows and 35 columns; the 2021 data has 583567 rows and 35 columns.
For VAERSVAX, the 2020 data has 60225 rows and 8 columns; the 2021 data has 609373 rows and 8 columns.
For VAERSSYMPTOMS, the 2020 data has 61565 rows and 11 columns; the 2021 data has 779234 rows and 11 columns.

5.1.2. Variables characteristic analysis

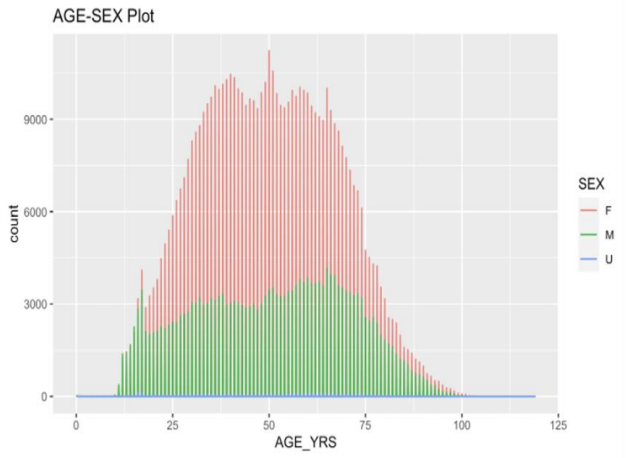
We try to have a more intuitive understanding of the data, so we plot some variables.



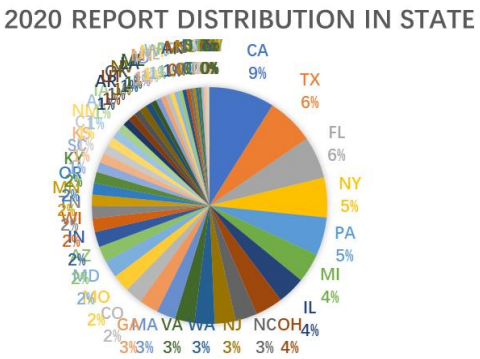
Plots show females are more than twice as male to report the adverse effects, which contradicts our previously expect that the sex may be equally distributed. This may be because females are more concerns about their body's condition and may be more willing to report.

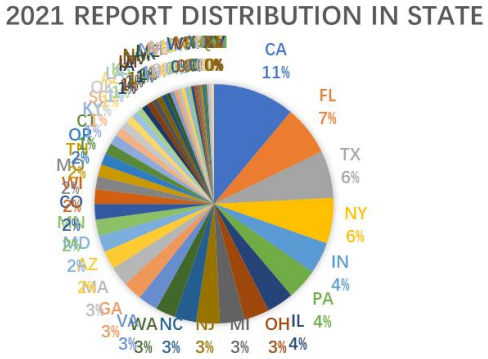


Plots above show middle age people's number are the largest, which follows our previous expectation. Since we guess the middle age group's vaccination number may also be the largest, the inverse effect number may also be the largest, which may approximately follow the normal distribution. However, in 2020's plot, the number of young people is also very high. We find Moderna doesn't allow people under 18 to vaccine and Pfizer doesn't allow people under 12. This means we may need to pay attention to this variable and do some adjustment later.



We do more to investigate the Age-Sex-Vaccination Relationship. These plots show in people's report, Pfizer and Moderna are mentioned most frequently. This may be because people vaccine these two kinds of medicine most frequently.





From analysis on state distribution, we find CA, FL, TX, NY's reporting numbers are higher than other states. This may be because people in these states are more willing to have vaccination.

5.2. References

[1] "VAERS Data Use Guide." *Vaccine Adverse Event Reporting System*, https://vaers.hhs.gov/docs/VAERSDataUseGuide_en_September2021.pdf

[2] Suling, Marc, and Iris Pigeot. "Signal Detection and Monitoring Based on Longitudinal Healthcare Data." *MDPI*, Multidisciplinary Digital Publishing Institute, 13 Dec. 2012, https://www.mdpi.com/1999-4923/4/4/607.

[3] "Proportional Reporting Ratios - Statistical Methods of Signal Detection: Pharmacovigilance." *pharmacy180.Com*, http://www.pharmacy180.com/article/proportional-reporting-ratios-3174/.

