Exploratory Data Analysis of Covid Vaccine Safety Data

Nikhil Gangaramani[1], Sahil Lotya[2], Akansha Ahuja[3], Shivani Kulkarni[4], Sujata Khedkar[5], Devesh Rajadhyax[6]

{2018.nikhil.gangaramani@ves.ac.in[1], 2018.sahil.lotya@ves.ac.in[2], 2018.akansha.ahuja@ves.ac.in[3], 2018.shivani.kulkarni@ves.ac.in[4], sujata.khedkar@ves.ac.in[5], devesh.rajadhyax@cerelab.com[6]}

Computer Engineering, Vivekanand Education Society's Institute of Technology Mumbai, India[1-4],
Vivekanand Education Society's Institute of Technology Mumbai, India[5],
Cere Labs Pvt. Ltd. Mumbai, India[6]

Abstract. Adverse Drug Reaction (ADR) is a harmful or unpleasant reaction from the use of a medical product. Whenever a vaccine is discovered, it goes through a series of clinical trials. Though these trials are useful to detect ADRs using vaccine safety data, a system to visualize the analysis of the adverse effects of a vaccine considering the patient's condition would be more helpful to prevent mishaps. To simplify and speed up the process of analysis of ADRs from texts this system implements Exploratory Data Analysis and Natural Language processing techniques wherein the system can generate graphs of analysis of each considered parameter, process textual data for analysis, and also extract medical terms from text using Named Entity Recognition eventually identifying and analyzing potential data of the patient that can lead to an adverse reaction.

Keywords: Adverse Drug Reaction, vaccines, Topic Modeling, Latent Dirichlet Allocation, Named entity recognition, POS tagging, pharmacovigilance

1 Introduction

Keeping a track of Adverse Drug Reaction (ADR) data as well as patient information is critical for post-market drug safety monitoring. Problems arise in times of a pandemic like COVID-19 when there wasn't enough time for the detection of every ADR. People of different age groups, suffering from various illnesses and different regions of the world react differently to vaccines. It is very important to acknowledge every available piece of information regarding the patient, vaccine, vaccination procedure, and adverse events that took place.

Vaccines manufactured are supplied to different parts of the world. Considering the difficulty to keep a track of ADRs, an international, fast, and robust pharmacovigilance system is needed. NER (Named Entity Recognition), polarity checker, classification algorithms, and

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318171

other approaches can be utilized to accelerate the process. Identification of fatal ADRs can help vaccine manufacturers and drug regulators to take necessary steps like drug recall, updated safety guidelines, etc. Analysis of ADRs data can help understand the demographic of people who are likely to be affected by ADRs. Such people can be warned of the potential side effects or can be given alternatives.

If data continues to be collected and stored regularly, an analysis and prediction system can be scaled worldwide and discrepancies can also be handled in underdeveloped countries such as the differences created due to the unreported cases.

Starting with Exploratory Data Analysis with the available data, the system can detect patterns and relationships among a large number of variables in the dataset. It is essential to consider every possible attribute of the dataset that could help professionals analyze things faster. This system focuses on evaluating textual data aspects such as POS tag frequency, medical term extraction, sentiment polarity, and linguistic features.

2 Related Work

In order to achieve the work's purpose of doing exploratory data analysis on vaccine safety data and extracting usable data for further study, we looked at the following publications and articles.

The first reference [1] analyses the frequencies of SARS- CoV-2 infection in cohorts of patients stratified by blood type, age, and race-based on a statistical analysis of 137,037 individuals from the Mayo Clinic electronic health record (EHR) database. The system's drawback was that the dataset could include factors like medical history and symptom review.

To better understand the distribution of Covid-19 in India, Sarvam Mittal [2] built a statistical model. It looked at statistics on Covid-19 transmission in India compared to other nations, as well as age and state-specific transmission, Covid- 19 symptoms, and future epidemic forecasts in India.

