# Case study for AstraZeneca Data Science Solutions interview

## Problem Statement

Drug-drug interactions are an important type of adverse event with the seriousness ranging from mild to life-threatening. While drug-drug interactions can be tested in pharmacological interaction studies or in clinical studies, a comprehensive analysis of these interactions is possible only after the drug is adopted across the cohort it is intended for. The FDA Adverse Event Reporting System (FAERS) database contains information about the patient's clinical information, the diseases they have, the dosage of drugs used, any adverse reactions along with the seriousness and time to event information. We therefore use the information in the database to explore two questions related to drug interactions and the associated adverse reactions:

1. Which medicinal products are often taken together?
2. Do the medicinal product combinations identified in part (1) result in any statistically significant adverse reactions?

This information obtained from these questions can be used in both preclinical and clinical studies. In late stage clinical trials, the information can be used for patient selection by excluding patients who use drugs that can cause adverse events when taken with the drug under trial. Additionally, the information can also be used during drug development to determine if a particular medicinal product is often taken with other medicinal products and can therefore result in adverse reactions.

## Approach

The analysis scripts can be accessed at: https://github.com/VishwaNellore/Case_Study.

The analysis scripts are organized into a series of modules which can all be run using the script main.py. Here, the motivation behind each analysis script, assumptions made and the results are briefly described.

1. Data Pre-processing Module:

The OpenFDA API restricts the number of files that can be downloaded per day. Therefore, it was easier to write a script to download the files. The files contain a lot of information. For the purpose of answering the questions defined in the problem statement, the drugs and adverse reactions for each patient were retained for future analysis. This eased any memory requirement required locally and greatly speeded up the process of getting the data.

Furthermore, the FAERS data was examined for repeats and non-alphanumeric values, which needed to be removed as their presence can substantially alter the results.

2. Drug Combinations Module:

This module identifies the different medicinal product combinations taken by the patients whose information is reported to the FDA. The Apriori algorithm (Agarwal et. al., Fast Algorithms for Mining Association Rules, VLDB, 1994) was used to identify the medicinal product combinations. The Apriori algorithm is frequently used to study items frequently purchased together by observing customers' purchase transactions. Here this is equivalent to studying drugs frequently taken together by observing FDA patient records.

The algorithm uses a bottom up approach, which starts with the frequency at which a single item is observed. This item is then extended one item at a time. At each step, the algorithm prunes out item sets that fall below a "support", which is a user defined value indicating how many instances of an item set should be observed for it to be extended for further analysis. When no further extensions to item sets are possible, the algorithm terminates.

The Apriori algorithm therefore works on the following principles:

- All subsets of a frequent item set should also be frequent.
- If an item set is infrequent, all it's supersets will be infrequent.

Pros of the Apriori algorithm

1. It calculates exact occurrences for each drug combination with sufficient support.
2. It is easy to interpret the results of the algorithm.
3. It can be used on large item sets.

Cons of the Apriori Algorithm

1. With a large number of candidates, the algorithm can be computationally expensive.

2. Significant Adverse Reactions Module:

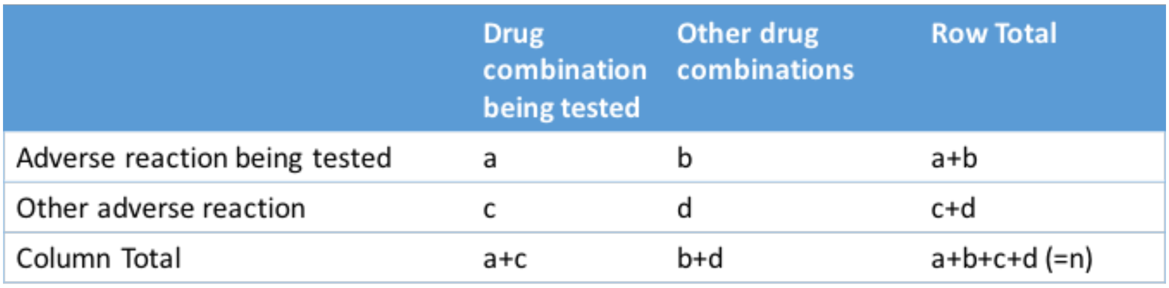
This module first creates a dictionary with the medicinal product combinations as the keys and all the reported adverse reactions for that combination as the values.

Next, to determine if the adverse reactions associated with a medicinal product combination are significant, the questions asked were as follows:

- How do we determine if an adverse reaction is significantly associated with a drug combination? For instance, if we see an increase in the adverse reaction, even across 3 or 4 patients, we associate that adverse reaction with the drug combination. But if it is only seen in a few patients, it can be because of something else.
- Do we establish significance by how many times an adverse reaction is seen across patients taking a combination of drugs? Then what is that number?

