Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 148
DOI: 10.11159/icsta22.148

Identifying Safety-Vaccine Association for COVID-19 Vaccines from VAERS

Jianping Sun[1]
[1]Department of Mathematics and Statistics, University of North Carolina at Greensboro
116 Petty Building, Greensboro, NC 27402, USA
j_sun4@uncg.edu

Abstract **-** Vaccine Adverse Event Reporting System (VAERS) is a national early warning system to detect possible safety issues in US licensed vaccines. However, the data mining task for VAERS is quite challenging mainly due to the high dimensionality of the data set and complex confounding presented among adverse events and vaccines. In addition, the inconsistent data quality and the rarity of cases and events add more difficulties to the analysis of VAERS data. Tan et. al. (2022+) [1] proposed a 3-step strategy to detect safety associated vaccines by using zero-inflated Poisson regression model with variable selection from VAERS. In this paper, I discuss the safety issue related with COVID-19 vaccines that were reported to VAERS from year 2020 to current with the implementation of a modified version of 3-step strategy.

Keywords**:** Safety-Vaccine Association, COVID-19 Vaccine, VAERS, Zero-Inflated Poisson Model

1. Introduction

Established in 1990 and co-managed by Centers for Disease Control and Prevention (CDC) and U.S. Food and Drug Administration (FDA), the Vaccine Adverse Event Reporting System (VAERS) is a post marketing surveillance program. VAERS collects information about adverse events (AEs), i.e., possible harmful side effects, that occur after administration of vaccines. This system is expected to serve as a national early warning system to detect possible safety problems in U.S.-licensed vaccines [2]. VAERS reports are submitted by healthcare providers, manufacturers, vaccine recipients, and other sources. The data is publicly available, organized by year and updated every quarter. In a typical year, about 40,000 new reports are added into VAERS with about 3,000 different AEs, 70 vaccines involved, and 20 other variables, such as age, gender, hospitalization and so on.

VAERS database is a unique information source for rapid signal detection of potential safety issues that could be further evaluated by other safety systems, such as CDC's Vaccine Safety Datalink (VSD) [3]. However, being a passive data collection system, VAERS has limitations including reporting bias, inconsistent data quality and completeness, lack of control (unvaccinated) comparison group, all of which make VAERS inappropriate for assessing the causal relationship. In addition, the nature of high dimensionality, complex confounding, and rarity of events in VAERS data make it even harder to identify vaccines that maybe potentially alarming.

A number of methodology work has been developed to detect the signal of safety-vaccine association using VAERS or similar data ([4,5,6]). In particularly, Tan et. al. (2022+) [1] proposed a 3-step strategy to detect safety associated vaccines by using zero-inflated Poisson regression model with variable selection and illustrated the proposed method by analysing the VAERS data between years 2014 to 2019 as an example.

Since FDA approved emergency use authorization (EUA) of the first COVID-19 vaccine in Dec. 2020, huge number of reports have been added to VAERS: 49,683 reports in 2020, 747,911 reports in 2021, and 78,617 in 2022 by March 11. Most of these reports (788,624 out of a total of 876,211 reports) involved COVID-19 vaccines that are currently approved in U.S. Hence, in the following part of this paper, I will discuss the safety issue related with COVID-19 vaccines that were reported to VAERS from year 2020 to current, via a modified 3-step strategy proposed above.

2. Method

In the 3-step strategy, the authors focused on identifying vaccines that are associated with hospitalization (yes/no and length of stay) in the VAERS data. The usage of this single outcome other than the other AEs is because that hospitalization belongs to serious adverse events and is a more important clinical outcome than most of the other AEs. In addition, hospitalization reports tend to be more accurate and reliable compared with the other AEs. Consequently, in this analysis, I also focus on the outcome of hospitalization, especially the length of stay (LOS).

Due to the challenges of rarity and high dimensionality in VAERS data, in the 3-step strategy, the authors first screened all AEs and vaccines by excluding variables with low occurrence (e.g., < 0.01%), and then used a computation efficiency vaccine-AE association method, such as Fisher's exact test, to keep only vaccines and AEs that are associated.

To further filter the vaccines and AEs that were identified from the above step, in the second step, the authors applied a univariate screening procedure to examine the association between a given vaccine (or AE) and outcome variable by using linear regression model (for LOS) and logistic regression (for hospitalization yes/no). In this step, to account for complicated dependence and confounding among AEs and vaccines, principal components (PCs) for AEs and for vaccines were obtained after removing low-occurrence ones, and the calculated PCs were added in the linear or logistic regression as covariates.

