Journal of Applied Pharmaceutical Science 2026. Article in Press
Available online at http://www.japsonline.com
DOI: 10.7324/JAPS.2026.245748
ISSN 2231-3354

Development and evaluation of Python language processed automated disproportionality analysis system for FAERS database

Avinash Laddha[1,6#], Deepak Gurram[1#], Rajalakshmi Rajendran[1], Muhammed Rashid[1,2], Pooja Gopal Poojari[1,3], Sreedharan Nair[1], Sohil Khan[1,4], Krishnan Subramonian[5], Madamanchi Chandra Vardhan[5], Richa Jackeray[6], Girish Thunga[1,7*]

[1]Department of Pharmacy Practice, Manipal College of Pharmaceutical Sciences, Manipal Academy of Higher Education, Manipal, India.
[2]Department of Pharmacotherapy, College of Pharmacy, University of Utah, UT, USA.
[3]Department of Pharmacy Practice, Srinivas College of Pharmacy, Valachil, Mangalore, India.
[4]Pharmacotherapeutics and Evidence Based Practice, School of Pharmacy and Medical Sciences, Griffith University, Gold Coast, Queensland, Australia.
[5]Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India.
[6]Clinical Reporting, Novo Nordisk Global Business Services, Bengaluru, India.
[7]Centre for Toxicovigilance and Drug Safety, Manipal College of Pharmaceutical Sciences, Manipal Academy of Higher Education, Manipal- 576104.

ARTICLE HISTORY

Received on: 04/03/2025
Accepted on: 15/10/2025
Available Online: XX

Key words:
Adverse event, disproportionality analysis, FAERS, risk assessment.

ABSTRACT

The limitations of existing disproportionality analysis methods for Food and Drug Administration Adverse Event Reporting System (FAERS) data highlight the need for automated tools for efficient data mining and analysis. This study aimed to develop and validate an automated Python-based tool for FAERS data processing. The methodology included: (i) Automation development and signal detection using Python and (ii) Validation through traditional disproportionality analysis. Public FAERS quarterly extract files from 2004 (Q1) to 2021 (Q4) were accessed, matched, and deduplicated using a multi-step approach. The cleaned dataset was analysed using a contingency table to compute reporting odds ratios, PRR, Chi-square statistics, and 95% confidence intervals. Validation confirmed that results from the automated tool matched traditional analysis. Using Remdesivir as an example, we identified 12,777 adverse events and 256 safety signals. The tool offers multiple advantages: simplified coding, minimal storage requirements, cloud-based execution (Google Colab), accessibility for non-technical users, rapid processing, tailored signal detection, and automated FAERS updates. The extreme deduplication process ensures refined results, aligning with pre-defined criteria. This validated tool can significantly enhance pharmacovigilance research and regulatory decision-making, with potential for broader adoption through user-friendly desktop interfaces.

1. INTRODUCTION

An adverse drug reaction (ADR) is defined as any unwanted or harmful effect of a drug or medicinal product administered at its therapeutic dose for any indicated use. The World Health Organization defines Pharmacovigilance (PV) as the science and discipline concerned with detecting, understanding, assessing, and preventing ADRs or any other drug-related problems [1]. In addition to drugs, PV now encompasses herbals, traditional and complementary medicines, blood products, biologicals, medical devices, and vaccines [2]. Post-marketing surveillance (PMS) and spontaneous ADR reporting are crucial for identifying potential risks and rare ADRs that may not have been detected during clinical trials [3,4]. The Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is one of the largest spontaneous ADR reporting databases globally, managed by the FDA [5]. FAERS contains adverse event reports, medication error reports, and product quality complaints related to drugs and therapeutic biologics. These reports can be submitted voluntarily

*Corresponding Author
Girish Thunga, Department of Pharmacy Practice,
Manipal College of Pharmaceutical Sciences,
Manipal Academy of Higher Education, Manipal, India, 576104
Phone number: +91-9880151127
E-mail address: girish.thunga@manipal.edu

#Authors contributed equally

or mandatorily by industries, hospitals, or the public [6]. A systematic assessment of ADRs using statistical tools, known as data mining, can reveal unusual or unexpected product-event combinations [7]. The Council for International Organizations of Medical Sciences working group VIII recommends data mining for signal detection. These generated signals can then trigger further clinical investigations to gather additional safety evidence for medicinal products [7,8]. Various tools have been developed to assist in data mining, de-duplication, and data curation. These tools align with standard terminologies such as SNOMED-CT, RxNorm, or MedDRA [5,8,9]. Among them, Banda *et al*. developed a data curator to clean FAERS data, but this tool treats multi-ingredient drugs as single entities, limiting its accuracy [8]. Another attempt by Khaleel *et al*. [5] involved a pre-calculated disproportionality analysis using standardized drug names as per the RxNorm vocabulary, normalized to a single active ingredient level. However, the program included all suspect drugs—primary, secondary, and concomitant—potentially exaggerating the reporting odds ratio (ROR). The authors acknowledged these limitations in the dataset's de-duplication process [5]. To the best of the authors' knowledge, the dataset developed by Khaleel *et al*. [5] remains the latest publicly available clean version of the FAERS database.

Given the limitations of existing tools and the time-consuming nature of traditional methods, we aim to develop an automated tool using Python to perform disproportionality analysis. This tool streamlines the entire process, from initial data extraction and conversion into a machine-readable format to case matching, file merging, de-duplication, and statistical analysis to assess actual causal relationships between drugs and specific ADRs. The tool has a simple coding structure and a user-friendly interface, making it accessible to non-technical personnel. It efficiently stores data and operates on cloud-based platforms such as Google Colab, reducing processing time and resource consumption. Additionally, the tool allows investigators to customize ADR analyses according to their requirements, ensuring refined results. One of its key advantages is the automatic updating of the data system, which keeps the information current. Furthermore, the tool's performance is validated by comparing its output with traditional disproportionality analysis conducted manually, ensuring accuracy and reliability throughout the de-duplication process.

2. METHODOLOGY

This study follows a two-phase methodology: (i) the development of an automated tool using Python for signal detection and disproportionality analysis and (ii) validation through traditional disproportionality analysis. The tool was developed and validated between January 2022 and March 2023. It has been filed for a patent under application number 202441050326, dated July 1, 2024.

2.1. Data sources

The publicly available international FAERS, now referred to as the FAERS database (https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html), was used for data collection. This spontaneous reporting system serves as a source for safety signal generation, aiding in the tracking and monitoring of new or emerging adverse events and medication errors [5].

2.2. Data extraction

We accessed the publicly available quarterly extract files from FAERS covering the period from 2004 (Q1; January–March) to 2021 (Q4; October–December). Each extract includes reports received by FAERS within a specific quarter of the year. In total, we downloaded 61 quarterly extract files, which are available in ASCII or XML format. Each quarterly extract consists of seven data files: DEMOyyQq (demographic information), DRUGyyQq (drug details), INDIyyQq (indication for use), OUTCyyQq (patient outcome), REACyyQq (adverse event/reaction), RPSRyyQq (reporter source), and THERyyQq (therapy details, including start and end dates).

2.3. Data retrieval and merging

The data files were retrieved starting from the year of the drug's first marketing approval, which was determined using the official FDA website, Drugs@FDA (https://www.accessdata.fda.gov/scripts/cder/daf/). The initial step in data retrieval involved searching for the specific drug of interest. Each FAERS report is uniquely identified by a primary ID, which serves as a link between all seven data files. The primary ID is a combination of the Case ID and Case Version assigned to each case. The Case ID is a unique number that identifies a FAERS case, while the Case Version represents the safety version number associated with that case.