Chiahui Yu et al. [3] propose a framework that focuses on data analysis and visualization for productive exploration of data about adverse drug reaction (ADR) surveillance. The methodology was tested using data from social media (AskaPatient.com) and data from the FDA Adverse Event Reporting Systems (FAERS). The framework has a 79% F1 score and 95% accuracy, 87% precision, 73% recall. The discovery's weakness is that it should have been based on a larger dataset with variables like medical history and allergies.

AbeedSarker et al. [4] has proposed a methodical review for the detection and extraction of ADR from social media data and their relevance to pharmacovigilance. They also describe a possible mechanism for tracking ADR via social media. They classified the research based on a variety of factors, including the principal ADR detection method, the corpus size, alternative methodologies depending on the social network or type of social media from which the data was gathered, availability, and evaluation criteria. The system's disadvantage is that user-generated text contains (e.g., misspellings, abbreviations, and irregular phrase formation), which limits the performance of lexicon-based techniques.

Jean-Louis Montastruc et al. [5] examine the advantages and disadvantages of disproportionality analysis for ADR identification. The primary use of the method is to confirm (or not) a probable association predicated on a pharmacological hypothesis between a specific drug and an ADR. The proposed method should be used for signal detection and should not be considered as a quantitative risk

Further, analysis was done on research work based on the dataset we used. Jian-Jian Ren [6] presents a statistical analysis of vaccine-adverse-event data based on the Vaccine Adverse Event Reporting System (VAERS) dataset, which contains over 500,000 post-vaccination reports from US- approved vaccines. Their research tries to discover trends in how all reported adverse event symptoms are linked to vaccines. The major drawback of the study is that 24 years of data is analyzed together which can make results inaccurate due to changing trends and particular vaccine data not being analyzed.

The next one gave a detailed view of the data and highlighted the importance of the source of data [7]. The drawback was it had a smaller number of rows for developing a system. Both technological and sociological variables were considered in the research.

Another implementation had target variables of analysis as death, SARS-CoV-2 positive patients, and hospital admission status [8]. Most significantly, it took into account the patient's medical history. The disadvantage was that the data had not been thoroughly cleaned prior to the primary implementation. One model including Weka implementation followed by classification algorithms gave the highest accuracy of 99.9% with the decision tree classifier and took only 0.08 seconds to build [9].

Lastly, an analysis performed in R on the population, vaccines, age groups, and particular ADRs revealed trends and results showed 15% of reported AEs were classified as severe [10]. All of these studies were successful, however, the total number of rows considered was only 6745 [10]. The percentages will fluctuate dramatically in subsequent rounds of the investigation. These papers contain the dataset's interim results and analysis and hence deserve more examination.

3 Proposed Model

3.1. Dataset

The data set used for the system analysis is the Vaccine Adverse Event Reporting System (VAERS) dataset. Three CSV files are included with the dataset. As of September 2021, the dataset's overall size was estimated to be around 8 Lakh 25 thousand. The dataset contains

ADRs reported due to three vaccines which are Pfizer Biotech, Moderna, and Janssen. We have also found an Indian dataset which is released by the Ministry of Health and Family Welfare's (MoHFW) immunization department. At the national level, a special group has been constituted to assess the cause of AEFIs (Adverse Event Following Immunization) of COVID-19 vaccinations. The special group comprises medical experts from different specializations who give their opinion on every AEFI report. The advantage of this dataset is that the data is screened by medical experts and hence the system can get better analytics from such types of data. The data is recently released and hence is low in number. We're looking into the dataset and will examine it once enough information is made public.

3.2. Implementation

Clinical studies are conducted by healthcare professionals to detect ADRs before selling medications, however, they are frequently limited in number. As a result, when the medications are released on the market, post-market drug safety monitoring is essential to help detect ADRs.

