Predicting hospitalisation after Covid-19 vaccination

Elia Al Geith
elia.al-geith@student.uni-tuebingen.de

Matthias Blum
m.blum@student.uni-tuebingen.de

Abstract

In this project, we analyze the VAERS data set. We attempt to use the available data to answer the following question: How likely is it to be hospitalized after complete vaccination? Based on the number of adverse events that can occur after full vaccination and based on patient information such as age, gender, and pre-existing conditions, we attempt to predict if a patient will be hospitalized or not. For model training, we use logistic regression and decision trees. [1]

1 General approach

The first step in our approach is reading the input files in separate data-frames. The data-frames are pre-processed by taking selected features and convert these features to numerical values which are used to compute the correlation between them and the target like hospitalisation. The last step is to train a Logistic Regression model and evaluate it.

2 Data Exploration

VAERS is used to continually monitor reports to determine whether any vaccine has a higher than expected rate of adverse events.

Disclaimer. VAERS data are from a passive surveillance system and represent voluntary and unconfirmed reports of health events occurring after vaccination. Such data are subject to the limitations of underreporting, concurrent administration of multiple different vaccines, reporting bias, and lack of incidence rates in unvaccinated comparison groups. It is important to note that no cause-and-effect relationship was established for any reported event. This incompleteness must be taken into account in the analysis.

Purpose. VAERS researchers use procedures and analytical methods that help us closely monitor vaccine safety. When a problem arises, action can be taken. The main purpose of the database is to serve as an early warning or signaling system for adverse events that were not detected during pre-market testing.

Data format. VAERSDATA provides information about the person. VAERSSYMPTOMS lists all the symptoms of that person. VAERSVAX specifies all the information about that persons vaccination status.

Pitfalls. VAERS data contain strong biases. Incidence rates and relative risks for certain adverse events cannot be calculated [4]. Note that because of the incompleteness, one cannot normalize the absolute values of the adverse events and therefore cannot compare them. This is important because various groups are trying to exploit the VAERS database to make COVID19 vaccinations look bad [3][5][2]. As we realized these pitfalls, our initial question was overthrown. Hence, we mainly stick to predicting hospitalization from VAERS features.

[1] The github repository for the paper https://github.com/Eg07/data_literacy_ws_21_22

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

To gain understanding of the data set, we visualize the target columns and the quote of vaccination among the patients.

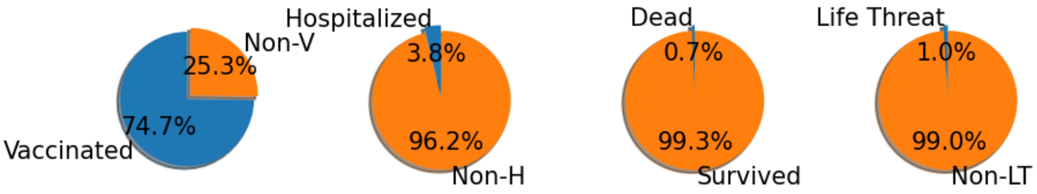

Figure 1: Number of vaccination , life-threated, deaths and, hospitalisation rate among patients

We see 74.7% of patients are vaccinated in the figure above. From this quote, 3.8% were hospitalised after taking the vaccine. 0.7% have died because of the vaccination, and 1% had a life-threatening situation

3 Data Cleaning and Pre-processing

Firstly, we start by selecting interesting columns which we think play a primary role by causing one of our target columns. After the selection, we start the data cleaning. As the last step, we preprocess the data by converting it to numerical data type to use for the correlation analysis and model fitting.

Selecting Data. After the loading each file in a separate data-frame, we select the the following features:

2021VAERSVAX. We choose only covid-19 vaccines. The selected vaccines have to be given in 3 doses fashion (like BioNTech and Moderna vacancies), or the chosen vaccine have to be JANSSEN.

2021VAERSSYMPTOMS. We select all symptoms columns and count non-null values. The count of non-null values is add to a new column called NUM_OF_SYM_SUB.

2021VAERSDATA. There are 35 VAERS features. Since not all features make sense in our context, a preselection is made. Features that have something to do with the biology or health status of the patient remain. The remaining are the following features: AGE_YRS, SEX, DIED, L_THREAT, HOSPDAYS, X_STAY, DISABLE, OTHER_MEDS, CUR_ILL, HISTORY, BIRTH_DEFECT, ALLERGIES, NUM_OF_SYM_SUB. These features describe age, gender, death after vaccination, current life-threatening illness, number of days hospitalized, prolongation of existing hospitalization, disability, current medication, current condition, chronic or long-standing illness, congenital disabilities, allergies and number of submitted symptoms.