2.4. Matching and de-duplication

2.4.1. Matching with primary IDs

There can be multiple reports for a single case, sharing the same Case ID but having different Primary IDs. The first version of the case represents the initial report, while all subsequent follow-up reports are assigned an increasing version number (e.g., Version 1, Version 2, Version 3, and so on). The latest version of a case contains the most updated information, whereas all previous versions are considered duplicates. To eliminate such duplicates, we selected reports with the highest Primary ID and Case Version, as recommended by the FDA [10].

In some instances, reporters may fail to correctly link follow-up versions to their corresponding initial reports. This results in duplicate reports with different Primary IDs and Case IDs but containing identical information. To address this, we implemented a second deduplication step. Cases were identified as duplicates and removed if they matched across the following data fields: PROD_AI (active ingredient), AGE, SEX, EVENT_DT (event date), REPORTER_COUNTRY, and PT [preferred term (PT) for the adverse event]. Only records with complete information in these fields were considered for deduplication. Cases with missing values in any of these columns were excluded from this process to prevent the accidental removal of unique cases where two or more field values were unavailable [8].

2.4.2. Matching FAERS and LAERS

Legacy adverse event reporting system (LAERS) was the predecessor to FAERS, operating between 2004Q1 and 2012Q3. Unlike FAERS, LAERS used ISR, Case, and Foll_seq in place of Primary ID, Case ID, and Case Version,

respectively. FAERS replaced LAERS in 2012Q4. Notably, ISR was not a concatenated key of Case and Foll_seq; instead, it was a unique identifier for an AERS report and served as the primary link field between files. To ensure consistency in data handling, we referred to the LAERS versus FAERS ASCII Tag Comparison Table provided in *ASC_NTS.DOC* within the quarterly extract files. This document outlines differences in data elements between the two systems (e.g., ISR in LAERS was renamed as PrimaryID in FAERS). We retrieved data from corresponding elements in AERS files and aligned them with their respective data elements in FAERS, as reported previously [8]. The detailed LAERS vs. FAERS ASCII tag comparisons are presented in Table 1 below.

2.4.3. Extreme de-duplication

To ensure consistency in data retrieval between LAERS and FAERS, we selected data elements based on FAERS standards. The output file contains FAERS data elements, with corresponding LAERS data mapped accordingly. An active ingredient search was performed in the *prod_ai* column within the DRUG file, and only drugs classified as primary suspects in the *ROLE_CODE* column were included. Reports categorized as secondary suspects, concomitant, or interacting were excluded.

Once reports with the desired active ingredient and primary suspect classification were retrieved, they were matched with corresponding data from the indication, reaction, outcome, and therapy files using the *Primary ID*. The report source file was excluded to avoid potential duplication, as a single report may be submitted by multiple reporters (e.g., consumer, healthcare professional, and manufacturer) [11].

For correlation between drug, indication, and therapy files, sequence numbers were matched using *drug_seq* (DRUG file), *indi_seq* (INDICATION file), and *dsg_drug_seq* (THERAPY file). A single case may contain multiple drugs, indications, and therapy dates.

For example, a case with *Case ID 17896499* and *Primary ID 178964993* contained seven reported drugs, including one primary suspect drug, two secondary suspects, and four concomitant drugs. The details of sequence numbers and role codes for multiple drugs in a single case are provided in Table 2.

Similarly, each case included indications and therapy dates, each assigned a unique sequence number. Cases with missing indication and therapy information were also included in the study. If multiple outcomes were reported for a single case, only one outcome was considered, based on the most severe or serious event, to prevent duplication from selecting multiple outcomes per case. The severity ranking follows the Guidance on E2B(M) Data Elements for Transmission of Individual Case Safety Reports provided by the USFDA [12]. The prioritized list of outcomes, ranked by severity, is presented in Table 3. For instance, if a case included reported outcomes such as death, life-threatening event, or hospitalization, only the most severe outcome (death) was considered the final outcome.

This approach was adopted to ensure the accuracy, consistency, and clinical relevance of the data used in our analysis. In spontaneous reporting systems such as FAERS and LAERS, a single adverse event case may involve multiple drugs, indications, outcomes, and reporters. Without careful filtering, such complexity can lead to overrepresentation or duplication of cases, inflating associations, and introducing bias. By including only primary suspect drugs, we focused on those most likely responsible for the adverse event, thereby minimizing confounding from co-administered medications. The use of sequence numbers across files allowed precise matching of drug–indication–therapy relationships within a case, ensuring contextual integrity of the clinical scenario. Furthermore, selecting only the most severe outcome per case helped avoid distortion in outcome frequency, which could occur if multiple outcomes from a single case were counted

Table 1. LAERS versus FAERS tag comparisons.

LAERS	FAERS
Isr	Primary ID
Case	Case ID
Foll_seq	Case version
Gndr_cod	Sex
Drug name	Prod_ai
Drug_seq (INDI FILE)	Indi_drug_seq
Drug_seq (THERAPY FILE)	Dsg_drug_seq
Outc_code	Outc_cod

Table 2. Sequence numbers and role codes of multiple drugs reported to a single case.

PRIMARYID	DRUG_SEQ	ROLE_CODE	DRUGNAME
178964993	1	PS	Remdesivir
178964993	2	SS	Remdesivir
178964993	3	SS	Vancomycin
178964993	4	C	Cobicistat\Elvitegravir\Emtricitabine\ Tenofovir Alafenamide Fumarate
178964993	5	C	Bictegravir Sodium\Emtricitabine\ Tenofovir Alafenamide Fumarate
178964993	6	C	Enoxaparin
178964993	7	C	Cefepime Hydrochloride

C = concomitant; PS = primary suspect; SS = secondary suspect.

Table 3. Priority list of outcomes selected for an individual case as per severity.

Priority	Outc_cod
DE (results in death)	7
LT (life-threatening)	6
HO (requires inpatient hospitalization or prolongation of existing hospitalization)	5
DS (results in persistent or significant disability or incapacity, as per reporter's opinion)	4
CA (is a congenital anomaly or birth defect)	3
RI (require intervention)	2
OT (other medically important condition)	1

separately. This prioritization reflects real-world clinical relevance and aligns with regulatory guidance on PV reporting, ultimately strengthening the robustness and interpretability of our results.

2.5. Output file generation

After merging all the files, the final output file was generated, incorporating data elements from all sources. These include the primary ID, case ID, case version, follow-up code (if_code), age, age code (age_cod), sex, sex code (sex_cod), weight (wt), weight code (wt_cod), rechallenge code (rechal), dechallenge code (dechal), event date (event_dt), manufacturer report date (mfr_dt), FDA received date (fda_dt), reporter country, role code (role_cod), active ingredient (prod_ai), drug name (drugname), drug sequence number (drug_seq), route of administration (route), dose as reported (dose_vbm), PT for adverse event (pt), indication sequence number (indi_drug_seq), indication (indi_pt), outcome code (outc_cod), therapy sequence number (dsg_drug_seq), therapy start date (strt_dt), and therapy end date (end_dt). This comprehensive dataset integrates demographic, drug, indication, adverse event, outcome, and therapy details for each case, ensuring a structured and uniform format for further analysis.