- What is a significant difference between taking a drug individually to taking a combination of drugs. As the number of drugs in a combination increases, the number of adverse reactions also increase. How do we control for the size of drug combination?
- Is this adverse reaction significant in this drug combination compared to other drug combinations?
- If we ask how often an adverse reaction is seen in a drug combination, common adverse reactions such as nausea will appear more significant than rare adverse events. Therefore, we need to estimate how often an adverse reaction is seen relative to other drug combinations. This is important with respect to the problem statement that we defined. If an adverse event comes up very often regardless of which drug combination is being used, we cannot really exclude patients easily based on that adverse event.

To provide an estimate of significance of an adverse reaction in a drug combination, after accounting for the size of the drug combination and the frequency with which the adverse reaction is seen in other drug combinations, the Fisher's exact test was used. In addition to satisfying the above requirements, the Fisher's test is exact and works even when the sample sizes are small. The contingency table to determine if an adverse reaction is association with a drug combination is shown below:



| | Drug combination being tested | Other drug combinations | Row Total |
|---|---|---|---|
| Adverse reaction being tested | a | b | a+b |
| Other adverse reaction | c | d | c+d |
| Column Total | a+c | b+d | a+b+c+d (=n) |

The p-value from the Fisher's exact is then calculated using: $p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$

The one sided Fishers exact test is used again keeping in mind the problem statement. It is important to identify if an adverse event is seen significantly more number of times in this drug combination compared to the others. This information will help in choosing patients for clinical trials.

On the other hand, if we are trying to use this data to select patients for trials, we may not be interested in why the adverse events are high for certain patients. And this means it does not matter if the patient is taking a large number of drugs and therefore reports a high number of adverse events. That patient should be excluded. Therefore, to compute a score of how significant adverse events are with a particular drug combination compared to other drug combinations, the Proportional Reporting Ratio (PRR) statistic is used. PRR is a ratio of the number of patients reporting an adverse event for a specific drug compared to the number of patients who report the same adverse event for different drugs. Using the same contingency table given above, the PRR is defined mathematically as follows:

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

Based on these results, a python script is written such that for any medicinal product that a user enters as input, an output is displayed with all the associated medicinal product combinations and the statistically significant adverse reactions.

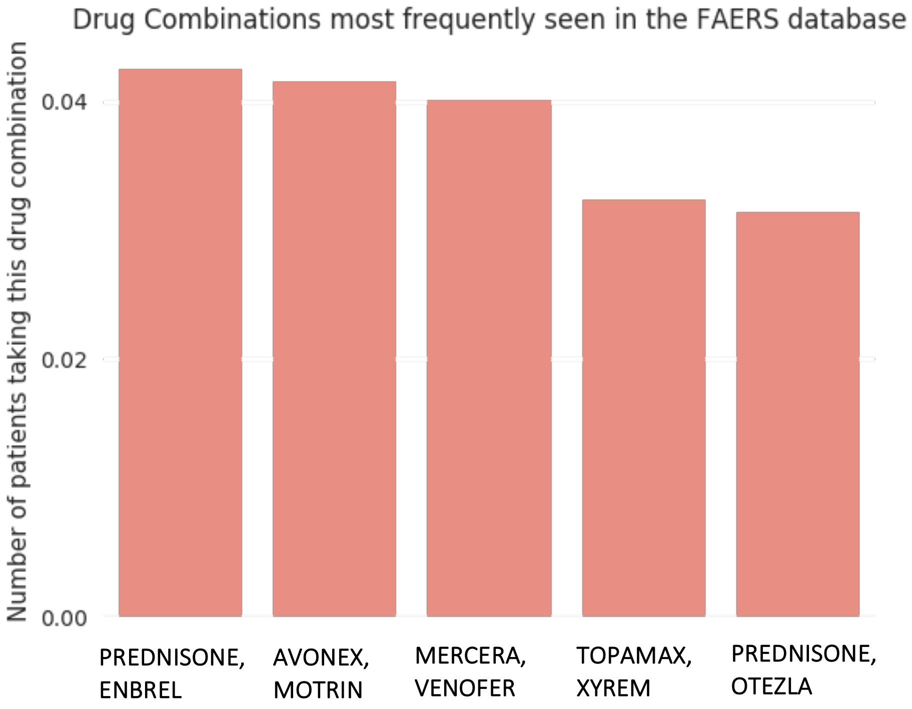
## Sanity Checks:

At several points in the pipeline of scripts, I used the following sanity checks to make sure the scripts are running as intended:

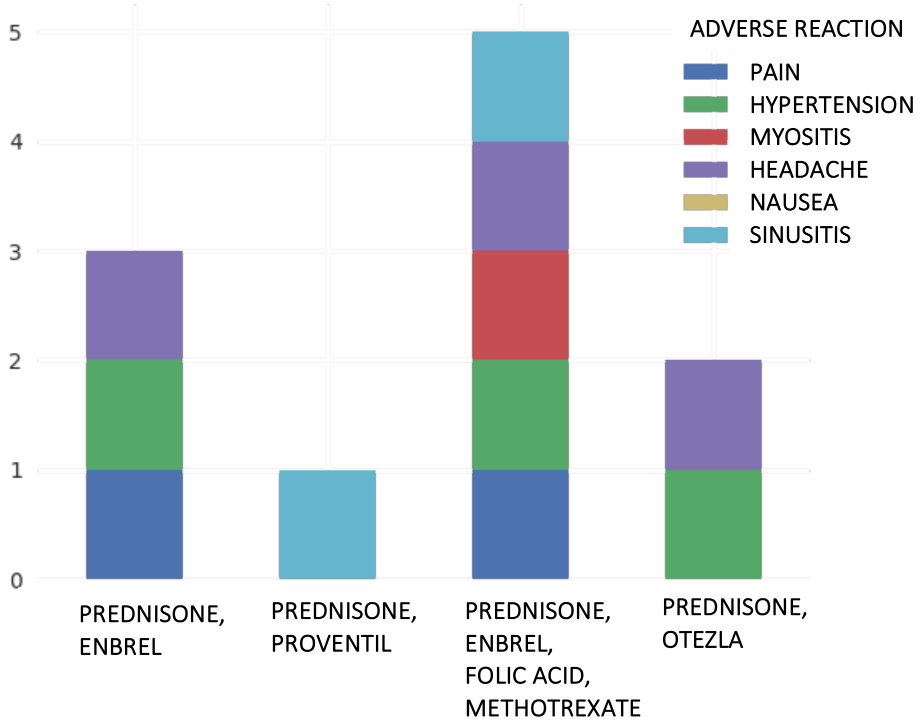
- Total number of FAERS reports
- The number of individual drugs, adverse reactions
- Number of drug combinations identified
- Number of adverse reactions that came up as significant
- The expected frequencies across the contingency table

## Results

After using the Apriori algorithm, the script shows the top medicinal product combinations and the number of patients in FAERS taking each of the medicinal product combinations as shown below. Please note that this is only an example from a limited amount of data. The full list is generated and stored in the output directory provided by the user.



For any medicinal product that a user enters as input, an output is displayed with all the associated medicinal product combinations and the statistically significant adverse reactions.



## Limitations of the Data

Several aspects the FDA database can benefit from improvements in the quantity and quality of data:

1. The analysis is restricted to the limited amount of data that is posted to the FAERS.
2. There could be errors in reporting the data.
3. As detailed on the FDA website, there is no certainty that a reported adverse event was actually due to the product used.
4. If multiple medicinal products are used in combination, the data does not provide enough detail to say with the certainty that an adverse reaction is caused by a certain medicinal product or by certain combinations of medicinal products.
5. Also the FDA website says that when a user is submitting information, if medicinal product name is not available, they can put in the active substance name. The medicinal product names should be standardized. If two or more names are used, they need to be converted to the same name or the results of this analysis will likely change.

## Future Work

### Additional Analysis:

1. Further analysis of adverse reactions associated with drug combinations can be done using the FDA database:
    a. Does the severity of the adverse events change disproportionally when a combination of medicinal products are used versus a single medicinal product.
    b. Does time to onset of adverse events change when a combination of medicinal products are used versus a single medicinal product.
2. Frequency of medicinal products taken together are only based on the data reported by FAERS. Data can also be accessed from other sources such as Electronic Health Records (EHR) using text mining. Integration of good quality data can generate results that are closer to true estimate across a cohort.

### Validation of the results:

The results presented here indicate only statistical significant associations and do not imply causality between the medicinal product combination and adverse reaction. Further rigorous analysis needs to be done to determine causality:

1. We need to test the model to make sure we are not reporting noise. Data from Drug Bank, Medscape and FDA labelling information can be mined using web-scraping and text mining to figure out if the medicinal product combinations are known to have the adverse reactions found here. Model accuracy can be described by Positive Predictive Value (PPV), Sensitivity, for each class.
2. To further determine causality for future medicinal product design, clinical trials need to be performed.

## Other Insights that can be Drawn from FAERS

1. For a medicinal product with the same indication, see if customers get the AstraZeneca brand or a competitor's brand. The results from this can be further analyzed to discover the reason customers choose a brand:
    a. Number of adverse reactions
    b. Severity of adverse reactions
    c. Access to medicinal product based on geography
    d. Access to medicinal product based on insurer from EHR data
2. As dosage of medicinal product changes, how does both number and seriousness of the adverse reactions change. This can potentially be used for future clinical trial design.
3. Are certain medicinal products/adverse events associated with gender.
4. Are adverse events proportional to age.
5. Does the severity of adverse reactions depend on the indication.