The columns of medical history, current illnesses and, token medicines are too noise and need to be cleaned. We follow this strategy by denosing these columns:

1. Select the noised column and show the unique values
2. Build a Hash map
3. add the redundant values as a key of the hash map
4. replace the same the redundant

A sample of the hash map which we use to filter the data of the column other medicines is shown below:

```
{'none':'None', 'unknown':'None', 'No':'None','N/a':'None',
'None reported':'None', 'Na':'None'}
```

Data Pre-processing. We convert the columns with binary values into 0 1 valued columns. The age and days of hospitalisation are continuous. Therefore, they converted to float data type. The number if the symptoms is converted to integer value which is in the interval $[0, 5]$

4 Correlation Analysis

After the first calculation of the correlation of the preselected features, logically, X_STAY, L_THREAT and HOSPDAYS had the most considerable correlation. X_STAY and HOSPDAYS correlate strongly with HOSPTIAL since a vaccinee must be hospitalized for them to come true. L_threat is a significant factor because patients with life-threatening illnesses are strongly susceptible to any influence that challenges the body´s immune system. It was decided not to include these features because they affect the correlation of the other features and thus complicate their interpretation. The subsequently calculated correlation can be seen in the following figure.

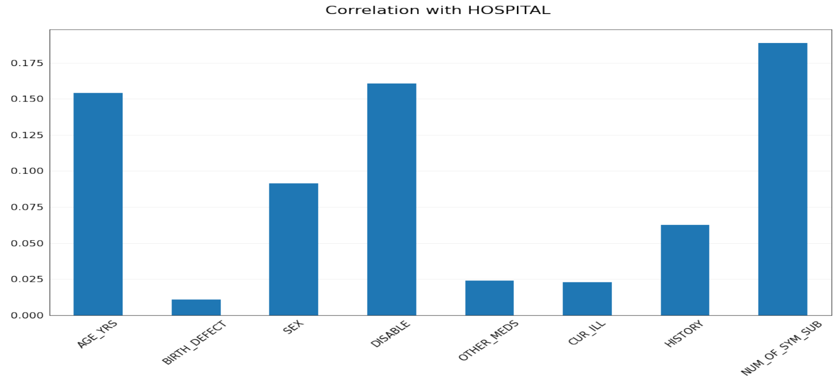

Figure 2: **Correlation of HOSPITAL and other features.** The graph shows the correlations between various VAERS features and HOSPITAL - whether a vaccinee was hospitalized or not.

It is now apparent that age, sex, disability, chronic disease, and reported symptoms correlate more strongly with hospitalization.

5 Model Training

To predict if a patient will be hospitalised or not, which is a classification problem, we use two classifiers and compare their results. The first classifier is the Logistic Regression, which we discussed during the lecture. The other classifier is the Decision Tree which is a hierarchical classifier. This kind of classifier has two main steps [1]:

- Divide the training space along feature dimensions
- Construct a decision tree

Generally, a decision tree is created according to the following algorithm:

Algorithm 1 General decision tree algorithm

```
Require: $D$ set of labeled instances; $F$ set of features
Ensure: $T$ tree with labelled leaves
  procedure GROWTREE($D, F$)
      if $D$ is homogeneous then
          return the class of $D$ as labeled class
      end if
      $S \leftarrow$ bestSplit $(D, F)$
      Split $D$ into subsets $D_i$
      for each $D_i \in D$ do
          if $D_i \neq \varnothing$ then
              $T_i \leftarrow$ GROWTREE($D_i, F$)
          elseLabel $T_i$ with the class of $D$
          end if
      end for
  end procedure
```

To train the models mentioned above, we use cross validation for parameter selecting where we first split the data set into train and test subsets. The training subset is used for fitting the model. After fitting, validation is done according to the accuracy of the model. In the case of binary classification sklearn uses the Jaccard Similarity. It states that, the number of observations in both sets is divided by the number in either set. The results of the training are shown in the table below.

model name	accuracy mean	accuracy std
Decision Tree	0.919052	0.002052
Logistic Regression	0.709235	0.003346

Table 1: Training results

As we can see, the decision tree classifier matches better than the logistic regression classifier because the decision tree classifier divides the space into smaller and smaller regions. In contrast, Logistic Regression fits a single line to divide the space precisely into two.

6 Conclusion

After becoming more familiar with the data, many of the ways we imagined working with the data set turned out to be naive. The incompleteness of the data set prevents us from performing reliable stochastic analyses to calculate, for example, the absolute probability of hospitalization based on various VAERS features of a patient. We then focused on training our models for binary prediction of hospitalization based on these features. For this purpose, the data had to be prepared elaborately. After correlation analysis, we used the most important features for training from the prepared data. Logistic regression and decision tree both showed an accuracy of over 70%. However, the decision tree performed significantly better in comparison, predicting hospitalization with about 91% accuracy. A classifier like this could certainly be used in an application.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Jessica McDonald. Increase in covid-19 vaers reports due to reporting requirements, intense scrutiny of widely given vaccines.

[3] Jonathan Jarry M.Sc. Don't fall for the 'vaers scare' tactic.

[4] Frederick Varricchio, John Iskander, Frank Destefano, Robert Ball, Robert Pless, M Miles Braun, and Robert T Chen. Understanding vaccine safety information from the vaccine adverse event reporting system. *The Pediatric infectious disease journal*, 23(4):287–294, 2004.

[5] WayneTheDBA. Vaers summary for covid-19 vaccines through 01/07/2022.