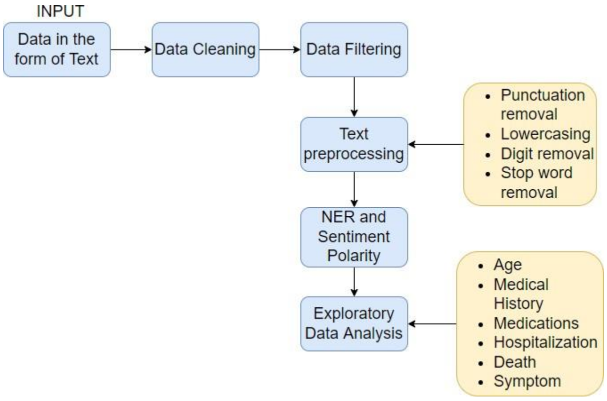

Fig.1 Modular Diagram of the System

- **Data Preprocessing**

The dataset used by the system contains three CSV files. In these studies, the system employed the VAERS id attribute to merge the datasets. As the vaccination campaign in the United States began in December 2020, the system used data from that year. As VAERS reports ADRs on all kinds of vaccines, the system eliminated the non-COVID vaccines entries as the system's main focus is on COVID-19 vaccines. The VAERS dataset included a lot of duplicate entries, therefore those were removed. VAERS discontinued a few attributes so the system dropped those entries. The age attribute of about 58 thousand entries was blank so to tackle that the system filled such entries by the median age of the dataset. The system discarded entries with blank values for critical parameters including VAX DATE (Date of Vaccination), ONSET DATE (Date on which ADR symptoms became noticeable), and Symptoms (Symptoms described by the patient). The following are the steps for filtering the data:

1. Stop-word removal using NLTK library
2. Removal of newline characters
3. Tokenization i.e., converting sentences into a list of words and removing punctuations
4. Lemmatization of the words

After filtering, only nouns, verbs, adjectives, and adverbs were kept using POS (Parts Of Speech) tagging.

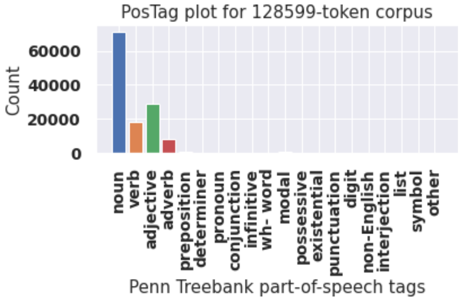

Fig.2. POS tagging

Topic modeling is a sort of statistical modeling used to find abstract subjects in a collection of texts. Latent Dirichlet Allocation is a technique used by the system (LDA). Each document is considered as a set of topics, each of which has its own set of keywords. Topics are nothing more than a collection of notable keywords or terms that are most likely to appear in a topic, allowing you to figure out what the topic is about. The system also consists of bigrams and trigrams model - bigrams are two words frequently occurring together in a document. Trigrams are groups of three words that frequently appear together.

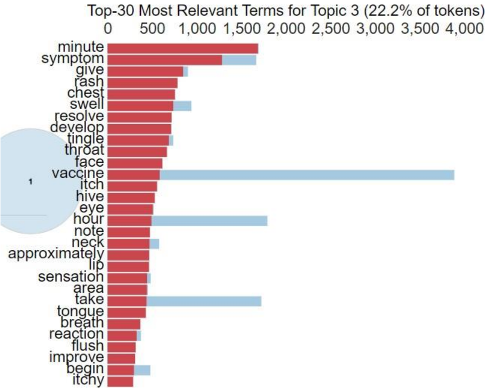

Fig. 3 Topic modeling of 30 most relevant terms

- **Data Extraction**

For Named entity recognition (NER), the system used the Scispacy pertained model “En_ner_bc5cdr_md: A spaCy NER model trained on the BC5CDR corpus. According to the researchers, the BC5CDR corpus has 1500 PubMed articles [11] with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease correlations. It helps to detect the disease and chemical-related words present in the text. This has helped with data trend analysis.