2.6. Data mining

Careful assessment and systematic analysis of adverse events reported in PV databases with large datasets using statistical tools is referred to as data mining. This approach identifies the reporting proportion of a specific adverse event (e.g., bacterial infection) by comparing it to the reporting proportion of all other adverse events. The resulting disproportionality between the selected adverse event and all other adverse events in the database can provide evidence suggestive of a causal relationship [13].

2.7. Statistical analyses

The data for the searched active ingredients is generated as a CSV output file. These data are then utilized for statistical analysis using a contingency table to calculate the ROR, Proportional Reporting Ratio (PRR), Chi-square statistics, and the 95% confidence interval (CI) (lower and upper limits). Briefly, for each drug–event pair in the FAERS dataset, a 2 × 2 contingency table was constructed to classify cases into four groups: reports with both the drug and the event (a), drug without event (b), event without drug (c), and neither drug nor event (d). Based on these counts, the ROR and PRR were calculated, along with 95% CIs for ROR using standard error estimates. The statistical tools currently used in data mining by the FDA include the Multi-Item Gamma Poisson Shrinker (MGPS), PRR, and ROR [13].

2.8. Signal detection

2.8.1. ROR

The ROR is a data mining statistical tool used to quantify the disproportionality of an adverse event. It represents the odds of a specific adverse event being reported for a particular drug compared to all other adverse events for the same drug, relative to the occurrence of the same adverse event with all other drugs in the database. A higher proportion of reporting for a specific adverse event results in a higher ROR.

An adverse event is considered a signal when the number of cases exceeds 3, the ROR is greater than 1, and the lower limit of its 95% CI is greater than 1 [10,14]. The 95% CI provides point estimates of the ROR and defines its precision. The ROR is calculated using the following equation [15]:

The ROR [14] and 95%CI [15] were calculated using the following equation:

$$ROR = \frac{A/B}{C/D}$$

$$SE(\ln ROR) = \sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$$

$$95\%CI = e^{\ln(ROR)\pm 1.96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}}$$

- 95% CI Lower bound - eLn(ROR)-1.96√1/a+1/b+1/c+1/d.
- 95% CI upper bound - eLn(ROR)+1.96√1/a+1/b+1/c+1/d.

2.8.2. PRR

The PRR is calculated by comparing the proportion of a specific adverse event reported for a particular drug to the proportion of the same adverse event reported for all other drugs in the database. PRR serves as a key tool in signal detection and helps measure the potential association between a drug and an adverse event. A higher PRR indicates a stronger signal.

A signal is considered present when the PRR is at least 2, the number of cases is at least 3, and the Chi-square statistic is at least 4, as proposed by Evans *et al*. [4]. PRR 95% CI of PRR was calculated using the following equation [4]:

$$PRR = \frac{A/(A+B)}{C/(C+D)}$$

- 95% CI upper bound - eLn(PRR)-1.96√1/a+1/b+1/c+1/d.
- 95% CI lower bound - eLn(PRR)+1.96√1/a+1/b+1/c+1/d.

Chi-square statistic was calculated using the following equation:

Chi-square statistic = X2 = (AD − BC)2 (A + B + C + D)/ [(A + B)(C + D)(A + C)(B + D).

The workflow diagram of data retrieval, data merging, and statistical analyses is provided in Figure 1.

3. RESULTS

3.1. Development of automation and validation for disproportionality analysis from FAERS database using python language

Since the manual process of traditional disproportionality analysis using the FAERS database is time-consuming and requires significant manpower, we incorporated Python to automate the entire process. The automation was executed using Google Colab, following the same methodological approach as traditional

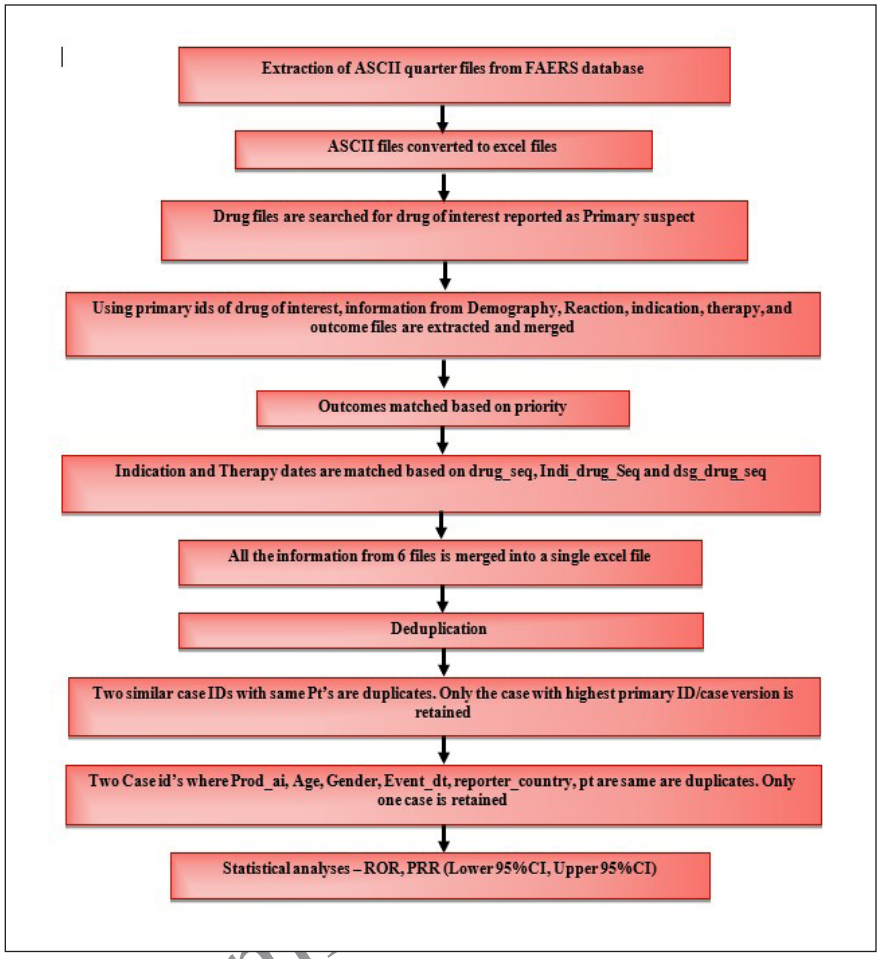


Figure 1. The workflow diagram of data retrieval, data merging, and statistical analyses.

disproportionality analysis. To validate this approach, we selected "remdesivir" as a test case and processed FAERS data through the automated system, comparing the results with traditional manual disproportionality analysis. The dataset for remdesivir, spanning from its inception to December 2021, was used for data extraction. The detailed procedure is outlined below.

3.1.1. Development of automation using python language

These ASCII files can be processed using databases such as Oracle, MS Access, SQL, IBM DB2, or SAS. However, to overcome time constraints, we developed a Python script to automate the process. Using an algorithmic approach, we retrieved FAERS data and performed statistical analyses efficiently. The developed algorithm downloads FAERS data and converts ASCII files into Excel format. It extracts information from multiple data categories, including demographics, drug details, indications, patient outcomes, reactions, report sources, and therapy details, consolidating them into a single Excel file. The data retrieval process follows a structured series of steps, including active ingredient search, primary suspect search, data merging, case deduplication, and statistical analyses.

3.1.2. Data retrieval

The developed algorithm can be executed using Google Colab or a Python application. It is integrated with a Megadrive that stores all quarterly extract files downloaded from the FAERS website. A total of 61 ASCII files are stored in this drive, serving as the data source for merging and statistical analyses. The primary objective of the algorithm is to merge all relevant columns from the FAERS files using primaryID as the unique key. This ensures accurate data linkage across different datasets. Figure 2 illustrates the ASCII entity relationship diagram, depicting the columns present in the quarterly extract files.

