Brief summary of adopted approach

This is a brief, bulleted summary of the approach that I took to explore and analyse the OpenFDA Adverse Event Reporting System. I collated information as I progressed and these notes were for my own reference. I include it in this repository in case it is of interest.

Initial data exploration

- Become familiar with openFDA website and API
- Learn query formatting and syntax
- Examine JSON structure of returned reports
- Understand magnitude of the dataset using a combination of Jupiter notebook, direct API queries and the interactive web app on the openFDA website
- Review the complete list fields and identify potentially interesting/important fields (see Table 1)
- Rule out those which have significant numbers of blanks e.g. pharmacological classes frequently contained too little data.

Table 1: reduced list of features to consider for data analysis and modelling

Drug Active substances Action drug Additional drugs Delivery route Doseage Drug indication Treatment duration Medical product Manufacturer Pharmacological class Pharmacological effect Drug type	**Patient** Age group Age at onset of adverse effect Sex Weight Reaction outcome
Primary Source Qualification Country	**General** Report type Seriousness

Exploratory data analysis on the entire population

Due to the significant size of the dataset, obtaining and examining the entire recordset was not possible. As a result, initial exploratory data analysis consisted of repeatedly querying the database and producing some simple, descriptive plots in order to understand:

- The structure of the nested fields and database as a whole (e.g. number of records)
- The more commonly used fields (i.e. those with the most and least missing entries)
- The type of data (predominantly categorical with a few numerical fields) and the number of categories present in each field
- The way in which the database is used e.g. who contributes the data, how frequently are reports made, what is the nature of reports
- Basic properties and trends in the database based on a few highlighted features

Define the problem

- Consider a range of problems that might be interesting to stakeholders (i.e. AstraZeneca), for example:
    - Are some (classes of) drugs more likely to result in adverse responses?

    - Does manufacturer affect frequency of reported adverse effects?
    - Are certain groups of patients more likely to experience adverse effects?
    - Do certain drug delivery methods results in a higher likelihood of adverse responses?
    - What factors determine the severity of an adverse effect?
- For these questions, determine which have sufficient data to address the problem and are achievable in the timeframe given limited resources
- Decide on an exploration into factors affecting the severity of adverse responses in young people. Break this down into the following structure:
    - Exploratory data analysis examining trends in patient and drug characteristics
    - Detailed modelling of pediatric dataset

Data collection and cleanup

- Queries via the API limited to return 100 reports at a time so define a function to repeatedly query the API to build up a dataset of required size
- Import data into pandas dataframe and define functions to collapse nested lists/dictionaries to obtain a flattened dataframe
- Clean the data:
    - Select features with potential impact on outcome (severity of response) and drop other columns
    - Map missing values to 0 where appropriate
    - Drop any columns with >40% missing values
    - Reformat data types
    - Format ages as years
    - Remove outliers
    - Convert 1s and 2s to 1s and 0s

Paediatric Data Analysis

- Repeat the process of EDA for this new data which is a subset of the total population
- Conduct descriptive statistics on the dataset and analyse age and sex distributions
- Consider the features that may influence severity of adverse response in children
- Conduct hypothesis testing using Pearson chi^2 tests and z tests to determine correlations with severity at 99% significance level

---

Modelling

- Consider the necessity for predictive modelling in this task
- Consider the utility and relative complexity of implementation of a range of models
- Decide to implement a simple ML model for illustrative purposes using logistic regression