[4] Ramya, and Ramya Ramya is pharmacovigilance trainer. “43. Reporting Odds Ratio (ROR).” *Pharmacovigilance*, 5 Mar. 2020, https://allaboutpharmacovigilance.org/43-reporting-odds-ratio-ror.
[5] Sun, Li, et al. “Parallel ADR Detection Based on Spark and BCPNN.” *Tsinghua Science and Technology*, vol. 24, no. 2, 2019, pp. 195–206., https://doi.org/10.26599/tst.2018.9010074.
[6] Xiao, Nan, et al. *Rank Aggregated Signal Detection for Vaers Data*, https://nanx.me/rankv/.
[7] Pihur, Vasyl, et al. “Weighted Rank Aggregation of Cluster Validation Measures: A Monte Carlo Cross-Entropy Approach.” *Bioinformatics*, vol. 23, no. 13, 2007, pp. 1607–1615., https://doi.org/10.1093/bioinformatics/btm158.
[8] Garbade, Dr. Michael J. “Understanding K-Means Clustering in Machine Learning.” *Medium*, Towards Data Science, 12 Sept. 2018, https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.
[9] *PRR Computations*, https://docs.oracle.com/health-sciences/empirica-signal-811/ESIUG/PRR_Computations.htm#MiniTOCBookMark6.

5.3. Codes

#Code for data cleaning:

```
library(tidyverse)
#load original datasets
t01<-read_csv("./data/2020VAERSDATA.csv")
t02<-read_csv("./data/2020VAERSVAX.csv")
t03<-read_csv("./data/2020VAERSSYMPTOMS.csv")
t11<-read_csv("./data/2021VAERSDATA.csv")
t12<-read_csv("./data/2021VAERSVAX.csv")
t13<-read_csv("./data/2021VAERSSYMPTOMS.csv")
#clean data
t1<-rbind(t01,t11)%>%filter(RECVDATE<='09/23/2021')%>%
  dplyr::select(VAERS_ID,CAGE_YR,SEX)%>%filter(SEX!="U")%>%drop_na()
t2<-
rbind(t02,t12)%>%filter(VAX_TYPE=='COVID19'&(VAX_MANU=='MODERNA'|VAX_MANU=="PFIZER\\BIONTECH") & (VAX_DOSE_SERIES%in%c('1','2')))%>%
```

```
  dplyr::select(VAERS_ID,VAX_MANU,VAX_DOSE_SERIES)%>%drop_na()
t3<-
rbind(t03,t13)%>%unite('SYMPTOM',SYMPTOM1,SYMPTOM2,SYMPTOM3,SYMPTOM4,
SYMPTOM5,sep = '|',na.rm =T)%>%group_by(VAERS_ID)%>%
  summarise(SYMPTOM=paste0(SYMPTOM,collapse = '|'))%>%
  dplyr::select(VAERS_ID,SYMPTOM)%>%drop_na()
t3$SYMPTOM<-lapply(t3$SYMPTOM, function(x) gsub("(i|I)njection site","Vaccination site",
x))
t3<-as.data.frame(lapply(t3, unlist))
#combine data based on VAERS_ID
df<-t1%>%inner_join(t2,by='VAERS_ID')%>%inner_join(t3,by='VAERS_ID')
#detect interchange of vaccine products and exclude those ids
interchange<-
df%>%distinct()%>%group_by(VAERS_ID)%>%mutate(n=n())%>%filter(n>1)%>%group_by(
VAERS_ID,VAX_MANU)%>%mutate(n1=n())%>%filter(n1==1)
exid<-unique(interchange$VAERS_ID)
df1<-df%>%filter(!(VAERS_ID%in%exid))
```

#Code for dpa:

#prepare for dpa

```
df1<-df1%>%dplyr::select(VAERS_ID,VAX_MANU,SYMPTOM)
dpa_pre<-
df1%>%separate_rows(SYMPTOM,sep='[|]')%>%group_by(SYMPTOM,VAX_MANU)%>%m
utate(n=n())%>%dplyr::select(-VAERS_ID)%>%distinct()%>%drop_na()
library(PhViD)
#PRR
symptoms <- as.PhViD(dpa_pre,MARGIN.THRES = 100)
prr <- PRR(symptoms,DECISION = 3,RANKSTAT = 2)
prr_result<-prr$SIGNALS[order(prr$SIGNALS$`LB95(log(PRR))`,decreasing = T),]
row.names(prr_result) <- NULL
#ROR
```

```
ror<-ROR(symptoms,DECISION = 3,RANKSTAT = 2)
ror_result<-ror$SIGNALS[order(ror$SIGNALS$`LB95(log(ROR))`,decreasing = T),]
row.names(ror_result) <- NULL
#BCPNN
bcpnn<-BCPNN(symptoms,DECISION = 3,RANKSTAT = 2)
bcpnn_result<-bcpnn$SIGNALS[order(bcpnn$SIGNALS$`Q_0.025(log(IC))`, decreasing =
TRUE),]
row.names(bcpnn_result) <- NULL
```