3.1.3. Data extraction, merging, de-duplication, and output generation

The algorithm consists of multiple cells, each performing a specific function in the process of data merging and statistical analysis. The *changes cell* (Fig. 3A) allows users to input the drug to be extracted and specifies the removal of *role_code* values other than *primary suspect* cases. To include secondary suspect, concomitant, or interacting cases, the corresponding *role_codes* can be deleted in this cell. Figure 3B

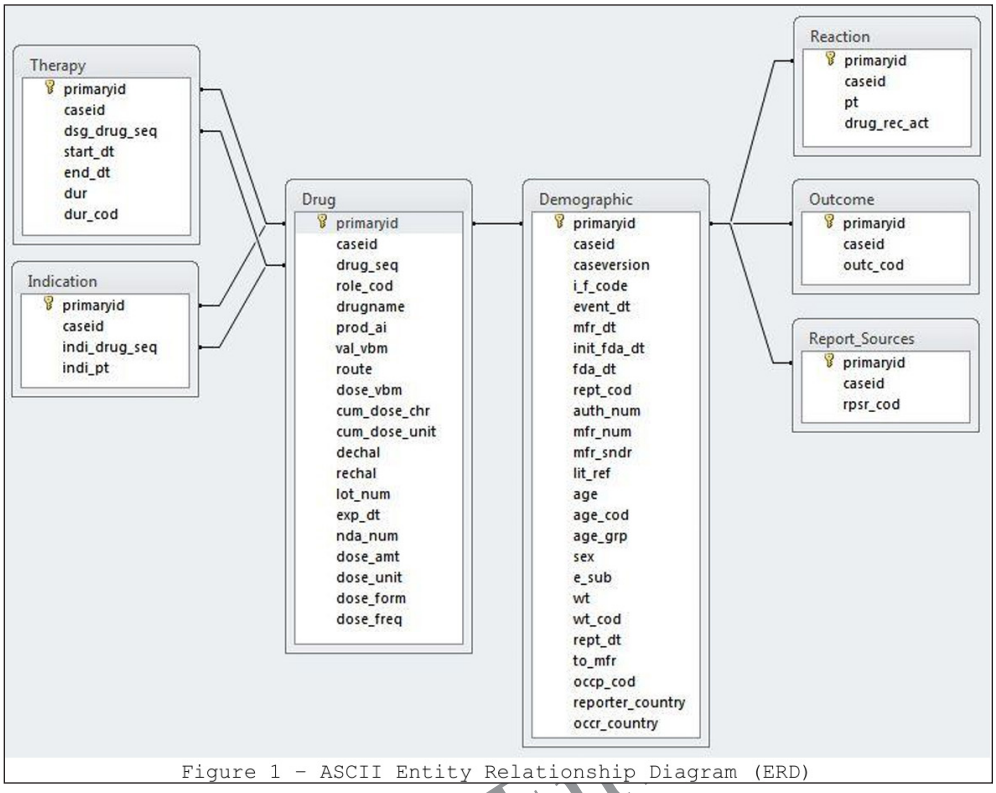


Figure 2. ASCII entity relationship diagram**.** The algorithm contains different cells that perform different functions in the process of data merging and statistical analyses. Figure 2 shows the changes cell where the input of the drug extracted is given. It also shows the role_code removed, other than the primary suspect cases. To include the secondary suspect, concomitant, and interacting cases role_codes can be deleted in the changes cell.

displays the merged columns between different files and the *rename directory*, ensuring uniformity of data elements between LAERS and FAERS. Figure 3C illustrates the prioritization of case severity based on outcome classification. The merging of drug, indication, and therapy files using sequence matching is performed in Figure 3D. The deduplication process, which removes duplicate cases, is handled in Figure 3E. Finally, the processed data can be downloaded as *FinalDrugsData.zip* under the *files* section, as shown in Figure 3F.

3.1.4. Statistical analysis

A separate set of Python codes was developed for statistical analysis and the calculation of causality measures such as ROR and PRR, following the traditional disproportionality analysis approach. Initially, all the codes and functions need to be executed as before, except those related to the *CHANGES* cell and statistical analysis (Fig. 4A). Next, necessary modifications should be made in the *CHANGES* cell, temporary directories should be created, and drug names should be extracted (Fig. 4B). This is followed by retrieving A–B data, creating local directories, moving A–B data to the drive, and extracting the PTs for each drug (Fig. 4C). Similar steps are then applied to C–D data, and the A–D datasets are retrieved from the drive (Fig. 4D). Finally, all required values are calculated based on A–D datasets (Fig. 4E). The results can be converted into a local format, and the entire process can be stored and downloaded as a ZIP file for offline access.

This automation can be modified to any drug of interest and event with simple modifications in the code. This includes the modification in the matching and de-duplication criteria.

3.2. Validation through traditional disproportionality analysis

To validate the output obtained from the algorithm, manual validation has been conducted through traditional disproportionality analysis.

3.2.1. Data retrieval, matching, and de-duplication

The validation process ensured that the data retrieved from FAERS using Python automation matched the results obtained through manual processing. In the first step, all primary IDs of Remdesivir that were identified as primary suspects were extracted from the drug file. These IDs were then compared with the algorithm's output, which initially identified 5,900 unique primary IDs through manual extraction. However, 207 primary IDs were removed during the deduplication process. A manual check confirmed that all 207 cases met the deduplication criteria.

In the second step, the extracted primary IDs were matched with the corresponding PTs from the reaction file. This resulted in a total of 13,200 PTs. The algorithm deleted 423 PTs during deduplication, and manual verification confirmed that these deletions adhered to the deduplication criteria. The deletions primarily resulted from retaining only the highest case version when case IDs were duplicated or from cases meeting

```
DRUG_FOLDER = "drugs"
TMP_ZIP_FILE = "tmp/zip/data.zip"
DATA_FOLDER = "tmp/data"
################################################################################
# REQUIRED_DRUGS = {  "sitagliptin": ["janumet", "janumet xr", "januvia", "juvisync"],
#                     "saxagliptin": ["onglyza", "kombiglyze xr"],
#                     "linagliptin": ["jentadueto", "tradjenta"],
#                     "vildagliptin": ["galvus", "galvus met"],
#                     "alogliptin": ["kazano", "nesina", "oseni"],
#                     "dapagliflozin": [],
#                     "canagliflozin": ["invokana"],
#                     "empagliflozin": [],
#                     "ertugliflozin": []
#                   }

REQUIRED_DRUGS = {
                    "remdesivir": []
                  }


################################################################################
REMOVE_ROLE_CODE = ["SS", "C", "I"]
################################################################################
REMOVE_COLUMNS = ["val_vbm", "cum_dose", "cum_dose_unit",
            "lot_num", "exp_dt", "nda_num", "dose_form", "caseid_y",
            "auth_num", "mfr_num", "mfr_sndr", "lit_ref", "e_sub", "rept_dt",
            "to_mfr", "occr_country", "caseid_y", "caseid_x", "rpsr_cod",
            "cum_dose_chr",
            "init_fda_dt", "dose_amt", "dose_unit", "dose_freq", "drug_rec_act"]
################################################################################
```

