<https://www.fda.gov/food/compliance-enforcement-food/cfsan-adverse-event-reporting-system-caers>

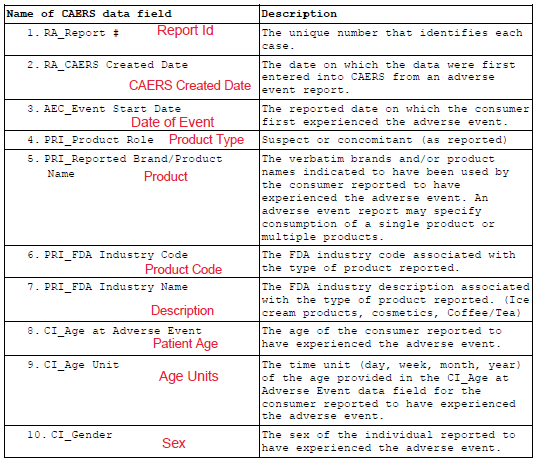
The dataset contains **‘Adverse event reports’** of **‘Dietary supplements, Cosmetics, Foods, and other Products**’ submitted in a database by either consumers (patient/parents/doctors) and health care practitioners or sometimes both.

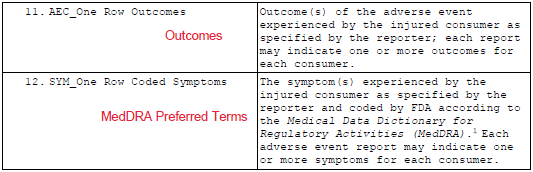
There are two datasets in this site. The variables in both the datasets are same.

Dataset 1: One dataset has records from Jan-2004 to Dec-2013 (**12 variables, 80024 observations**)

Dataset 2: Second dataset has records from Jan-2014 to Mar-2019 (**12 variables, 50055 observations**).

Variable Description (the one in red is the name mentioned in the downloaded csv files)





**4 ALGORITHMS: mandatory 4 algorithms has to be done in R**

**Questions of Interest:**

**# Report ID**

* Total number of reports reported to FDA
* Does total number of reports corresponds to total number of patients facing issues
* Are there any chances of duplicate reports since it is reported either by the patient/consumer or their HCP (healthcare professional/doctor)
* Are there reports of patients who have consumed multiple products and experienced adverse events? Can something like a system could be built to identify them?

**# CAERS Created Date**

* What is the change in number of reports reported against year?
* Has there been any drastic change in numbers? If so why?
* Find the difference between the time the report was created and the actual time of Adverse event and check the difference in time for different age range.
* Which day/week of the week/month the AEs reported are more in numbers? Is there a tendency/trend in people to report as the week goes on (reasons might be- trying to pass time on weekends, bad weekend, hangover and over eating on weekends, start of the week)

**# Date of Event**

* Has the number of adverse events increased or decreased over years?
* If it has increased, is it because more events are occurring or more events are being reported? And vice versa (in case it has decreased)

**# Product Type**

* Which type (suspect/concomitant) has more counts of adverse events?
* Can a causality be concluded on suspect/concomitant wrt to adverse events (MedDRA)

**# Product/Product Code**

* Total number of products reported
* Most common product against which Adverse events has been reported maximum
* Does the most common product in the list means the one which gets sold much?
* Has the report for the most common product has been consistent over the years (2004-2019, Mar)...Can do Time series
* Form clusters based on outcome
* A patient can have one Report ID with various Adverse events, product. So here, classify products based on Report Id for individual patients.
* Predict and forecast products sold
* Occurrences of common product based on key word in Description
* Most selling product and least selling product
* Which Product code has the maximum number of products assigned?
* How has the product evolved since 2004 to present? Has there been a change in taste, selection of product etc?

**# Description**

* What was the most common 'Description' wrt to Adverse event?
* Sentiment analysis of Product description
* Does each product description correspond to individual product code and product?

**# Patient Age**

* Which age was the most affected ones in the list?
* Find the min and max age, and then see - Can age be grouped into Age group class and check for the most affected group?
* What was the most common Adverse event in all the age group? Was the Adverse event against the same product or different in different age group? Which type of customers bought it more?
* Which age range seems to be the most dangerous?
* How babies of different age are affected? Is it because they really experience AEs or is it because their parents are worried?
* Which age group suffers the most severe outcomes?

**# Sex**

* Which gender has reported more Adverse events?
* Common AE reported by male and females respectively
* Common product against which AE has been reported (both male and female)

**# MedDRA Preferred Terms**

* Total number of adverse events reported
* The most common AE/Symptoms
* Did all the common symptoms take place due to the same product?
* AE as per year/month/weekday
* AEs per product
* Can the number of AEs reported be predicted based on present?
* Based on MedDRA PT term classify these based on the 'System Organ Class'
* Were the reported adverse events more or the number of patients/consumers more?

**# Outcomes**

* Some of the outcome seems to be interlinked. Need to find out such types and find out the best possible way to classify
* Classification can be done based on severity
* Clustering based on serious and non-serious.
* Which age group has the maximum serious counts? Can the count be predicted based on present? (Maybe Null hypothesis)
* The data has mentioned about the condition of patient and has also stated for some if treatment was offered. Classification/Clustering can be made based on this, as in which age group preferred treatment etc.
* Check if SVM works a model designed.