Eudra-Vigilance Data Project

By Craig Paardekooper (September 13th 2024)

WHAT IS EUDRA-VIGILANCE ?

Eudra-vigilance is the European equivalent of VAERS – it is a database of adverse reaction reports for all vaccines, substances and products used in the European Union (includes Astazeneca data) since 2001.

URL : https://www.adrreports.eu/en/disclaimer.html

PURPOSE OF THE EUDRA-VIGILANCE DATA PROJECT

The purpose of this project is to provide public access to Eudra-vigilance data for analysis of drug safety. The resulting CSV dataset is the largest **publicly available** dataset of its kind in the world, consisting of input from **4,497,850 unique people**, covering 19,265 unique symptoms relating to 4122 unique drug brand names, and 4387 unique ingredient lists, some being variations for the same drug.

SINGLE DRUG EFFECTS

The dataset consists of **7,044,225** individuals who submitted reports.

However, I only look at those reports where a single drug was administered to an individual .There are **5,333,575** reports where only 1 drug is administered.

And I only look at reports where the name of the drug is provided. There are **4,497,850** reports where the name of the drug is provided.

Finally, I applied a minimal sample size of 100 symptom records per drug, and excluded all drugs with fewer records. This was so I could draw reliable statistical conclusions.

The final dataset has –

1. **4,486,467** individual people records
2. **14,987,664** drug-symptom associations –
3. **19,247** unique symptoms
4. **1,504** unique drug brand names, and
5. **3,163** unique ingredient lists.
6. Each drug is represented by a name consisting of a single word

This dataset will empower analysts to rank many drugs by the incidence of any adverse symptom – and aide informed choice.

Several practical examples will be provided to demonstrate the reliability of this dataset in identifying drugs with high adverse outcomes.

Downloads

Original set of 4088 csv files - here

Data files concatenated into one big file - here

One big file after data has been cleaned and organised – here

A fourth dataset is available. It is a pivot table showing the safety signals for any of 19,245 symptoms for 1504 different drugs. It also shows every drug sorted in order of its safety signal for any chosen symptom. This dataset has a size of only 4Mb (zipped).

Download

Search

A search engine has been created that ranks drugs by the incidence of any chosen symptom or provides a list of all symptoms for any drug.

Search Dataset - here

AUTHENTICITY OF EACH RECORD

ICSR means "Individual Case Safety Report" – showing that each row is an individual person. There are **4,486,467** unique case reports in the final dataset.

You can view the report hosted on the EMA website, by putting the ID number into the URL below.

https://dap.ema.europa.eu/xmlpserver/PHV%20DAP/Reports/ICSR.xdo?_xpf=&_xt=form&SR_ID=10011128660&_xpt=1&_xf=pdf

This is the URL for the report with ID 10011128660. It is a pdf hosted on the European Medical Association website.

This shows that the downloaded data is not made up. Rather, each row of data has a report that you can verify using the link provided.

Each downloaded report looks like this –

EVPM ICSR(s) **Individual Case Safety Report Form** EudraVigilance

General Information

EudraVigilance Local Report Number	EU-EC-10011128660
Sender Type	Not available
Sender's Organisation	Senders Organisation is not displayed
Type of Report	Spontaneous
Primary source country	European Economic Area
Reporter's qualification	Non-Healthcare Professional
Case serious?	No

Patient

Age Group	Age Group (as per reporter)	Sex
18-64 Years		Female

Reaction / Event

MedDRA LLT	Duration	Outcome	Seriousness[1]
Menstrual cycle abnormal		Not Recovered/Not Resolved	
Tachycardia		Not Recovered/Not Resolved	
Vision blurred		Not Recovered/Not Resolved	
Pain chest		Not Recovered/Not Resolved	

Drug Information

Role[2]	Drug	Duration	Dose	Units in Interval	Action taken
S	COMIRNATY - TOZINAMERAN		1.0 {DF}	Total	Not applicable

Drug Information *(cont.)*

Info[3]	Drug	Indication	Pharm. Form	Route of Admin.
	COMIRNATY - TOZINAMERAN	COVID-19 immunisation		Intramuscular use

WHO INPUT THE DATA ?

In the original dataset consisting of **7,044,225** individual reports, there is a QUALIF column that provides the credentials of the inputers.

```
1  f1['QUALIF'].value_counts()
```

Healthcare Professional	4492180
Non Healthcare Professional	2523665
Not Specified	28380

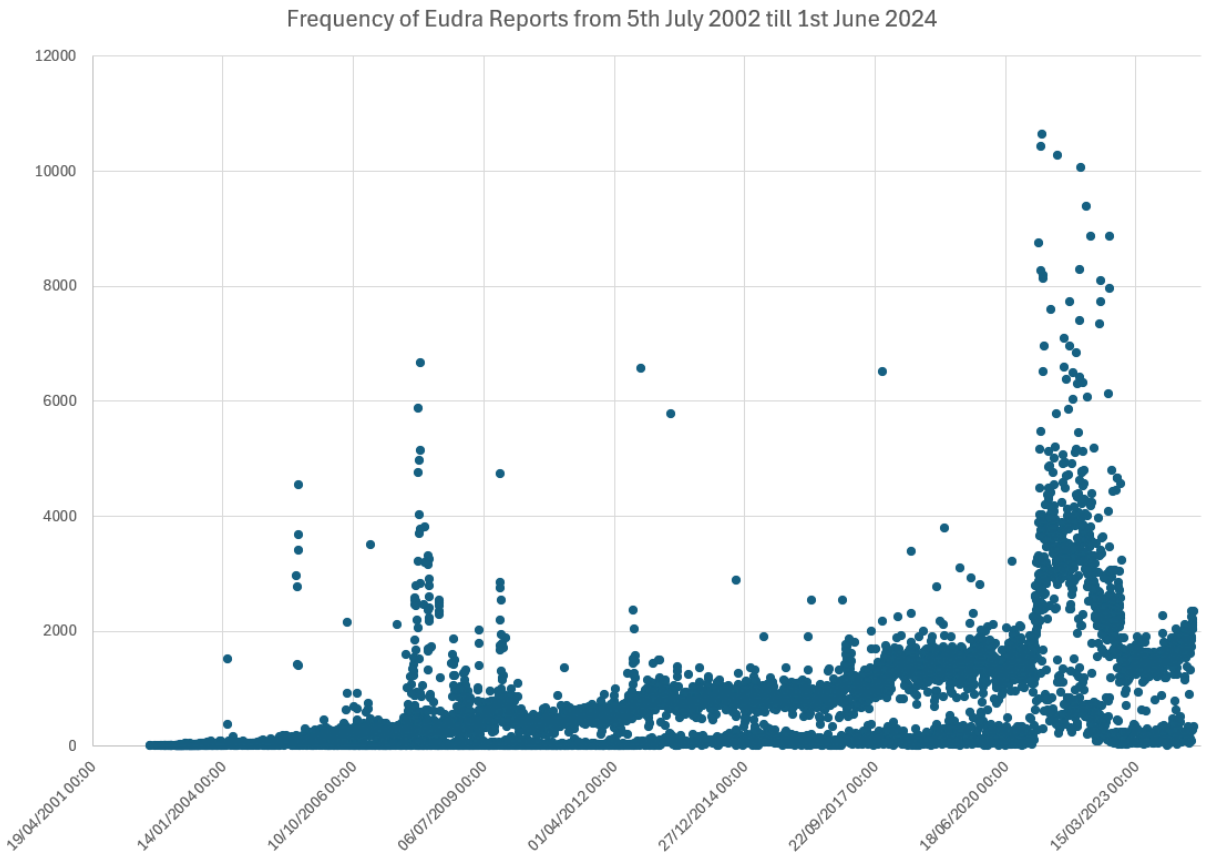
So most reporters were healthcare professionals

NOTE : 63% of reports in Eudra were submitted by healthcare **professionals**, and a further 35% of reports were submitted by non-healthcare **professionals** - most likely admin staff in the employ of hospitals, private medical practices, or care services. So this data is to be taken seriously.

```
DateRange = frame['DATE'].value_counts().reset_index().rename(columns={"index": "DATE", "DATE": "FREQ"})
DateRange.to_csv(r"C:\Users\User\Documents\AAAcovid\dates.csv")
```

This is the date when the report was received. Reports cover a time span from 5th July 2002 till 1st June 2024, a period of 22 years.