(A)

```
################################################################################
REMOVE_ROLE_CODE = ["SS", "C", "I"]
################################################################################
REMOVE_COLUMNS = ["val_vbm", "cum_dose", "cum_dose_unit",
            "lot_num", "exp_dt", "nda_num", "dose_form", "caseid_y",
            "auth_num", "mfr_sndr", "lit_ref", "e_sub", "rept_dt",
            "to_mfr", "occr_country", "caseid_y", "caseid_x", "rpsr_cod",
            "cum_dose_chr",
            "init_fda_dt", "dose_amt", "dose_unit", "dose_freq", "drug_rec_act"]
################################################################################
RENAME_DICT = {
                    "isr": "primaryid",
                    "case": "caseid",
                    "foll_seq": "caseversion",
                    "gndr_cod": "sex",
                    "drugname": "prod_ai",
                    "drug_seq": {
                                    "INDI": "indi_drug_seq",
                                    "THER": "dsg_drug_seq",
                                    "DRUG": "drug_seq"
                                },
                    "outc_code": "outc_cod"
                }
################################################################################
```

Create directories as required

```
os.makedirs("tmp/zip", exist_ok=True)
os.makedirs("tmp/data", exist_ok=True)
for i in REQUIRED_DRUGS:
    os.makedirs(os.path.join(DRUG_FOLDER, i), exist_ok=True)
```

(B)

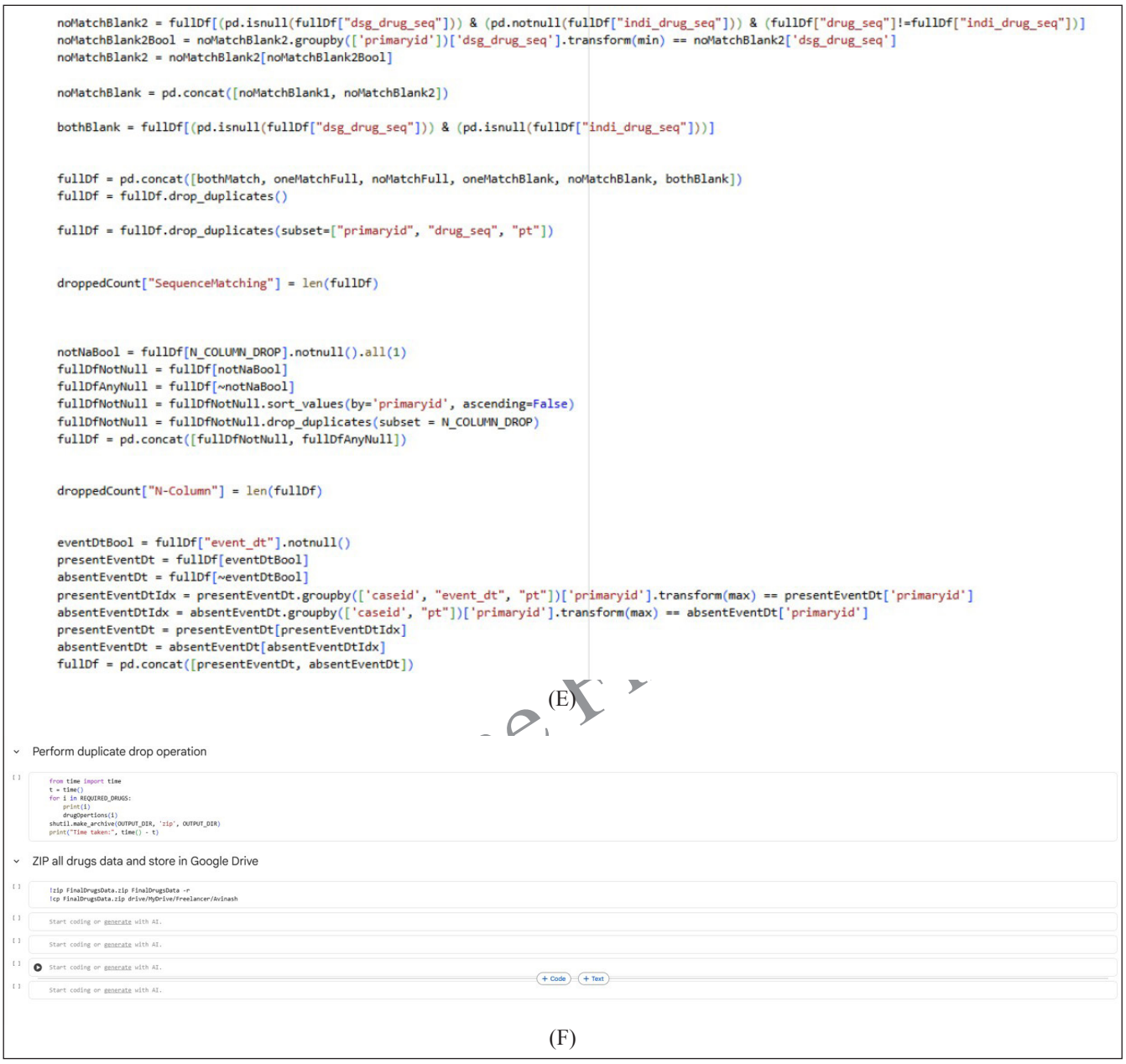
```
def drugOpertions(drugname):
    droppedCount = {}

    drugData = os.listdir("drugs/"+drugname)
    alldf = [pd.read_csv(os.path.join("drugs/"+drugname, i),) for i in drugData]
    fullDf = pd.concat(alldf)


    droppedCount["Original"] = len(fullDf)

    ### Rogue column is added sometimes
    if 'Unnamed: 0' in fullDf.columns:
        fullDf.drop('Unnamed: 0', axis=1, inplace=True)

    ### Priority filtering after merge; getting rows with highest priority
    priorityDict = {"OT":1, "RI":2, "CA":3, "DS":4, "HO":5, "LT":6, "DE":7}
    invPriorityDict = dict(map(reversed, priorityDict.items()))
    notMissingCodDf = fullDf[~fullDf["outc_cod"].isnull()]
    notMissingCodDf = notMissingCodDf.replace({"outc_cod": priorityDict})
    # print(len(notMissingCodDf))
    notMissingCodDfIdx = notMissingCodDf.groupby(['primaryid', 'caseid'])['outc_cod'].transform(max) == notMissingCodDf['outc_cod']
    notMissingCodDf = notMissingCodDf[notMissingCodDfIdx]
    # print(len(notMissingCodDf))
    notMissingCodDf = notMissingCodDf.replace({"outc_cod": invPriorityDict})
    missingCodDf = fullDf[fullDf["outc_cod"].isnull()]
    fullDf = pd.concat([notMissingCodDf, missingCodDf])
    # print(len(fullDf))

    droppedCount["CodPriority"] = len(fullDf)

    bothMatch = fullDf[(fullDf["drug_seq"]==fullDf["dsg_drug_seq"]) & (fullDf["drug_seq"]==fullDf["indi_drug_seq"])]

    oneMatchFull1 = fullDf[(pd.notnull(fullDf["dsg_drug_seq"])) & (pd.notnull(fullDf["indi_drug_seq"])) & (fullDf["drug_seq"]==fullDf["dsg_drug_seq"]) & (fullDf["drug_seq"]!=fullDf["indi_drug_seq"])]
    oneMatchFull1Bool = oneMatchFull1.groupby(['primaryid'])['indi_drug_seq'].transform(min) == oneMatchFull1['indi_drug_seq']
    oneMatchFull1 = oneMatchFull1[oneMatchFull1Bool]
```