4 Results and Discussion

4.1. Analysis of VAERS dataset

As the VAERS dataset is a dataset that has ADRs of all types of vaccine the system found that 92.62% of the data is of COVID-19 vaccine data. As seen in Figure 4, the remaining 7.38% of the data is for non-COVID-19 vaccination. Such data of non-COVID-19 vaccines is dropped before further analysis.

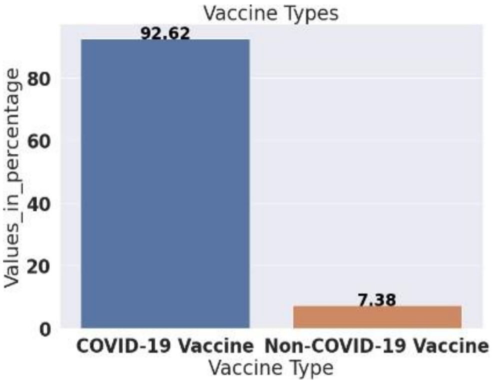

Fig.4. Types of Vaccines

Figure 5 depicts the gender distribution in the system dataset. We clearly see the percentage of female patients is significantly high i.e., 68.78%.

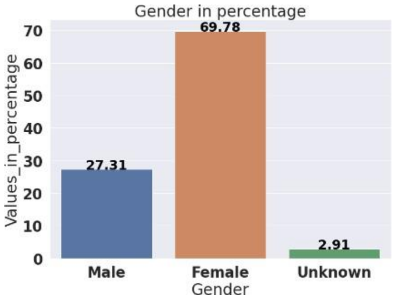

Fig.5. Gender Percentage

Figure 6 shows the percentage of hospitalizations in the dataset. The percentage of patients admitted to the hospital is 5.01 percent.

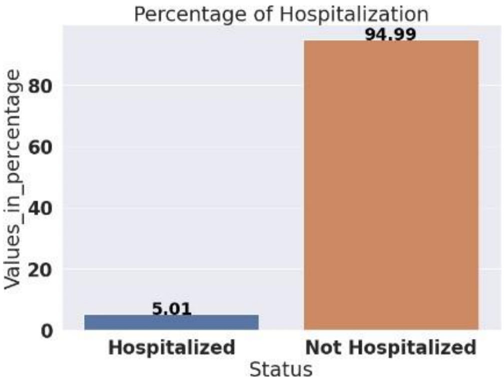

Fig.6. Hospitalization

Figure 7 shows the daily ADR count which is reported in the United States.

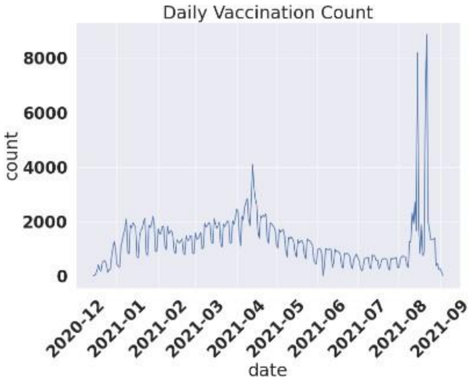

Fig.7. Daily ADR Count

Figure 8 depicted the death % in the dataset. ADRs claimed the lives of 1.03 percent of the patients in the study.

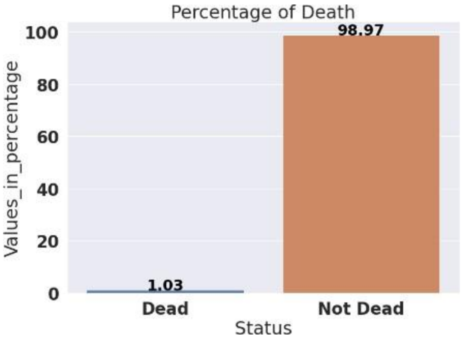

Fig.8. Death Percentage

Figure 9 shows that there is an equal distribution in normal reactions in all age groups. whereas we can see more hospitalizations in the age group above 50 similarly majorities of the dead patients belong to the age group of 50 to 75.