The spread of dates is interesting -



DATE	FREQ	>1000
01/06/2024 00:00	329	
31/05/2024 00:00	2342	>1000
30/05/2024 00:00	2201	>1000
29/05/2024 00:00	2147	>1000
28/05/2024 00:00	2032	>1000
27/05/2024 00:00	2112	>1000
25/05/2024 00:00	272	
24/05/2024 00:00	2047	>1000
23/05/2024 00:00	2038	>1000
22/05/2024 00:00	2342	>1000
21/05/2024 00:00	2242	>1000
20/05/2024 00:00	1725	>1000
19/05/2024 00:00	79	
18/05/2024 00:00	252	
17/05/2024 00:00	2101	>1000
16/05/2024 00:00	1882	>1000
15/05/2024 00:00	1900	>1000
14/05/2024 00:00	1843	>1000
13/05/2024 00:00	1705	>1000
12/05/2024 00:00	124	
11/05/2024 00:00	319	
10/05/2024 00:00	1709	>1000
09/05/2024 00:00	1610	>1000
08/05/2024 00:00	1839	>1000
07/05/2024 00:00	2011	>1000
06/05/2024 00:00	1316	>1000
05/05/2024 00:00	38	
04/05/2024 00:00	84	
03/05/2024 00:00	1963	>1000
02/05/2024 00:00	1842	>1000
01/05/2024 00:00	892	
30/04/2024 00:00	2081	>1000
29/04/2024 00:00	1792	>1000

The graph appears to split into two streams, a lower one and an upper one. The lower one is made up of weekend dates when input was less, and the upper is made of weekday dates, when input was greater.

I have created a spreadsheet of the frequency of reports for each date from 2022 till 2024. You can download the spreadsheet here.

Notice that the high frequency of reports comes in clusters of 5 days. This is because data inputting mainly occurs during weekdays rather than at weekends.

This is further evidence that data is being input by professionals working in the health industry, rather than by ordinary members of the public, who would tend to input during out-of-work hours or at the weekend.

STEPS FOR THE CREATION OF THE DATASET

All the data is publicly available, but the following steps automate the gathering of it and the organising of it into useful fields. I have completed all the steps for you. I am describing the steps in case you wish to repeat / confirm them for yourself.

Step 1 : Data Download

Wouter Aukema (http://www.aukema.org/) wrote code for automatically reading and downloading the data from Eudra-Vigilance. The download takes about 4 hours using high speed internet (30 mb / sec). The python code used for downloading is available here - python . The downloaded data consisted of 4088 csv files.

If you are not a Python programmer, you can skip downloading all the files yourself, since I have already done that and made all 4088 files available as a downloadable zip file here - Dropbox (complete unformatted tables)

Step 2 : Reading the Files with Python

Once I had the files downloaded, the next stage was to read them. They were all .csv files, however the files did not all use the same encoding. Some csv files used "ISO-8859-1" encoding, other files used "UTF-8" encoding, others used "windows 1292" encoding, and for some files the encoding was unknown.

To find out which ones I could open, I looped through the files, trying to read each one with one encoding, then another – here - file reading.

In many cases the files would not open in Python, so I had to open each one manually by importing it into an excel spreadsheet, then saving the sheet as a csv. The csv files were then read into Python and concatenated into one large file. The method of concatenation is shown here - manual conversion

You can skip these step too, if you wish, since I have done it for you.

Here is the code I used find out which encoding each file had -

```
import chardet

fname = r"C:\Users\User\Documents\ABCOVID\substances_21755_2019_Serious_European_Economic_Area_Healthcare_Professional_Femal

# Step 2: Read CSV File in Binary Mode
with open(fname, 'rb') as f:
    data = f.read()

# Step 3: Detect Encoding using chardet Library
encoding_result = chardet.detect(data)

# Step 4: Retrieve Encoding Information
encoding = encoding_result['encoding']

# Step 5: Print Detected Encoding Information
print("Detected Encoding:", encoding)
```

Otherwise I used this process for detecting encoding –

```
import pandas as pd
import seaborn as sb
import time
time.sleep(3)
import os, re

counterT = 0
counterA = 0
counterB = 0
badlistCov = []
goodlistCov = []
frameListCov = []
for root,dirs,files in os.walk(r"C:\Users\User\Documents\ABCOVID"):
    for fname in files:

                try:
                        frame = pd.read_csv(r"C:\Users\User\Documents\ABCOVID\\" + fname, sep="\t", encoding = "Windows-1252
                        counterA += 1
                        frameListCov.append(frame)
                        goodlistCov.append(fname)
                except:
                        counterB += 1
                        badlistCov.append(fname)


DataFrameCov = pd.concat(frameListCov)
print(counterA)
print(counterB)
```

As you can see, this code told me which files could be opened using a specific encoding, and which files could not. In addition, it concatenated the files that could be opened into one big file.

Repeating this with other encodings eventually led to the creation of one big file out of the many.

STEP 3 Concatenating the files into one massive file

Concatenation of the files was performed when reading the files. You can see that as each file is read, it is appended to a framelist, and this is eventually written to a csv file - file reading.

However, I could also do the concatenation separately like this -

I create a master dataframe

```
import pandas as pd
#create a master dataframe

master_dataframe = pd.DataFrame()

print(master_dataframe.shape[0])
```

Then I loop through, reading each file into a dataframe of its own, and concatenating them to the master dataframe.

```
for x in range(1,113):

    data = pd.read_csv(r"C:\Users\User\Documents\AAAcovid\cov" + str(x) + ".csv", encoding="ISO-8859-1")

    frames = [master_dataframe, data]
    master_dataframe = pd.concat(frames)
    print("File " , x , " - " ,master_dataframe.shape[0])
```

The final master dataframe is then saved to a csv file

```
master_dataframe.to_csv(r"C:\Users\User\Documents\AAAcovid\master.csv")
```

Step 4 : Looking at the Data Fields

Here are all the fields in the original downloaded csvs.