(C)

```
oneMatchFull1 = fullDf[(pd.notnull(fullDf["dsg_drug_seq"])) & (pd.notnull(fullDf["indi_drug_seq"])) & (fullDf["drug_seq"]==fullDf["dsg_drug_seq"]) & (fullDf["drug_seq"]!=fullDf["indi_drug_seq"])]
oneMatchFull1Bool = oneMatchFull1.groupby(['primaryid'])['indi_drug_seq'].transform(min) == oneMatchFull1['indi_drug_seq']
oneMatchFull1 = oneMatchFull1[oneMatchFull1Bool]
del oneMatchFull1["indi_drug_seq"]
oneMatchFull1["indi_drug_seq"] = ""
del oneMatchFull1["indi_pt"]
oneMatchFull1["indi_pt"] = ""

oneMatchFull2 = fullDf[(pd.notnull(fullDf["dsg_drug_seq"])) & (pd.notnull(fullDf["indi_drug_seq"])) & (fullDf["drug_seq"]!=fullDf["dsg_drug_seq"]) & (fullDf["drug_seq"]==fullDf["indi_drug_seq"])]
oneMatchFull2Bool = oneMatchFull2.groupby(['primaryid'])['dsg_drug_seq'].transform(min) == oneMatchFull2['dsg_drug_seq']
oneMatchFull2 = oneMatchFull2[oneMatchFull2Bool]
del oneMatchFull2["dsg_drug_seq"]
oneMatchFull2["dsg_drug_seq"] = ""
del oneMatchFull2["start_dt"]
oneMatchFull2["start_date"] = ""
del oneMatchFull2["end_dt"]
oneMatchFull2["end_date"] = ""

oneMatchFull = pd.concat([oneMatchFull1, oneMatchFull2])

noMatchFull = fullDf[(pd.notnull(fullDf["dsg_drug_seq"])) & (pd.notnull(fullDf["indi_drug_seq"])) & (fullDf["drug_seq"]!=fullDf["dsg_drug_seq"]) & (fullDf["drug_seq"]!=fullDf["indi_drug_seq"])]
del noMatchFull["dsg_drug_seq"]
noMatchFull["dsg_drug_seq"] = ""
del noMatchFull["indi_drug_seq"]
```

(D)

Figure 3. Continued



```
noMatchBlank2 = fullDf[(pd.isnull(fullDf["dsg_drug_seq"])) & (pd.notnull(fullDf["indi_drug_seq"])) & (fullDf["drug_seq"]!=fullDf["indi_drug_seq"])]
noMatchBlank2Bool = noMatchBlank2.groupby(['primaryid'])['dsg_drug_seq'].transform(min) == noMatchBlank2['dsg_drug_seq']
noMatchBlank2 = noMatchBlank2[noMatchBlank2Bool]

noMatchBlank = pd.concat([noMatchBlank1, noMatchBlank2])

bothBlank = fullDf[(pd.isnull(fullDf["dsg_drug_seq"])) & (pd.isnull(fullDf["indi_drug_seq"]))]


fullDf = pd.concat([bothMatch, oneMatchFull, noMatchFull, oneMatchBlank, noMatchBlank, bothBlank])
fullDf = fullDf.drop_duplicates()

fullDf = fullDf.drop_duplicates(subset=["primaryid", "drug_seq", "pt"])


droppedCount["SequenceMatching"] = len(fullDf)



notNaBool = fullDf[N_COLUMN_DROP].notnull().all(1)
fullDfNotNull = fullDf[notNaBool]
fullDfAnyNull = fullDf[~notNaBool]
fullDfNotNull = fullDfNotNull.sort_values(by='primaryid', ascending=False)
fullDfNotNull = fullDfNotNull.drop_duplicates(subset = N_COLUMN_DROP)
fullDf = pd.concat([fullDfNotNull, fullDfAnyNull])


droppedCount["N-Column"] = len(fullDf)


eventDtBool = fullDf["event_dt"].notnull()
presentEventDt = fullDf[eventDtBool]
absentEventDt = fullDf[~eventDtBool]
presentEventDtIdx = presentEventDt.groupby(['caseid', "event_dt", "pt"])['primaryid'].transform(max) == presentEventDt['primaryid']
absentEventDtIdx = absentEventDt.groupby(['caseid', "pt"])['primaryid'].transform(max) == absentEventDt['primaryid']
presentEventDt = presentEventDt[presentEventDtIdx]
absentEventDt = absentEventDt[absentEventDtIdx]
fullDf = pd.concat([presentEventDt, absentEventDt])
```

(E)

Perform duplicate drop operation

```
from time import time
t = time()
for i in REQUIRED_DRUGS:
    print(i)
    drugOpertions(i)
shutil.make_archive(OUTPUT_DIR, 'zip', OUTPUT_DIR)
print("Time taken:", time() - t)
```

ZIP all drugs data and store in Google Drive

```
!zip FinalDrugsData.zip FinalDrugsData -r
!cp FinalDrugsData.zip drive/MyDrive/Freelancer/Avinash
```

Start coding or generate with AI.

Start coding or generate with AI.

Start coding or generate with AI.

+ Code + Text

Start coding or generate with AI.

(F)

Figure 3. Data extraction, merging, de-duplication, and output generation process through using Python language.

the six-column deduplication criteria (prod_ai, Age, Gender, event_dt, reporter country, and PT).

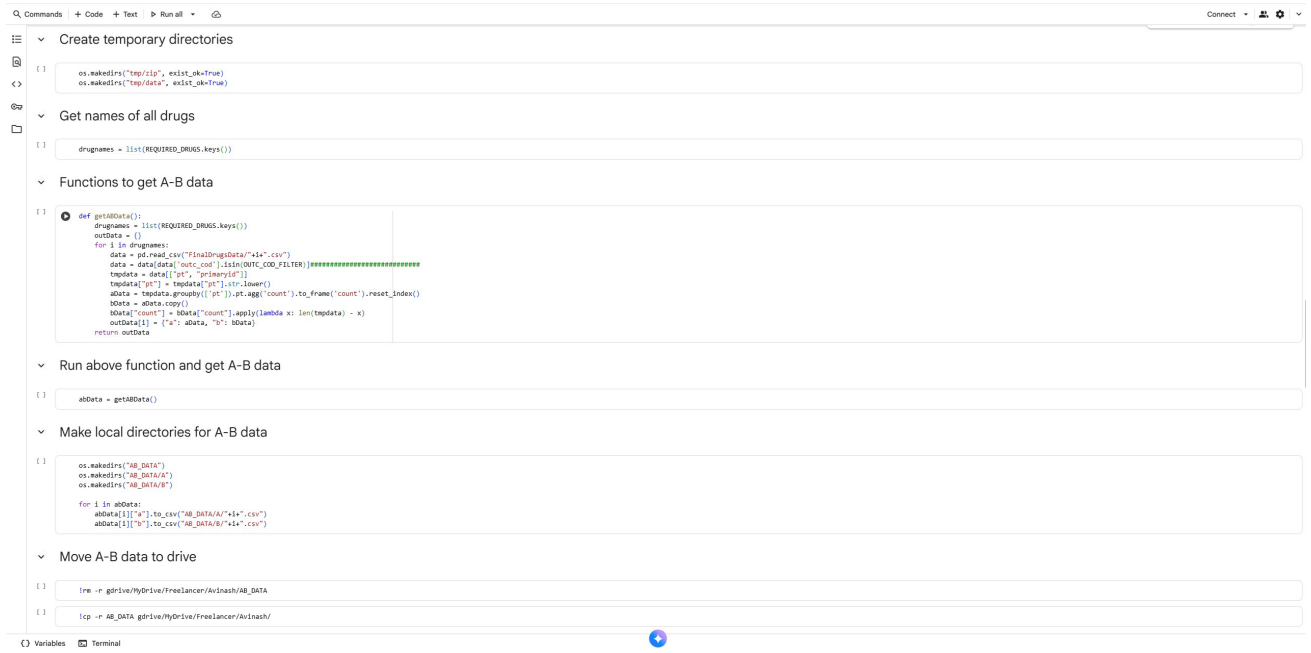
For validation, all cases with Remdesivir as a primary suspect were manually extracted from the FAERS quarterly extract files (Fig. 5A). The deduplication process involved retaining the highest version of a case. If a case ID had multiple versions, only the version with the highest primary ID was retained, while older versions were deleted—except when an older version contained a PT not present in the highest version, in which case that PT was retained.