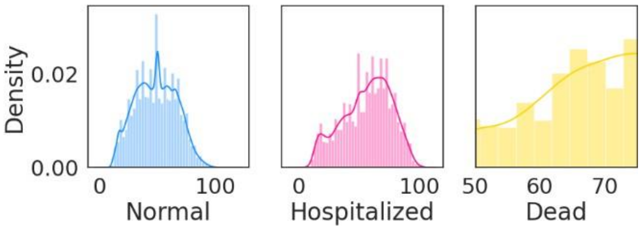

Fig.9. Age Distribution

Figure 10 shows the number of entries for each vaccine present in the dataset. We can clearly see that the Moderna vaccine has shown the highest number of ADRs in the system dataset.

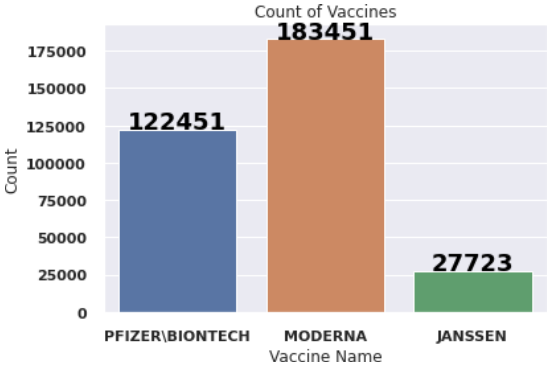

Fig.10. Count of vaccines from different Manufacturers

4.2. Analysis of attributes

The entire VAERS data is divided into 3 categories which are normal, hospitalized, and dead patients for the analysis.

- History of patients
  Fig. 11 consists of the patients who suffered from normal adverse reactions majorly had a history of depression, hypertension, asthma, GERD, or anxiety

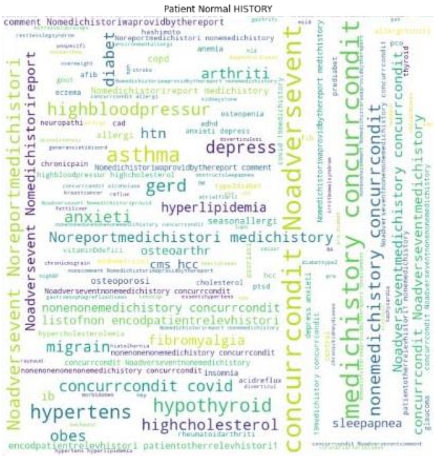

Fig.11. Normal patient’s history

Fig. 12 consists of the patients who were hospitalized after having adverse reactions majorly had a history of hypertension, hyperlipidemia, GERD, or hypothyroid.

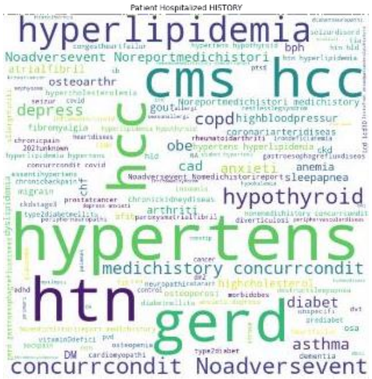

Fig.12. Hospitalized patient’s history

Fig. 13 consists of the patients who died after having adverse reactions majorly having a history of hypertension, hyperlipidemia, atrial fibrillations, Chronic obstructive pulmonary disease (COPD), GERD, or hypothyroid.