	EU Local Number	Report Type	EV Gateway Receipt Date	Primary Source Qualification	Primary Source Country for Regulatory Purposes	Literature Reference	Patient Age Group	Patient Age Group (as per reporter)	Parent Child Report	Patient Sex	Reaction List PT (Duration – Outcome - Seriousness Criteria)	Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	ICSR Form
0	EU-EC-10000301503	Spontaneous	2017-12-26 00:00:00	Non Healthcare Professional	European Economic Area	Not available	Not Specified	Adult	Yes	Female	Abdominal pain (n/a - Recovering/ Resolving -)...	PANTOPRAZOLE [PANTOPRAZOLE, PANTOPRAZOLE SODIU...	[OMEPRAZOLE] (C - Product used for unknown ind...	<a target="_blank" href="https://dap.ema.europ...
1	EU-EC-10000269438	Spontaneous	2017-12-21 00:00:00	Non Healthcare Professional	European Economic Area	Not available	65-85 Years	Elderly	No	Female	Hyperchlorhydria (n/a - Not Recovered/Not Reso...	[ALUMINIUM HYDROXIDE GEL, DRIED, MAGNESIUM CAR...	[BISOPROLOL FUMARATE, BISOPROLOL FUMARATE, ACE...	<a target="_blank" href="https://dap.ema.europ...
2	EU-EC-10000251211	Spontaneous	2017-12-19 00:00:00	Non Healthcare Professional	European Economic Area	Not available	Not Specified	Not Specified	No	Female	Abdominal pain (n/a - Not Recovered/ Not Resolv...	PANTOPRAZOLE [PANTOPRAZOLE, PANTOPRAZOLE SODIU...	[OMEPRAZOLE] (C - Gastrooesophageal reflux dis...	<a target="_blank" href="https://dap.ema.europ...
3	EU-EC-10000253124	Spontaneous	2017-12-19 00:00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Diffuse alopecia (n/a - Not Recovered/ Not Reso...	[RANITIDINE HYDROCHLORIDE] (S - Oesophagitis -...	Not reported	<a target="_blank" href="https://dap.ema.europ...
4	EU-EC-10000010647	Spontaneous	2017-11-22 00:00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Adult	No	Female	Gastritis (n/a - Recovered/ Resolved -)	[RANITIDINE, RANITIDINE HYDROCHLORIDE] (S - Ga...	Not reported	<a target="_blank" href="https://dap.ema.europ...

Here is a list of the columns

- Unnamed: 0
- EU Local Number
- Report Type
- EV Gateway Receipt Date
- Primary Source Qualification
- Primary Source Country for Regulatory Purposes
- Literature Reference
- Patient Age Group
- Patient Age Group (as per reporter)
- Parent Child Report
- Patient Sex
- Reaction List PT (Duration – Outcome - Seriousness Criteria)
- Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])
- Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])
- ICSR Form
- Reaction List PT (Duration – Outcome - Seriousness Criteria)

Step 5 : Identifying Null Values

As you can see, every other row contained null values.

```
import pandas as pd

frame = pd.read_csv(r"C:\Users\User\Documents\AAAcovid\total.csv")

frame.head()
```

EU Local Number	Report Type	EV Gateway Receipt Date	Primary Source Qualification	Primary Source Country for Regulatory Purposes	Literature Reference	Patient Age Group	Patient Age Group (as per reporter)	Parent Child Report	Patient Sex	Reaction List PT (Duration Outcome - Seriousness Criteria)	Suspect/ interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	Conco Admini Dr (Drug Indicat - [Dur F
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
EU-EC-10011128660	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Chest pain (n/a - Not Recovered/ Not Resolved -...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
EU-EC-10011128692	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Fatigue (n/a - Recovered/ Resolved With Sequela...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Counting Not-Nulls

```
frame.notnull().sum()
```

There are 10,882,388 rows in the initial dataset.

```
Unnamed: 0                                                                                    10882388
Unnamed: 0.1                                                                                  10882388
EU Local Number                                                                                8794672
Report Type                                                                                    8794672
EV Gateway Receipt Date                                                                        8794672
Primary Source Qualification                                                                   8794672
Primary Source Country for Regulatory Purposes                                                 8794672
Literature Reference                                                                           8794672
Patient Age Group                                                                              8794672
Patient Age Group (as per reporter)                                                            8794672
Parent Child Report                                                                            8794672
Patient Sex                                                                                    8794672
Reaction List PT (Duration  Outcome - Seriousness Criteria)                                    2087716
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    8794672
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    8794672
ICSR Form                                                                                      8794672
Reaction List PT (Duration – Outcome - Seriousness Criteria)                                     47471
Reaction List PT (Duration â€“ Outcome - Seriousness Criteria)                                   1381913
The query resulted in no rows                                                                        0
Reaction List PT (Duration â Outcome - Seriousness Criteria)                                   5277572
```

```
nan_count = frame.isna().sum()
print(nan_count)
```

```
EU Local Number                                                                    2087716
Report Type                                                                        2087716
EV Gateway Receipt Date                                                            2087716
Primary Source Qualification                                                       2087716
Primary Source Country for Regulatory Purposes                                     2087716
Literature Reference                                                               2087716
Patient Age Group                                                                  2087716
Patient Age Group (as per reporter)                                                2087716
Parent Child Report                                                                2087716
Patient Sex                                                                        2087716
Reaction List PT (Duration  Outcome - Seriousness Criteria)                        8794672
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    2087716
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    2087716
ICSR Form                                                                          2087716
Reaction List PT (Duration – Outcome - Seriousness Criteria)                       10834917
Reaction List PT (Duration â€“ Outcome - Seriousness Criteria)                       9500475
The query resulted in no rows                                                      10882388
Reaction List PT (Duration â Outcome - Seriousness Criteria)                       5604816
```

So, the number of records in each column where the value is not null is given by –

```
frame.count()
```

```
EU Local Number                                                                    8794672
Report Type                                                                        8794672
EV Gateway Receipt Date                                                            8794672
Primary Source Qualification                                                       8794672
Primary Source Country for Regulatory Purposes                                     8794672
Literature Reference                                                               8794672
Patient Age Group                                                                  8794672
Patient Age Group (as per reporter)                                                8794672
Parent Child Report                                                                8794672
Patient Sex                                                                        8794672
Reaction List PT (Duration  Outcome - Seriousness Criteria)                        2087716
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    8794672
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    8794672
ICSR Form                                                                          8794672
Reaction List PT (Duration – Outcome - Seriousness Criteria)                       47471
Reaction List PT (Duration â€“ Outcome - Seriousness Criteria)                       1381913
The query resulted in no rows                                                      0
Reaction List PT (Duration â Outcome - Seriousness Criteria)                       5277572
```

Step 6 : Removing Null Values

```
frame.dropna(subset=['EU Local Number'], inplace=True)
frame.head()
```

The EU Local Number is the report identification number. When we drop all NaN values from this column, we get

EU Local Number	Report Type	EV Gateway Receipt Date	Primary Source Qualification	Primary Source Country for Regulatory Purposes	Literature Reference	Patient Age Group	Patient Age Group (as per reporter)	Parent Child Report	Patient Sex	Reaction List PT (Duration Outcome - Seriousness Criteria)	Suspect/ interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	Concor Admini Dr (Drug Indicat - t [Dur R
EU-EC-10011128660	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Chest pain (n/a - Not Recovered/ Not Resolved -...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re
EU-EC-10011128692	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Fatigue (n/a - Recovered/ Resolved With Sequela...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re
EU-EC-10011128700	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Asthenia (n/a - Recovered/ Resolved -), <BR...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re
EU-EC-10011128707	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Menstrual disorder (n/a - Recovering/ Resolving...	COMIRNATY [TOZINAMERAN] (S - COVID-19 prophyla...	Not re
EU-EC-10011128714	Spontaneous	31/12/2021 00:00	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	Capillary fragility (n/a - Recovering/ Resolvin...	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not re

The counts for this column now match that for the other columns

Unnamed: 0	8794672
Unnamed: 0.1	8794672
EU Local Number	8794672
Report Type	8794672
EV Gateway Receipt Date	8794672
Primary Source Qualification	8794672
Primary Source Country for Regulatory Purposes	8794672
Literature Reference	8794672
Patient Age Group	8794672
Patient Age Group (as per reporter)	8794672
Parent Child Report	8794672
Patient Sex	8794672
Reaction List PT (Duration Outcome - Seriousness Criteria)	2087716
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	8794672
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	8794672
ICSR Form	8794672
Reaction List PT (Duration – Outcome - Seriousness Criteria)	47471
Reaction List PT (Duration â€“ Outcome - Seriousness Criteria)	1381913
The query resulted in no rows	0
Reaction List PT (Duration â Outcome - Seriousness Criteria)	5277572

Merging the 4 REACTION Columns

So, there are 4 columns where Reaction is specified, which contain the following numbers of records -

- 2087716
- 47471
- 1381913
- 5277572

Totalling 8,794,672

These 4 separate columns should be merged into a single column.