A specific case (Primary ID: 184498482) was reviewed (Fig. 5C). It was retained due to updated information, while an older version (Primary ID: 184498481) was deleted, except for a unique PT ("Sepsis") that was not reported in the latest version. Similarly, in another case (Case ID: 18050980), the algorithm retained the latest version (Primary ID: 180509803) with five PTs, deleting two duplicates ("Acute respiratory distress syndrome" and "Respiratory failure") (Fig. 5D). Another comparison (Fig. 5E) showed that cases 18059803 and 18144857 had identical values across all six deduplication criteria, leading to the retention of only one version.

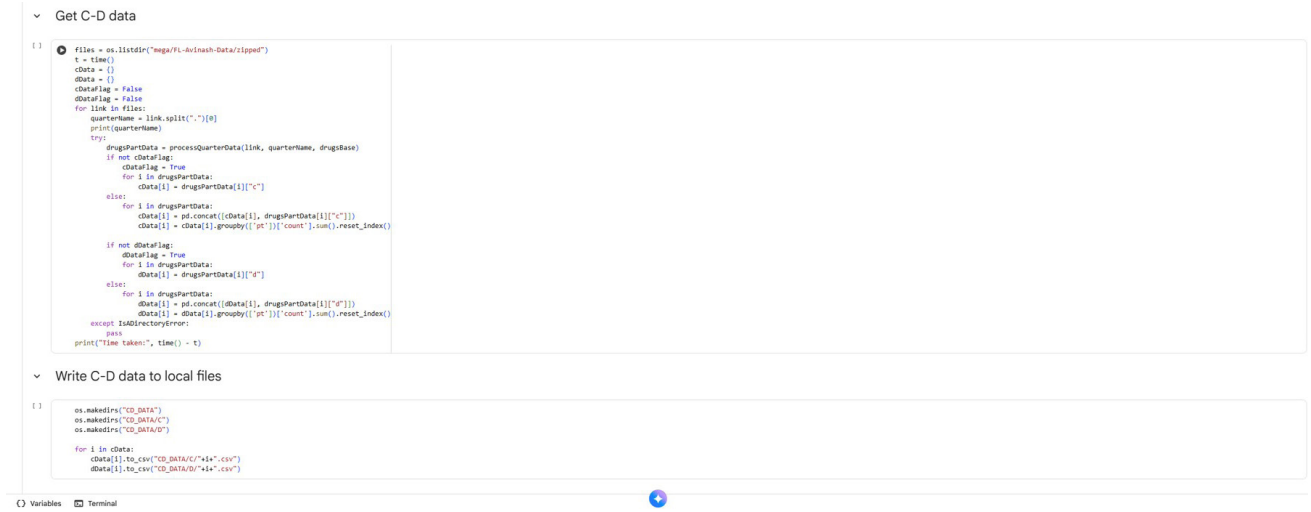
The validation confirmed that the automated Python-based retrieval and deduplication of FAERS data closely aligned with the manual process. Moreover, the deletions performed by the algorithm followed the predefined deduplication criteria, ensuring accurate and reliable results.

(A)

(B)



(C)



(D)

Figure 4. Continued

(E)

Figure 4. The statistical analysis for causality assessment using the Python language.

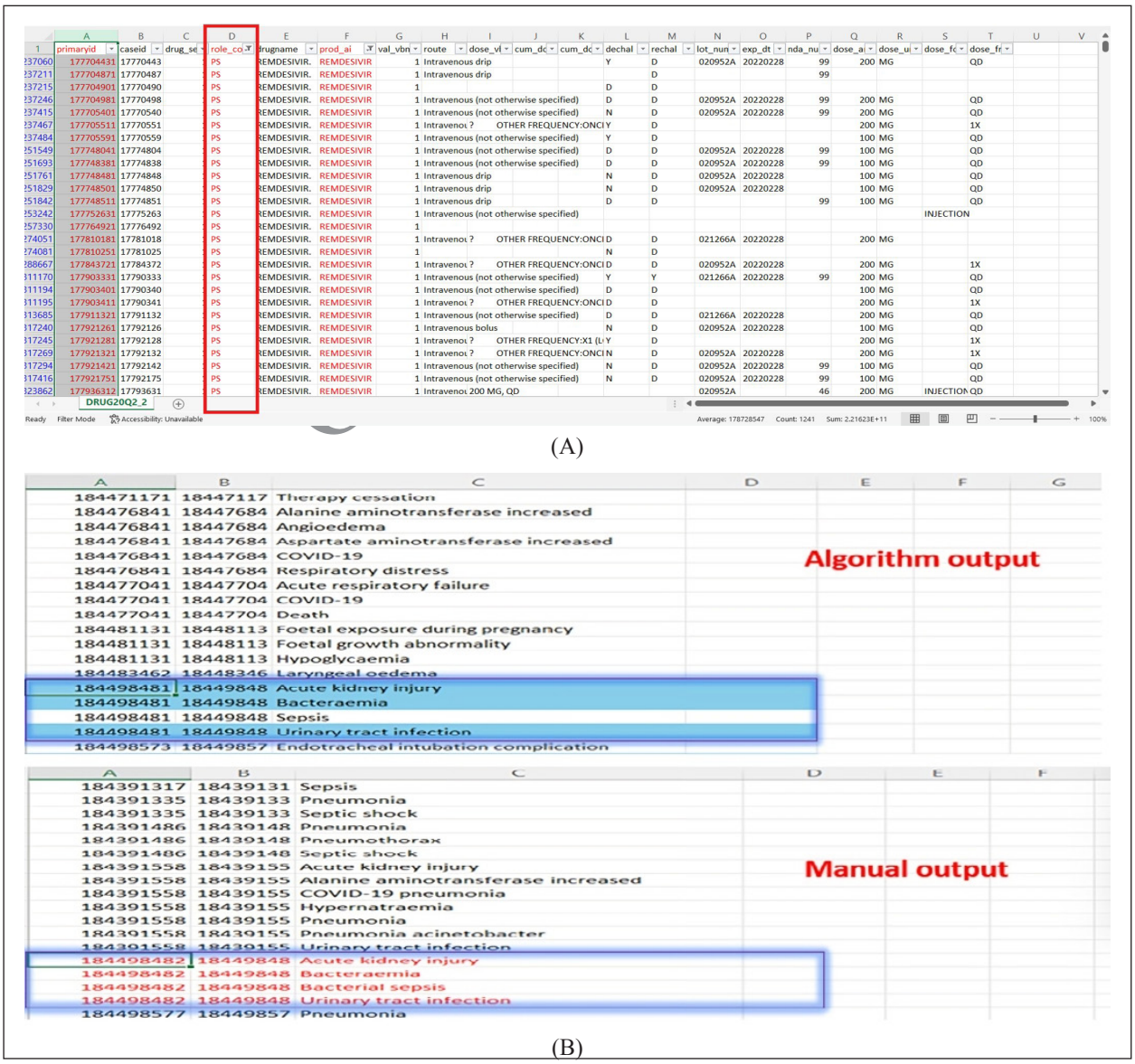
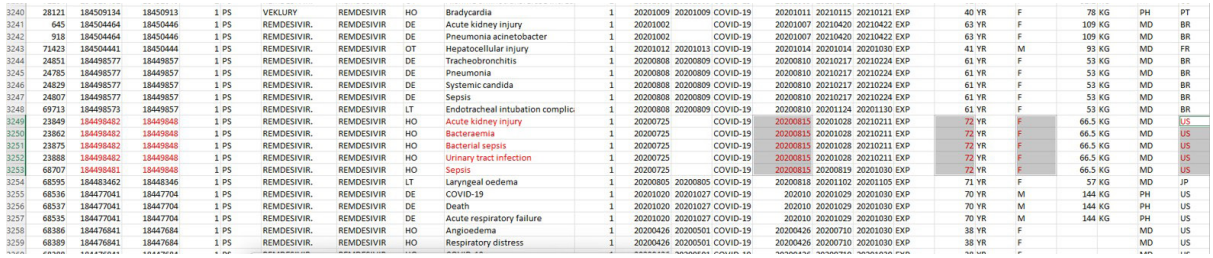