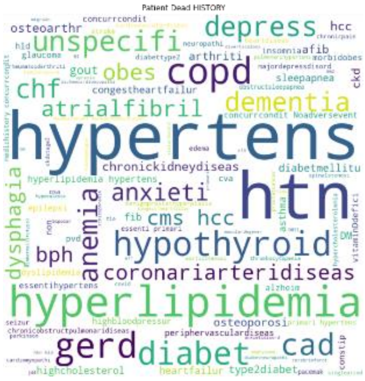

Fig.13. Dead patient’s history

- **Allergies of patients**

Fig. 14 patients that had normal adverse reactions majorly had allergies from sulfa, penicillin, latex, or codeine.

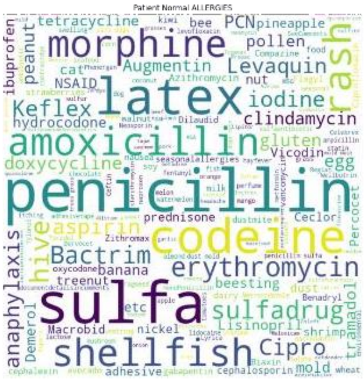

Fig.14. Normal patient’s allergies

Fig. 15 Shows that patients who were hospitalized after having adverse reactions majorly had allergies from codeine, latex, morphine, penicillin, or sulfa.

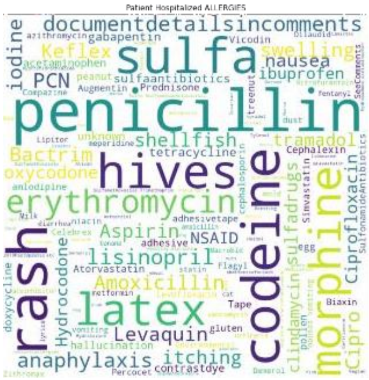

Fig.15. Hospitalized patient’s allergies

Fig. 16 Shows that patients who were dead after having adverse reactions majorly had allergies from codeine, penicillin, morphine, or lisinopril.

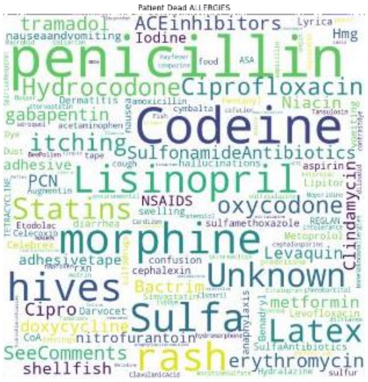

Fig.16. Dead patient’s allergies

- **Patient’s Medication**

Fig. 17 Patients suffering from a normal adverse reaction were majorly taking medication such as multivitamins, Vitamin D, fish oil, calcium.

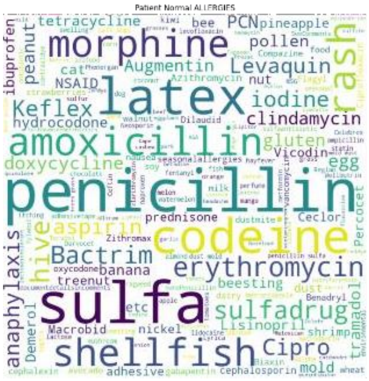

Fig.17. Normal patient's medication

Fig. 18 Shows that patients who were hospitalized after having adverse reactions were mostly taking medications atorvastatin, aspirin, levothyroxine, lisinopril, or multivitamins.

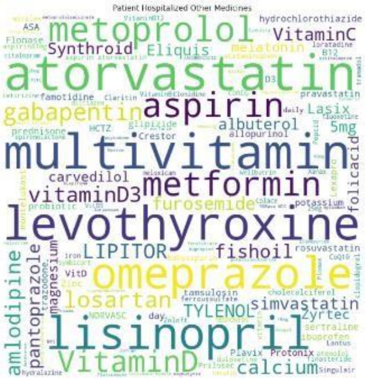

Fig.18. Hospitalized patient's medication

Fig. 19 Shows that patients who were dead after having adverse reactions were majorly taking medicines atorvastatin, gabapentin, aspirin, furosemide, or atorvastatin.