Before we can work with this data, we must merge the 4 reaction columns into a single column called REACTION, and then drop the 4 original reaction columns.

Here, I created a new dataframe called frame1, in case there were any issues with the merge, I did not want it to effect the original frame.

```
frame1 = frame.assign(REACTION = frame[frame.columns[12]].astype(str) + \
frame[frame.columns[16]].astype(str) + frame[frame.columns[17]].astype(str) + \
frame[frame.columns[19]].astype(str))
```

```
frame1 = frame1.drop([frame1.columns[1], frame1.columns[12],frame1.columns[16],frame1.columns[17], + \
frame1.columns[18],frame1.columns[19]], axis=1)
frame1.head()
```

	EU Local Number	Primary Source Qualification	Primary Source Country for Regulatory Purposes	Literature Reference	Patient Age Group	Patient Age Group (as per reporter)	Parent Child Report	Patient Sex	Suspect/ interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	Concomitant/ Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])	ICSR Form	REACTION
	EU-EC-10011128660	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not reported	<a target="_blank" href="https://dap.ema.europ...	Chest pain (n/a - Not Recovered/ Not Resolved -...
	EU-EC-10011128692	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not reported	<a target="_blank" href="https://dap.ema.europ...	Fatigue (n/a - Recovered/ Resolved With Sequela...
	EU-EC-10011128700	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	COMIRNATY [TOZINAMERAN] (S - COVID-19 immunisa...	Not reported	<a target="_blank" href="https://dap.ema.europ...	Asthenia (n/a - Recovered/ Resolved -), <BR...
	EU-EC-10011128707	Non Healthcare Professional	European Economic Area	Not available	18-64 Years	Not Specified	No	Female	COMIRNATY [TOZINAMERAN] (S - COVID-19 prophyla...	Not reported	<a target="_blank" href="https://dap.ema.europ...	Menstrual disorder (n/a - Recovering/ Resolving...

I confirmed that there were no longer any nulls in any of the columns

```
frame1.isnull().sum()
```

```
Unnamed: 0                                                                      0
EU Local Number                                                                 0
Primary Source Qualification                                                    0
Primary Source Country for Regulatory Purposes                                  0
Literature Reference                                                            0
Patient Age Group                                                               0
Patient Age Group (as per reporter)                                             0
Parent Child Report                                                             0
Patient Sex                                                                     0
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])      0
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    0
ICSR Form                                                                       0
REACTION                                                                        0
```

All columns were now of the same length

```
frame1.notnull().sum()
```

```
Unnamed: 0                                                                      8794672
EU Local Number                                                                 8794672
Primary Source Qualification                                                    8794672
Primary Source Country for Regulatory Purposes                                  8794672
Literature Reference                                                            8794672
Patient Age Group                                                               8794672
Patient Age Group (as per reporter)                                             8794672
Parent Child Report                                                             8794672
Patient Sex                                                                     8794672
Suspect/interacting Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])      8794672
Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - [Duration - Dose - Route])    8794672
ICSR Form                                                                       8794672
REACTION                                                                        8794672
```

Step 7 : Renaming Columns

To facilitate coding, I renamed these columns with easier names.

```
frame1.rename(columns = {frame1.columns[0]:'INDEX'}, inplace = True)
frame1.rename(columns = {frame1.columns[1]:'ID'}, inplace = True)
frame1.rename(columns = {frame1.columns[2]:'DATE'}, inplace = True)
frame1.rename(columns = {frame1.columns[3]:'QUALIF'}, inplace = True)
frame1.rename(columns = {frame1.columns[4]:'COUNTRY'}, inplace = True)
frame1.rename(columns = {frame1.columns[5]:'LIT'}, inplace = True)
frame1.rename(columns = {frame1.columns[6]:'AGE'}, inplace = True)
frame1.rename(columns = {frame1.columns[7]:'REP-AGE'}, inplace = True)
frame1.rename(columns = {frame1.columns[8]:'PARENT-REPORT'}, inplace = True)
frame1.rename(columns = {frame1.columns[9]:'GENDER'}, inplace = True)
frame1.rename(columns = {frame1.columns[10]:'DRUG'}, inplace = True)
frame1.rename(columns = {frame1.columns[11]:'CONCOM'}, inplace = True)
frame1.rename(columns = {frame1.columns[12]:'ICSR'}, inplace = True)
```

INDEX	8794672
ID	8794672
DATE	8794672
QUALIF	8794672
COUNTRY	8794672
LIT	8794672
AGE	8794672
REP-AGE	8794672
PARENT-REPORT	8794672
GENDER	8794672
DRUG	8794672
CONCOM	8794672
ICSR	8794672
REACTION	8794672

There were no issues with this stage, so I updated the original frame

```
frame = frame1.copy()
```

****

The concatenated data-frame is available as a download from Dropbox here.

Step 8 : Removing Duplicates

So we have 8,794,672 reports. I wanted to see if any records have been duplicated, which might happen when I was concatenating so many separate files. Duplicates may have arisen during the process of combining the 4088 csv files, if these csv files had any overlap.

In order to remove duplicates I used –

```
frame.drop_duplicates(inplace=True)
frame.count()
```

Each individual person has a unique ID. There are 7,044,225 unique IDs in the dataset.

```
frame['EU Local Number'].nunique()
```

```
7044225
```

So 7,044,225 reports remained after duplicates were removed.

Step 9 : Dropping Unnecessary Columns

My aim in this analysis is to rank all drugs according to the incidence of their effects. So to do this analysis, I only needed the ID column, DRUG column and the REACTION column. I drop all the other columns. This is done to save memory and avoid the computer crashing.

frame = frame[[frame.columns[1], frame.columns[10],frame.columns[13]]].copy()

Step 10 : Filtering for Single-drug Administrations

All rows where the DRUG column contains multiple drugs are removed. This is done to enable clearer identification of relationships between symptom and drug.