In the final step, a zero-inflated Poisson (ZIP) model was implemented to connect LOS with filtered vaccines. In the literature, it is recommended to assume a Poisson distribution for LOS, or a zero-inflated Poisson (ZIP) model when a large portion of LOS has value of 0 [7,8]. Specifically, denote $Y = (Y_1, Y_2, \dots, Y_n)$ as the LOS vector of n VAERS reports, $B_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ as the covariates in the ith report that are associated with the zero-part of the ZIP model, and $G_i = (x_{i1}, x_{i2}, \dots, x_{iq})$ as the covariates in the ith report that are associated with the count-part of the ZIP model. For the ith report, a ZIP model, which can be viewed as a mixture model with two components, is constructed, as

$$Y_i \sim \pi_i Poisson(0) + (1 - \pi_i) Poisson(\lambda_i). \tag{1}$$

In equation (1), $Poisson(0)$ represents a degenerated Poisson distribution with intensity rate 0, $Poisson(\lambda_i)$ denotes a Poisson distribution with a positive intensity rate. Here $logit(\pi_i) = B_i'\beta$ with β as the unknown zero-part regression coefficients, and $\log(\lambda_i) = G_i'\gamma$ with γ as the count-part regression coefficient to be estimated. Note that two vectors B_i and G_i may contain common components. In the 3-step strategy, B_i and G_i are the same by including both vaccines and PCs for vaccines. A LASSO penalty [9] was also imposed on the regression coefficients in the above ZIP model at the end, so that to select significant vaccines.

3. Results

In this paper, the focuses are detecting associations between LOS and COVID-19 vaccines based on the VAERS reports from year 2020 to March 2022. During this time, there are in total of 876211 reports on 11775 AEs and 68 vaccines, in which COVID-19 vaccines are separated by different brands (Janssen, Moderna, Pfizer-Biontech, and Unknown). Since the number of involved AEs and vaccines in this time period is smaller than previous years, and the main interested vaccines are COVID-19 vaccines, in the analysis, I used occurrence thresholds 1% for AEs and 0.05% for vaccines to keep 77 AEs and 32 vaccines after initial screening. Besides, due to vaccine-AE association is not primary focus here, the step of Fisher's exact test for vaccine-AE association was skipped.

To further filter vaccines, as suggested by the 3-step strategy, univariate screenings were conducted to test the association between a given vaccine and LOS and hospitalization (yes/no) by using linear and logistic regression, respectively. In both regression models, all 77 AEs and 4 PCs for vaccines (explaining >90% of total variance) were adjusted as covariates. The reason not using PCs for AEs is because that the dependency among AEs was not main interests here. Note, since vaccines are binary variables (1 for involved and 0 for not involved), the logistic PC method was used to obtain PCs [10,11]. In addition, permutation test was applied to get the significance of the association between outcomes and target vaccine. By using a cut-off value of 0.05, 27 vaccines (including all COVID-19 vaccines) were retained after this step.

Finally, the remained 27 vaccines, 4 vaccines PCs obtained from 32 vaccines that passed initial screening, and two other clinics variables, age and sex, were used for both B vector and G vector in the ZIP model (1). Note that variable age had a portion of missing values (~12%), I imputed it by the median age (50) among all reports during 2020 to now. Similarly, a number of variable sex (~5.6%) was missing and denoted as Unknown in the analysis. In addition, two reports that had outlier LOS (99999 days) were removed. So, at the end, a ZIP model was fitted by using 876,209 reports with 33 variables. The fitting results for the final ZIP model are summarized in Table 1 below.

Table 1: Results of ZIP model for 27 vaccines.

		Logistic Model Part			Poisson Model Part		
Freq	Vaccine	Estimate	Std.	P-value	Estimate	Std.	P-value
7.76%	COVID-19 Janssen	-0.97	0.128	3.52E-14	-0.49	0.044	< 2.00E-16
41.06%	COVID-19 Moderna	-0.02	0.116	8.32E-01	-0.53	0.042	< 2.00E-16
41.15%	COVID-19 Pfizer	-0.54	0.105	3.57E-07	-0.48	0.042	< 2.00E-16
0.22%	COVID-19 Unknown	-1.24	0.162	2.12E-14	-0.58	0.066	< 2.00E-16
0.17%	DTAP	-1.20	0.187	1.78E-10	-0.59	0.064	< 2.00E-16
0.10%	DTAPHE	-1.19	0.204	5.96E-09	0.09	0.082	2.79E-01
0.18%	DTAPIPV	-0.29	0.292	3.27E-01	-0.03	0.119	8.15E-01
0.15%	DTAPIPVHIB	-0.86	0.190	5.67E-06	0.62	0.073	< 2.00E-16
1.50%	FLU	0.19	0.284	5.08E-01	-0.53	0.105	5.60E-07
1.43%	FLU4	-0.19	0.275	4.85E-01	0.13	0.101	2.05E-01
0.06%	FLUA3	-0.37	0.247	1.31E-01	-0.72	0.110	6.37E-11
0.17%	FLUA4	0.38	0.206	6.89E-02	-0.17	0.079	3.02E-02
0.48%	FLUX	-0.22	0.275	4.33E-01	0.34	0.091	2.10E-04
0.28%	HEP	-0.30	0.176	9.15E-02	-0.42	0.065	8.03E-11
0.23%	HIBV	-0.46	0.183	1.19E-02	-0.58	0.073	1.42E-15
0.43%	HPV	-0.88	0.252	4.98E-04	0.19	0.084	2.71E-02
0.09%	IPV	-0.24	0.242	3.27E-01	2.30	0.046	< 2.00E-16
0.20%	MEN	-0.30	0.232	1.98E-01	-0.04	0.089	6.16E-01
0.33%	MMR	-1.35	0.190	9.92E-13	0.03	0.061	6.26E-01
0.30%	MMRV	-0.99	0.263	1.62E-04	-1.61	0.149	< 2.00E-16
0.29%	MNQ	-0.28	0.207	1.83E-01	-0.66	0.077	< 2.00E-16
0.34%	PNC13	-0.63	0.165	1.44E-04	-0.58	0.066	< 2.00E-16
0.54%	PPV	0.56	0.229	1.53E-02	-0.83	0.089	< 2.00E-16
0.26%	RV	-1.92	0.167	< 2.00E-16	-0.83	0.062	< 2.00E-16
0.12%	TYP	2.22	0.716	1.90E-03	0.13	0.253	6.11E-01
1.40%	UNK	0.13	0.157	4.04E-01	0.03	0.054	6.06E-01
3.16%	VARZOS	1.50	0.156	< 2.00E-16	0.60	0.056	< 2.00E-16