#Code for rank aggregation:

```
agg_pre<-function(data){
  df<-as.data.frame(data$SIGNALS)
  rownames(df) <- paste(df$`drug code`, df$`event effect`, sep = ": ")
  return(df)
}
df_prr<-agg_pre(prr)
df_ror<-agg_pre(ror)
df_bcpnn<-agg_pre(bcpnn)
common_signals <- Reduce(intersect, list(rownames(df_prr), rownames(df_ror),
rownames(df_bcpnn)))
length(common_signals)
agg<-function(df){
  df<-df[common_signals, ]
  df <- df[order(df[,5], decreasing = TRUE), ]
  return(df)
}
df_prr<-agg(df_prr)
df_ror<-agg(df_ror)
df_bcpnn<-agg(df_bcpnn)
ranks <- matrix(0, nrow = 3, ncol = length(common_signals))
ranks[1, ] <- rownames(df_prr)
ranks[2, ] <- rownames(df_ror)
```

```
ranks[3, ] <- rownames(df_bcpnn)
colnames(ranks) <- 1:ncol(ranks)
library("RankAggreg")
rankagg <- RankAggreg(
  ranks, k = 20,
  distance = "Spearman", method = "GA", maxIter = 2000,
  seed = 2021, verbose = FALSE
)
plot(rankagg)
rankagg$top.list
```

#Code for K-Means:

```
df1$SEX[df1$SEX=="F"]<-1
df1$SEX[df1$SEX=="M"]<-0
df1<-df1%>%mutate(SEX=as.integer(SEX))
df1<-na.omit(df1)
d <- df1 %>%
  select(CAGE_YR,SEX) %>%
  scale()
#Use elbow method by plotting to determine the number of k clusters
tot.withinss <- c()
for(k in 2:20){
  res <- kmeans(d, k)
  tot.withinss[k-1] <- res$tot.withinss
}
plot(2:20, tot.withinss, type="b")
#Create 15 clusters
res <- kmeans(d, 15)
df1$cluster <- res$cluster
head(df1)
```

#Code for Stratified PRR and Comparison with Unstratified PRR

```
s012345$prr <- rep(NA, dim(s012345)[1])
```

```
for (i in 1:dim(s012345)[1]) {
  nu <- 0
  de <- 0
    for (j in 1:length(unique(covid2021$cluster))) {
      nu <- nu + s012345[i, 2*j]
      de <- de + s012345[i, 2*j +1]
    }
  s012345$prr[i] <- nu/de
}
sum(s012345$prr>1 & s012345$prr != Inf)
# compare results
prr_result <- read.csv("prr_result.csv")
for (i in 1:length(prr_result$PRR)) {
  prr_result$name[i] <-  paste(prr_result$drug.code[i], prr_result$event.effect[i])
}
nameall <- prr_result$name[prr_result$PRR>1 & prr_result$PRR != Inf]
sum(prr_result$PRR>1 & prr_result$PRR != Inf)
length(intersect(nameall, s012345$name))
```

#Codes for EDA

```
qplot(x=SEX,data = t01)+ggtitle("Sex Distribution for Vaccination on 2020")+
theme(plot.title = element_text(hjust=0.5))
qplot(x=SEX,data = t11)+ggtitle("Sex Distribution for Vaccination on 2021")+
  theme(plot.title = element_text(hjust=0.5))
creat_plot_t01 = function(name,binwidth){
  return(ggplot(aes_string(x=name),data = t01,ylab = 'count')+geom_histogram(binwidth =
binwidth)+ggtitle("Age Distribution for Vaccination on 2020")+theme(plot.title =
element_text(hjust=0.5)))
}
creat_plot_t11 = function(name,binwidth){
```

```
  return(ggplot(aes_string(x=name),data = t11,ylab = 'count')+geom_histogram(binwidth =
binwidth)+ggtitle("Age Distribution for Vaccination on 2021")+theme(plot.title =
element_text(hjust=0.5)))
}
creat_plot_t01('AGE_YRS',0.7)
creat_plot_t11('AGE_YRS',0.7)
t_01 = t01%>% group_by(STATE)%>%
  summarise(counts = n())
max(t_01$counts)
t_11 = t11%>% group_by(STATE)%>%
  summarise(counts = n())
eda_symp<-function(data,column,t){
  fig<-data%>%filter(SYMPTOM %in% symp)%>%
    ggplot(aes(x=SYMPTOM,fill={{column}},y=`p (%)`))+
    geom_bar(position=position_dodge(0.75), stat="identity")+
    labs(title=paste0('ER of 8 Typical Symptoms According to\n ',t),
         x='Symptom')+
    theme(plot.title = element_text(hjust =0.5))+
    theme(axis.text.x = element_text(angle = 60, hjust = 1, size=12),
          axis.text=element_text(size=12))
  return(fig)
}
eda_symp(symp_ttl_manu,VAX_MANU,"Vaccine Manufacturer")
eda_symp(symp_ttl_sex, SEX,"Sex")
eda_symp(symp_ttl_age,age_group,"Age Group")
eda_symp(symp_ttl_series,VAX_DOSE_SERIES,"Number of Doses Administered")
```