**frame = frame[frame['DRUG'].str.contains('

')==False]**

There are **5,333,575** reports where only 1 drug is administered

Step 11 : Splitting the DRUGS column into INGREDIENTS and DRUGS

The original DRUG column contains the drug name and its ingredients, so I split the DRUG column into 2 columns – DRUG and INGREDIENTS

frame['NAME'] = frame['DRUG'].str.split('\[', expand = True)[0]

frame['INGREDIENTS'] = frame['DRUG'].str.split('\[', expand = True)[1]

Step 12 : Splitting the INGREDIENTS column into PATHOLOGY and INGREDIENTS

Then the INGREDIENTS column is further split into 2 columns – PATHOLOGY and INGREDIENTS

frame['INGREDIENT-LIST']=frame['INGREDIENTS'].str.split('\]', expand = True)[0]

frame['PATHOLOGY'] = frame['INGREDIENTS'].str.split('\]', expand = True)[1]

The initial columns are then dropped to eliminate duplication.

frame = frame.drop(['DRUG'], axis=1)

frame = df.frame(['INGREDIENTS'], axis=1)

Step 13 : Putting each SYMPTOM in a separate ROW

The REACTION column contains several symptoms all on one line, which makes it difficult to count the symptom frequencies. So I created a new row for each symptom –

```
symptomlist = []
symptom = []
drugg =  []
ingredients = []
pathology = []
dates = []
idlist = []

Name = ""


for index, row in frame.iterrows():
    symptoms = str(row['REACTION']).split("<BR><BR>")
    for i in symptoms:
        symptomlist.append(i)
        drugg.append(row['NAME'])
        ingredients.append(row['INGREDIENT-LIST'])
        pathology.append(row['PATHOLOGY'])
        idlist.append(row['ID'])




dict = {'ID': idlist, 'PATHOLOGY': pathology, 'DRUG': drugg, 'INGREDIENTS' : ingredients,'REACTION': symptomlist}
frame = pd.DataFrame(dict)

frame.head()
```

Step 14 : Splitting the REACTION column into SYMPTOM and SEVERITY

Each symptom record in the REACTION column also contains information on the outcome. So I split the REACTION column into 2 columns – SYMPTOM and SEVERITY

frame['SYMPTOM'] = frame['REACTION'].str.split('\(', expand = True)[0]

frame['SEVERITY'] = frame['REACTION'].str.split('\(', expand = True)[1]

The initial REACTION column is then dropped to eliminate duplication

Frame = frame.drop(['REACTION'], axis=1)

Finally, the columns are tidied up by removing unnecessary brackets –

frame['PATHOLOGY'] = frame['PATHOLOGY'].str.replace('\(S -',"")

```
frame['SEVERITY'] = frame['SEVERITY'].str.replace('\),',"")
frame['SEVERITY'] = frame['SEVERITY'].str.replace('\)',"")
```

Before proceeding I saved the existing frame under a new name Eudra. Doing this intermittently is a way of saving preceding work in case proceding steps have issues. I also saved frame as a csv, so I would not have to redo all steps each time I opened the project but could simply load the current csv.

```
Eudra = frame.copy()
```

Step 15 : Removing Rows where DRUG name is null

The DRUG column is null for some records, even though the INGREDIENTS column shows the ingredients. Consequently, if we are looking at DRUG-SYMPTOM associations, then the rows with no drug name should be removed first.

```
Eudra.dropna(subset=['DRUG'], inplace=True)
```

Step 16 : Cleaning the DRUG Column

The DRUGS column contain the name of the drug together with additional punctuation and other info such as dosage. These variations must be combined if the same drug name occurs in each variation.

```
Eudra['DRUG'] = Eudra.apply(lambda x: "SPIKEVAX" if "SPIKEVAX" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "MODERNA" if "MODERNA" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "VAXZEVRIA" if "VAXZEVRIA" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "COMIRNATY" if "COMIRNATY" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "TOZINAMERAN" if "TOZINAMERAN" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "ELASOMERAN" if "ELASOMERAN" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "GARDASIL9" if "GARDASIL 9" in x['DRUG'] else x['DRUG'] , axis=1)
```

```
Eudra['DRUG'] = Eudra.apply(lambda x: "ASTRAZENECA" if "ASTRAZENECA" in x['DRUG'] else x['DRUG'] , axis=1)
```

Since almost all of the drug names are the first word in the DRUG column, then I could remove all the superfluous information by selecting the first word as the drug name.

I used a loop, rather than trying to do this with a single line of code such as –

```
Eudra['DRUG'] = Eudra['DRUG'].split(' - ')[0]
```

Instead I used this loop. The loop uses much less memory

```
from tqdm import tqdm
druglist = []
for index, row in tqdm(Eudra.iterrows()):
    druglist.append(row['DRUG'].split(' ')[0])

Eudra['DRUG'] = druglist

```

I also removed any commas

```
Eudra['DRUG'] = Eudra['DRUG'].str.replace(",","")
```

Step 17 : Tidying up SYMPTOMS Column

Here I am tidying up the SYMPTOM column by removing unwanted spaces and text, and converting all to lower case.

```
Eudra['SYMPTOM'] = Eudra['SYMPTOM'].str.replace("nannannan","")
Eudra['SYMPTOM'] = Eudra['SYMPTOM'].str.replace("nannan","")
Eudra['SYMPTOM'] = Eudra['SYMPTOM'].str.strip()
Eudra['SYMPTOM'] = Eudra['SYMPTOM'].apply(str.lower)

```

Step 18 : Tidying up PATHOLOGY Column

I also tidied up the PATHOLOGY column by removing unwanted text after the hyphen.

I also wanted to remove null values, but did not want to remove entire rows with null values, since this would delete drug-symptom associations in the other columns, so I used fillna() instead of dropna()

```
Eudra[['PATHOLOGY']] = Eudra[['PATHOLOGY']].fillna("Not listed")
from tqdm import tqdm
pathologylist = []
for index, row in tqdm(Eudra.iterrows()):
    pathologylist.append(row['PATHOLOGY'].split(' - ')[0])

Eudra['PATHOLOGY'] = pathologylist
```

I looped through the column because simply using split() on the entire column at once used too much memory.

Step 19 : Splitting the SEVERITY Column into DURATION, EFFICACY and SAFETY

I decided to split this column into 3 separate columns – DURATION, EFFICACY and SAFETY.

EFFICACY indicated whether the recipient of the drug had full or partial or no recovery.

SAFETY indicated whether the symptom was life-threatening, disabling, resulted in hospitalization, resulted in death, resulted in another medically important condition, or a congenital anomaly.

```
Eudra['SEVERITY'] = Eudra['SEVERITY'].str.replace("nan","")
from tqdm import tqdm
durationlist = []
efficacylist = []
safetylist = []
for index, row in tqdm(Eudra.iterrows()):
    try:
        durationlist.append(row['SEVERITY'].split(' - ')[0])
    except:
        durationlist.append('na')
    try:
        efficacylist.append(row['SEVERITY'].split(' - ')[1])
    except:
        efficacylist.append('na')
    try:
        safetylist.append(row['SEVERITY'].split(' - ')[2])
    except:
        safetylist.append('na')

Eudra['DURATION'] = durationlist
Eudra['EFFICACY'] = efficacylist
Eudra['SAFETY'] = safetylist
```

The original SEVERITY column was then dropped to eliminate duplication –

```
Eudra = Eudra.drop('SEVERITY', axis=1)
```

Statistical Restrictions

Some of these drugs had very few symptom records. Small samples lead to inaccurate statistic, so I decided to exclude all drugs with less than 100 symptom records in total.

> *A sample size of **30** is fairly common across statistics as the minimum for applying the central limit theorem.6 The higher your sample size, the more likely the sample will be representative of your population set.*
>
> https://www.investopedia.com/terms/c/central_limit_theorem.asp

100 was chosen because there is an average of 3.34 symptoms per person report, and I want to base calculations on a minimum of 30 reports, which equates to a minimum of 100 symptom records for each drug.