4. Conclusion

Table 1 lists the frequency of each 27 vaccines involved among all 876209 reports. The first four rows in this table are the results for COVID-19 vaccines. It shows that Janssen and Pfizer-Biontech vaccines were negatively and significantly associated with being hospitalized (P-values = 3.52E-14 and 3.57E-07), and Moderna vaccine was not significantly associated with hospitalization (P-value=8.32E-01). Conditional on being hospitalized, all three brand COVID-19 vaccines were significantly associated with shorter hospital stay. The conclusions for the other 23 vaccines could be made similarly according to this table.

However, due to the limitations of VAERS, the above analysis may only be used to flag the potential associations, but not causalities. Reliable interpretation of such findings based on VAERS is still quite challenging. The hope is that the proposed method and analysis results could help to shape new hypotheses which could be examined through other more reliable and appropriate data sources.

References

[1] X. Tan, W. Wang, G. Liu, D. Zeng, G. Diao, J. Ibrahim, "Detecting Safety-Vaccine Association from VAERS Data with Complex and High Dimensional Confounding", *under review*, 2022+

[2] T.T. Shimabukuro, M. Nguyen, D. Martin, F. DeStefano, "Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS)," *Vaccine*, vol. 33, no. 36, pp. 4398-4405, 2015.

[3] R.T. Chen, J.W. Glasser, P.H. Rhodes, R.L. Davis, W.E. Barlow, R.S. Thompson, J.P. Mullooly, S.B. Black, H.R. Shinefield, C.M. Vadheim, S.M. Marcy, J.I. Ward, R.P. Wise, S.G. Wassilak, S.C. Hadler, "Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States," *Pediatrics*, vol. 99, no. 6, pp. 765–773, 1997.

[4] S.J. Evans, P.C. Waller, S. Davis, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports", *Pharmacoepidemiol Drug Saf.*, vol. 10, no. 6, pp. 483-486, 2001.

[5] L. Huang, J. Zalkikar, R.C. Tiwari, "Likelihood ratio test-based method for signal detection in drug classes using FDA's AERS database," *Journal of Biopharmaceutical Statistics*, vol. 23, no. 1, pp. 178-200, 2013.

[6] K. Nam, N.C. Henderson, P. Rohan, E.J. Woo, E. Russek-Cohen, "Logistic Regression Likelihood Ratio Test Analysis for Detecting Signals of Adverse Events in Post-market Safety Surveillance," *Journal of Biopharmaceutical Statistics*, vol. 27, pp. 990-1008, 2017.

[7] C.X. Feng and L. Li, "Modelling Zero Inflation and Overdispersion in the Length of Hospital Stay for Patients with Ischaemic Heart Disease," in *Advanced Statistical Methods in Data Science*, D.G. Chen, J. Chen, X. Lu, G.Y. Yi, H. Yu, Springer, 2016, pp. 35–53.

[8] J.X. Song, "Zero-inflated Poisson regression to analyze lengths of hospital stays adjusting for intra-center correlation," Communications in Statistics—Simulation and Computation, vol. 34, no. 1, pp. 235-241, 2005.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267-288, 1996.

[10] A.J. Landgraf, Y. Lee, "Dimensionality reduction for binary data through the projection of natural parameters," *Journal of Multivariate Analysis*, vol. 180, no. 3, 104668, 2020.

[11] S. Battaglino and E. Koyuncu, "A Generalization of Principal Component Analysis," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3607-3611