# Pre-Analysis Notes

The original dataset was extracted from the US FDA webpage. It consists on 3 files in txt format:

* exclusivity\_raw.txt
* patents\_raw.txt
* products\_raw.txt

Being my first health data project, it seems an interesting dataset to start with as it appears to be both manageable and sufficiently challenging. I hope this analysis brings impactful insights on FDA-approved products patents.

In order to come up with insightful questions to answer when analysing the dataset is crucial to understand all data fields. For that reason, we first revised the metadata provided by the FDA (see Metadata\_FDA\_Orange\_Book).

The aim of this project is to answer several questions:

* are those FDA-Approved products whose patent is almost to expire, which are the

The first thing to check was the meaning of each column, which was already posted on Kaggle. I have incorporated these metadata in a pdf document (see metadata folder).

The next step was to dive into the dataset. To explore the dataset we opted to use pandas, a python library very useful for data manipulation and data transformation (see PreAnalysis.