After applying this restriction, the number of drugs remaining in the dataset was 1504.

```
plus100 = Eudra['DRUG'].value_counts().reset_index().rename(columns={"index": "DRUG", "DRUG": "FREQ"})
druglist2 = []
drugcounter = 0
for index, row in plus100.iterrows():
    if row['FREQ'] >= 100:
            druglist2.append(row['DRUG'])
            drugcounter += 1
print(drugcounter)

Eudra = Eudra[Eudra['DRUG'].isin(druglist2)]
Eudra.nunique()
```

The Resulting Dataset

The resulting dataset is for DRUG-SYMPTOM associations.

It is called – drugversion.csv

The final dataset after removal of records where there was no drug name, has –

1. 4,486,467 individual people records
2. 14,987,664 drug-symptom associations –
3. 19,247 unique symptoms
4. 1,504 unique drug brand names, and
5. 3,163 unique ingredient lists.

Note : Each row is NOT a unique person. Rather it is a drug-symptom association. The ID column indicates the person. Each person takes a single drug. But that drug can have many symptoms and hence many rows. Each symptom has a different duration (DURATION), recovery rate (EFFICACY) and severity (SAFETY).

Files for Download

1. The 4088 original csv files are found in a master folder called - "eudra-vigilance3", with a size of 6.63 Gb uncompressed, and 690 Mb compressed.

Download (original 4088 files)

2. The 4088 files combined into a single file called – "total-08-06-2024" of size 4.78 Gb uncompressed, but compressed is 318 Mb.

Download (all columns - concatenated)

3. The dataset prepared and organized for analysis – "1504drugversion.csv"

Download (only columns for drug-reaction analysis – one symptom per row)

4. A fourth dataset is available. It is a pivot table showing the safety signals for any of 19,245 symptoms for 1504 different drugs. It also shows every drug sorted in order of its safety signal for any chosen symptom. This dataset has a size of only 4Mb.

Download

EXPLORING THE DATA – OTHER COLUMNS IN THE ORIGINAL DATASET

Who submitted these reports ?

```
1  f1['QUALIF'].value_counts()
```

Healthcare Professional	4492180
Non Healthcare Professional	2523665
Not Specified	28380

So most reporters were healthcare professionals

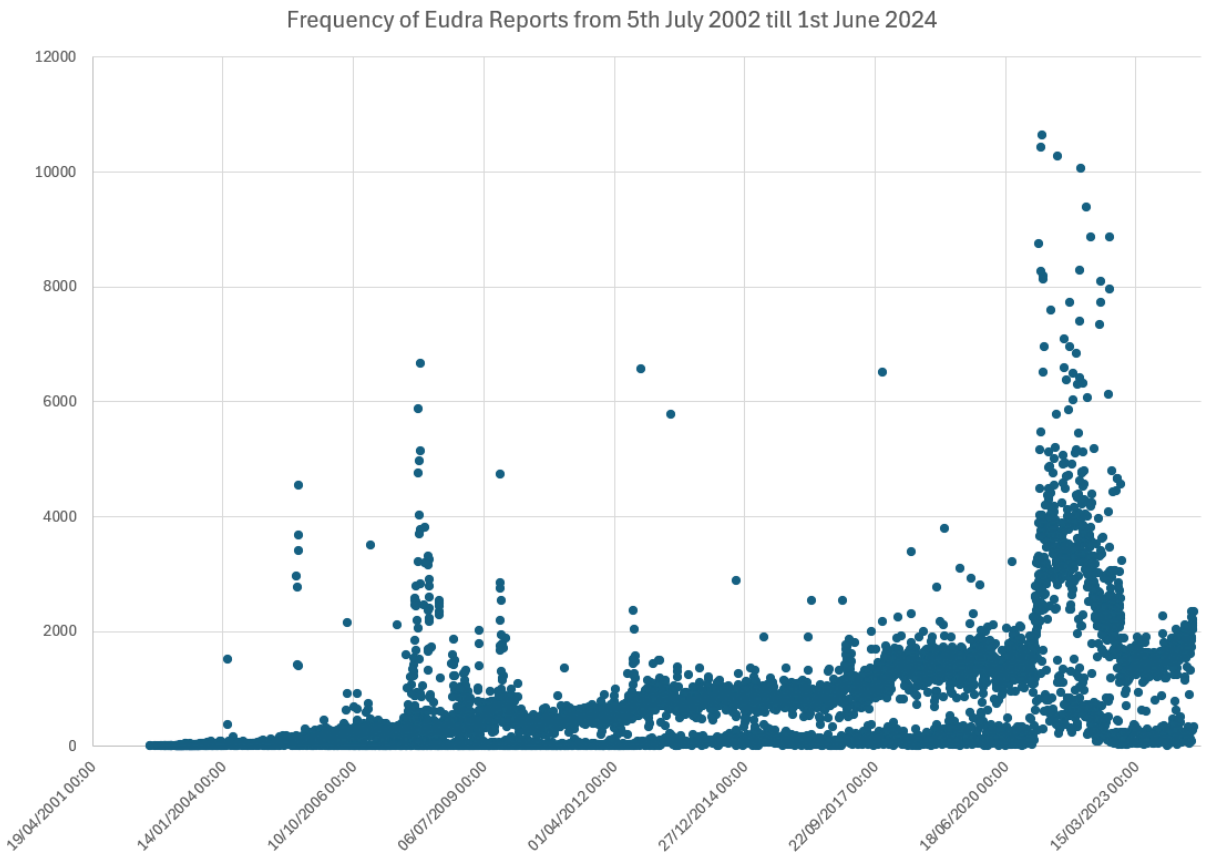
NOTE : 63% of reports in Eudra were submitted by healthcare professionals, and a further 35% of reports were submitted by non-healthcare professionals - most likely admin staff in the employ of hospitals, private medical practices, or care services. So this data is to be taken seriously.

The DATE Column

```
DateRange = frame['DATE'].value_counts().reset_index().rename(columns={"index": "DATE", "DATE": "FREQ"})
DateRange.to_csv(r"C:\Users\User\Documents\AAAcovid\dates.csv")
```

This is the date when the report was received. Reports cover a time span from 5th July 2002 till 1st June 2024, a period of 22 years.

The spread of dates is interesting –



As you can see, there is a gradual rise in number of reports from 2002 to 2024. There are also some sharp increases – in 2008 (Swine Flu Jab) and in 2021 (Covid Jab).

The incidence jumps up suddenly. It reached a peak on 27th March 2021 (Passover). It remains high for 1 whole year, then starts to drop in March 2022 By December 2022 it has returned to 2020 level. Both the abrupt beginning and abrupt end do not seem very virus-like.