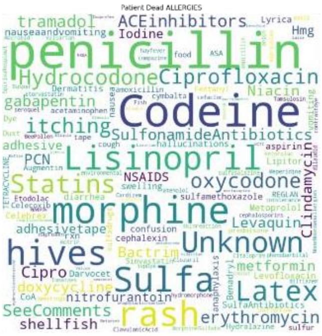

Fig.19. Dead patient's medication

5 Conclusion

India has reported 70,102 adverse events and 1,013 deaths following Covid Immunization (AEFI) till January 30, 2022, in over one year since the introduction of the vaccination program in January 2021, according to the government data. If the system has the right provision for testing the vaccines enough beforehand, all these complications can be easily avoided saving many lives. If time and circumstances do not permit, analysis can be done under post-marketing pharmacovigilance. Keeping this in mind, a model for this problem can help assess the situation faster. To simplify and speed up the process of analysis of ADRs from texts this system implements Exploratory Data Analysis on each considered parameter of a dataset and develops an interactive statistics visualization. Further, this study also performed analysis on various attributes such as the history of patients, allergies, age, and medications.

6 Future Scope

After exploratory data analysis, the system can extract medical terms from text using Named Entity Recognition. A semi-automatic algorithm to label the adverse drug reactions into three categories: minor reaction, major reaction, and deadly reactions can be developed. Further, models can be trained using classification algorithms.

References

[1] Pawlowski, C., Puranik, A., Bandi, H. et al. Exploratory analysis of immunization records highlights decreased SARS-CoV-2 rates in individuals with recent non- COVID-19 vaccinations. Sci Rep 11, 4741 (2021). https://doi.org/10.1038/s41598-021-83641-y

[2] Sarvam Mittal, 2020, An Exploratory Data Analysis of COVID-19 in India, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020)

[3] Li, Si & Yu, Chiahui & Wang, Yichuan & Babu, Yedurag. (2019). Exploring adverse drug reactions of diabetes medicine using social media analytics and interactive visualizations. International Journal of Information Management. 48. 10.1016/j.ijinfomgt.2018.12.007.

[4] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, Graciela Gonzalez,

[5] Montastruc JL, Sommet A, Bagheri H, Lapeyre-Mestre
M. Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. Br J Clin Pharmacol. 2011;72(6):905-908. doi:10.1111/j.1365- 2125.2011.04037.x https://doi.org/10.1186/s12911-021- 01402-3

[6] Ren, JJ., Sun, T., He, Y. et al. A statistical analysis of vaccine-adverse event data. BMC Med Inform Decis Mak 19, 101 (2019). https://doi.org/10.1186/s12911- 019-0818-8

[7] Mclachlan, Scott & Osman, Magda & Dube, Kudakwashe & Chiketero, Patience & Choi, Yvonne & Fenton, Norman. (2021). Analysis of COVID-19 vaccine death reports from the Vaccine

Adverse Events Reporting System (VAERS) Database Interim: Results and Analysis. 10.13140/RG.2.2.26987.26402.

[8] Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity 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 M.W. Quinn, Mohammad Ali Moni medRxiv 2021.04.16.21255618; doi: https://doi.org/10.1101/2021.04.16.21255618

[9] M. Abdulkareem, N., Mohsin Abdulazeez, A., Qader Zeebaree, D., & A. Hasan, D. (2021). COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms. Qubahan Academic Journal, 1(2), 100–105. https://doi.org/10.48161/qaj.v1n2a53

[10] A Report on the U.S. Vaccine Adverse Events Reporting System (VAERS) of the COVID-19 Messenger Ribonucleic Acid (mRNA) Biologicals Jessica Rose, PhD, MSc, BSc/

[11] https://www.ncbi.nlm.nih.gov/research/bionlp/Data/