DATE	FREQ	>1000
01/06/2024 00:00	329	
31/05/2024 00:00	2342	>1000
30/05/2024 00:00	2201	>1000
29/05/2024 00:00	2147	>1000
28/05/2024 00:00	2032	>1000
27/05/2024 00:00	2112	>1000
25/05/2024 00:00	272	
24/05/2024 00:00	2047	>1000
23/05/2024 00:00	2038	>1000
22/05/2024 00:00	2342	>1000
21/05/2024 00:00	2242	>1000
20/05/2024 00:00	1725	>1000
19/05/2024 00:00	79	
18/05/2024 00:00	252	
17/05/2024 00:00	2101	>1000
16/05/2024 00:00	1882	>1000
15/05/2024 00:00	1900	>1000
14/05/2024 00:00	1843	>1000
13/05/2024 00:00	1705	>1000
12/05/2024 00:00	124	
11/05/2024 00:00	319	
10/05/2024 00:00	1709	>1000
09/05/2024 00:00	1610	>1000
08/05/2024 00:00	1839	>1000
07/05/2024 00:00	2011	>1000
06/05/2024 00:00	1316	>1000
05/05/2024 00:00	38	
04/05/2024 00:00	84	
03/05/2024 00:00	1963	>1000
02/05/2024 00:00	1842	>1000
01/05/2024 00:00	892	
30/04/2024 00:00	2081	>1000
29/04/2024 00:00	1792	>1000
28/04/2024 00:00	97	

The graph appears to split into two streams, a lower one and an upper one. The lower one is made up of weekend dates when input was less, and the upper is made of weekday dates, when input was greater. Notice that weekend input also has a raised peak.

I have created a spreadsheet of the frequency of reports for each date from 2022 till 2024. You can download the spreadsheet here.

Notice that the high frequency of reports comes in clusters of 5 days. This is because data inputting mainly occurs during weekdays rather than at weekends.

This is further evidence that data is being input by professionals working in the health industry, rather than by ordinary members of the public, who would tend to input during out-of-work hours or at the weekend.

There is a surge in reports soon after the rollout of the Covid Jab. The surge is abrupt – rising to 2 x the 2020 rate of reports in a single month. and reaching a peak on March 27th. Then it remains high until March 2022. After that it begins to descend at a slower rate.

There is no surge evident during the 2020 "pandemic" period !! This suggests that the reports relate to adverse effects of the jab rather than adverse effects that preceded the jab.

Do viruses behave like this?

The magnitude of the explosion is about 2 times higher than the level immediately preceding it in 2020.

The COUNTRY Column

```
1 f1['COUNTRY'].value_counts()
```

European Economic Area	3681880
Non European Economic Area	3362277
Not Specified	68

The data is equally split between the European Economic Area, and the Non-European Economic Area.

It will be interesting to see how western Europe differs from eastern Europe.

The AGE Column

```
1 f1['AGE'].value_counts()
```

18-64 Years	3523899
Not Specified	1458975
65-85 Years	1440295
More than 85 Years	203196
12-17 Years	141390
2 Months - 2 Years	127431
3-11 Years	125836
0-1 Month	23203

These numbers of records are large, so an analysis of childhood effects is possible.

The REP-AGE Column

The REP-AGE column provides slightly different categories –

```
1 f1['REP-AGE'].value_counts()
```

Not Specified	4987662
Adult	1387925
Elderly	504569
Child	57388
Adolescent	48994
Infant	38191
Neonate (Preterm and Term newborns)	17264
Foetus	2232

It will be interesting to look at effects on foetus and neonates.

The PARENT-REPORT Column

```
1  f1['PARENT-REPORT'].value_counts()
```

```
No     6941192
Yes     103033
```

This is where parents submit a report on behalf of a child. There are 135586 such reports. Since these reports relate to children, this will be a valuable dataset for analysing effects on children.

The GENDER Column

```
1  f1['GENDER'].value_counts()
```

```
Female           4052779
Male             2607758
Not Specified     383688
```

Adverse effects are reported more frequently for females. Does this apply for all drugs, or only for the Covid jab?

The ratio of female to male is approximately 1 : 1.5, or 2 : 3. It will be interesting to see if this ratio varies for different drugs.

The DRUG Column

```
1  filter = f1[f1['DRUG'].str.contains('<BR>') == False]
2  filter['DRUG'].count()
```

```
5333575
```

This column lists the drugs suspected of causing the adverse reactions. In most cases only a single drug is listed.

- Number of reports with a single drug listed = 5,333,575 = 75%
- Number of reports with multiple drugs listed = 1,710,650 = 25%

When multiple drugs are administered, it is harder to attribute adverse effects to a particular drug. For this reason, I will analyse effects from single drug administration first.

Cases of multiple drug administration are mostly multiple doses of the Covid jab. It will be interesting to analyse this to see if there are dose dependent effects.

The CONCOM Column

This is similar to the DRUGS column, except it is for concomitant drugs taken, rather than for the suspected drug. These are drugs that are administered concomitantly with the suspected drug. Each drug entry also lists its PATHOLOGY TREATED, INGREDIENTS, DURATION and DOSE AMOUNT. These will have to be split out into separate columns, just as is needed for the DRUG column.

Concomitant drugs may suggest -

1. co-morbidities
2. additional drugs used to overcome the adverse effects of the suspected medicine.
3. Additional drugs used to compliment the suspected medicine

As with the DRUG column, reports with no concomitant drugs make for a clearer attribution of effects to the suspected drug. Such reports can be identified by "Not reported" or "NOT AVAILABLE".

```
FilterCONCOM = frame[frame['CONCOM'].str.contains('Not reported')]
```

The most common concomitant drugs are "anti-retro-virals". During the Covid "pandemic" anti-retro-virals were heavily prescribed as an anti-viral treatment for early onset of "Covid". This included -

- REMDESIVIR
- PAXLOVID (Ritonavir)
- TAMIFLU (Oseltamivir)

It is highly controversial why these drugs were used, since they have side-effects that are very serious including kidney damage, liver damage and bone damage. It is unclear why Tenofovir was being administered…

CONCOM	FREQ
[PARACETAMOL] (C - n/a - n/a - [n/a - n/a - n/a])	4486
[INFLUENZA VIRUS] (C - n/a - n/a - [n/a - n/a - n/a])	3391
[LEVOTHYROXINE SODIUM] (C - n/a - n/a - [n/a - n/a - n/a])	3307
[LEVOTHYROXINE, LEVOTHYROXINE SODIUM] (C - n/a - n/a - [n/a - n/a - n/a])	2886
BIKTARVY [EMTRICITABINE, TENOFOVIR ALAFENAMIDE, BICTEGRAVIR] (C - n/a - n/a -)	2773
DESCOVY [EMTRICITABINE, TENOFOVIR ALAFENAMIDE] (C - n/a - n/a -)	2331
GENVOYA [EMTRICITABINE, TENOFOVIR ALAFENAMIDE FUMARATE, COBICISTAT, ELVITEGRAVIR, EMTRICITABINE, TENOF	2219
COMIRNATY [TOZINAMERAN] (C - Immunisation - n/a - [n/a - 1{DF} - n/a])	1884
[LEVOTHYROXINE, LEVOTHYROXINE SODIUM] (C - Hypothyroidism - n/a - [n/a - n/a - n/a])	1845
[ETHINYLESTRADIOL, LEVONORGESTREL] (C - n/a - n/a - [n/a - n/a - n/a])	1843
[LETROZOLE] (C - n/a - n/a - [n/a - n/a - n/a])	1806
[ACETYLSALICYLIC ACID] (C - n/a - n/a - [n/a - n/a - n/a])	1803
[PREDNISONE] (C - n/a - n/a - [n/a - n/a - n/a])	1382
[DESOGESTREL] (C - Contraception - n/a - [n/a - n/a - n/a])	1167
BIKTARVY [EMTRICITABINE, TENOFOVIR ALAFENAMIDE, BICTEGRAVIR] (C - n/a - n/a -), GENVOYA [EMTRICITABI	1105
[WARFARIN SODIUM] (C - n/a - n/a - [n/a - n/a - n/a])	1073
IBUPROFEN [IBUPROFEN, IBUPROFEN SODIUM] (C - n/a - n/a - [n/a - n/a - n/a])	1015
[SERTRALINE, SERTRALINE HYDROCHLORIDE] (C - n/a - n/a - [n/a - n/a - n/a])	1014
[ETHINYLESTRADIOL, LEVONORGESTREL] (C - Contraception - n/a - [n/a - n/a - n/a])	990
COMIRNATY [TOZINAMERAN] (C - COVID-19 immunisation - Not applicable - [1d - 1{DF} - n/a])	979
BIKTARVY [EMTRICITABINE, TENOFOVIR ALAFENAMIDE, BICTEGRAVIR] (C - n/a - n/a -), DESCOVY [EMTRICITABI	979
[INSULIN, INSULIN HUMAN] (C - n/a - n/a - [n/a - n/a - n/a])	978
[OMEPRAZOLE] (C - n/a - n/a - [n/a - n/a - n/a])	966

The ICSR Column

ICSR means "Individual Case Safety Report" – showing that each row is an individual person. There are 7,044,225 unique case reports in the dataset.

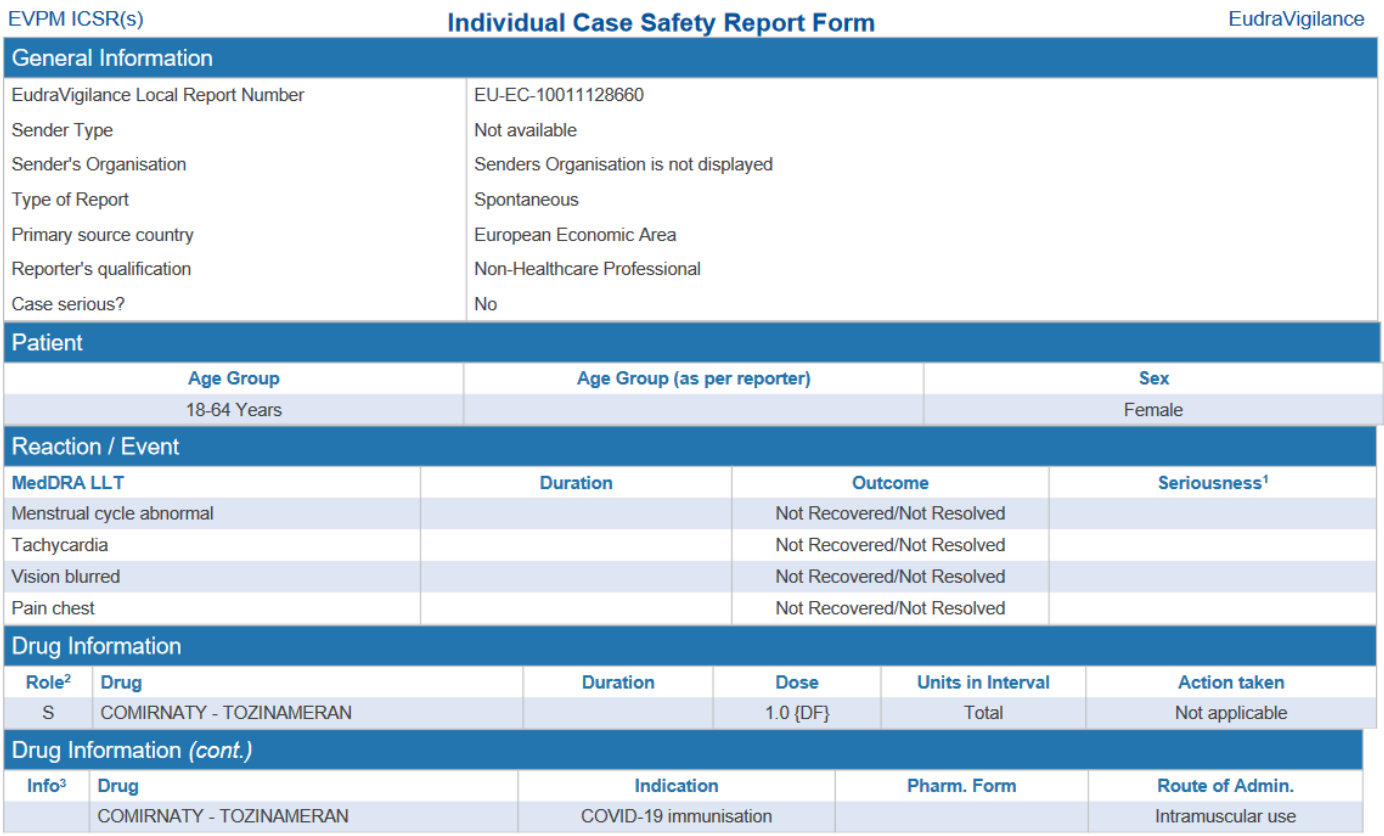
The ICSR column lists the URL for each report. For example –

https://dap.ema.europa.eu/xmlpserver/PHV%20DAP/Reports/ICSR.xdo?_xpf=&_xt=form&SR_ID=10011128660&_xpt=1&_xf=pdf

This is the URL for the report with ID 10011128660. It is a pdf hosted on the European Medical Association website.

This shows that the downloaded data is not made up. Rather, each row of data has a report that you can verify using the link provided.

Each downloaded report looks like this –



EVPM ICSR(s) **Individual Case Safety Report Form** EudraVigilance

General Information

EudraVigilance Local Report Number	EU-EC-10011128660
Sender Type	Not available
Sender's Organisation	Senders Organisation is not displayed
Type of Report	Spontaneous
Primary source country	European Economic Area
Reporter's qualification	Non-Healthcare Professional
Case serious?	No

Patient

Age Group	Age Group (as per reporter)	Sex
18-64 Years		Female

Reaction / Event

MedDRA LLT	Duration	Outcome	Seriousness[1]
Menstrual cycle abnormal		Not Recovered/Not Resolved	
Tachycardia		Not Recovered/Not Resolved	
Vision blurred		Not Recovered/Not Resolved	
Pain chest		Not Recovered/Not Resolved	

Drug Information

Role[2]	Drug	Duration	Dose	Units in Interval	Action taken
S	COMIRNATY - TOZINAMERAN		1.0 {DF}	Total	Not applicable

Drug Information *(cont.)*

Info[3]	Drug	Indication	Pharm. Form	Route of Admin.
	COMIRNATY - TOZINAMERAN	COVID-19 immunisation		Intramuscular use

Comparing Cancer Drugs - an example of drug analysis

I used the Version 2 dataset to determine the safety and efficacy of 152 cancer drugs.

Safety was measured by the number of associations with

- A fatal outcome
- A hospital outcome
- A life threatening event

each expressed as a percentage of the total number of associations

Efficacy was measured by the number of associations with

- An outcome of complete recovery
- An outcome of improvement
- An outcome of no recovery

each expressed as a percentage of the total number of associations

In addition I looked at the incidence of symptoms related to well-being during treatment –

- Pain
- Nausea
- Fatigue
- Diarrhea
- Hairloss
- Anaemia
- Neutropenia
- Cardiomyopathy
- Neuropathy

A high association did not mean that the drug caused the symptom – it could also arise if the drug was used to treat a pathology with the symptom. A high incidence simply identified POSSIBLE cause, and would require further research of clinical and experimental literature to determine if it was an ACTUAL cause.

The write-up for this study can be found here –

cancer-drugs-compared

And the downloadable spreadsheet can be found here –

cancer-drug-ratings-updated.csv