Figure 5. Continued



3240	28121	184509134	18450913	1	PS	VEKLURY	REMDESIVIR	HO	Bradycardia	1	20201009	20201009	COVID-19	20201011	20210115	20210121	EXP	40	YR	F	78	KG	PH	PT
3241	645	184504464	18450446	1	PS	REMDESIVIR.	REMDESIVIR	DE	Acute kidney injury	1	20201002		COVID-19	20201007	20210420	20210422	EXP	63	YR	F	109	KG	MD	BR
3242	918	184504464	18450446	1	PS	REMDESIVIR.	REMDESIVIR	DE	Pneumonia acinetobacter	1	20201002		COVID-19	20201007	20210420	20210422	EXP	63	YR	F	109	KG	MD	BR
3243	71423	184504441	18450444	1	PS	REMDESIVIR.	REMDESIVIR	OT	Hepatocellular injury	1	20201012	20201013	COVID-19	20201014	20201014	20201030	EXP	41	YR	M	93	KG	MD	FR
3244	24851	184498577	18449857	1	PS	REMDESIVIR.	REMDESIVIR	DE	Tracheobronchitis	1	20200808	20200809	COVID-19	20200810	20210217	20210224	EXP	61	YR	F	53	KG	MD	BR
3245	24785	184498577	18449857	1	PS	REMDESIVIR.	REMDESIVIR	DE	Pneumonia	1	20200808	20200809	COVID-19	20200810	20210217	20210224	EXP	61	YR	F	53	KG	MD	BR
3246	24829	184498577	18449857	1	PS	REMDESIVIR.	REMDESIVIR	DE	Systemic candida	1	20200808	20200809	COVID-19	20200810	20210217	20210224	EXP	61	YR	F	53	KG	MD	BR
3247	24807	184498577	18449857	1	PS	REMDESIVIR.	REMDESIVIR	DE	Sepsis	1	20200808	20200809	COVID-19	20200810	20210217	20210224	EXP	61	YR	F	53	KG	MD	BR
3248	69713	184498573	18449857	1	PS	REMDESIVIR.	REMDESIVIR	LT	Endotracheal intubation complic	1	20200808	20200809	COVID-19	20200810	20201124	20201130	EXP	61	YR	F	53	KG	MD	BR
3249	23849	184498482	18449848	1	PS	REMDESIVIR.	REMDESIVIR	HO	Acute kidney injury	1	20200725		COVID-19	20200815	20201028	20210211	EXP	72	YR	F	66.5	KG	MD	US
3250	23862	184498482	18449848	1	PS	REMDESIVIR.	REMDESIVIR	HO	Bacteraemia	1	20200725		COVID-19	20200815	20201028	20210211	EXP	72	YR	F	66.5	KG	MD	US
3251	23875	184498482	18449848	1	PS	REMDESIVIR.	REMDESIVIR	HO	Bacterial sepsis	1	20200725		COVID-19	20200815	20201028	20210211	EXP	72	YR	F	66.5	KG	MD	US
3252	23888	184498482	18449848	1	PS	REMDESIVIR.	REMDESIVIR	HO	Urinary tract infection	1	20200725		COVID-19	20200815	20201028	20210211	EXP	72	YR	F	66.5	KG	MD	US
3253	68707	184498481	18449848	1	PS	REMDESIVIR.	REMDESIVIR	HO	Sepsis	1	20200725		COVID-19	20200815	20200819	20201030	EXP	72	YR	F	66.5	KG	MD	US
3254	68595	184483462	18448346	1	PS	REMDESIVIR.	REMDESIVIR	LT	Laryngeal oedema	1	20200805	20200805	COVID-19	20200818	20201102	20201105	EXP	71	YR	F	57	KG	MD	JP
3255	68536	184477041	18447704	1	PS	REMDESIVIR.	REMDESIVIR	DE	COVID-19	1	20201020	20201027	COVID-19	202010	20201029	20201030	EXP	70	YR	M	144	KG	PH	US
3256	68537	184477041	18447704	1	PS	REMDESIVIR.	REMDESIVIR	DE	Death	1	20201020	20201027	COVID-19	202010	20201029	20201030	EXP	70	YR	M	144	KG	PH	US
3257	68535	184477041	18447704	1	PS	REMDESIVIR.	REMDESIVIR	DE	Acute respiratory failure	1	20201020	20201027	COVID-19	202010	20201029	20201030	EXP	70	YR	M	144	KG	PH	US
3258	68386	184476841	18447684	1	PS	REMDESIVIR.	REMDESIVIR	HO	Angioedema	1	20200426	20200501	COVID-19	20200426	20200710	20201030	EXP	38	YR	F			MD	US
3259	68389	184476841	18447684	1	PS	REMDESIVIR.	REMDESIVIR	HO	Respiratory distress	1	20200426	20200501	COVID-19	20200426	20200710	20201030	EXP	38	YR	F			MD	US

(C)

7019	180496881	18049688	Chills
7020	180497431	18049743	Aspartate aminotransferase increased
7021	180497881	18049788	Alanine aminotransferase increased
7022	180497881	18049788	Aspartate aminotransferase increased
7023	180497881	18049788	Blood albumin decreased
7024	180497951	18049795	Acute kidney injury
7025	180505991	18050599	Death
7026	180505991	18050599	Hypertension
7027	180505991	18050599	Hypoxia
7028	180505991	18050599	Pulmonary function test decreased
7029	180509803	18050980	Acute respiratory distress syndrome
7030	180509803	18050980	COVID-19
7031	180509803	18050980	Haematuria
7032	180509803	18050980	Proteinuria
7033	180509803	18050980	Respiratory failure
7034	180521001	18052100	COVID-19
7035	180521692	18052169	Death
7036	180521692	18052169	Pyrexia
7037	180523502	18052350	Alanine aminotransferase increased
7038	180533421	18053342	Alanine aminotransferase increased
7039	180533591	18053359	Cardiogenic shock
7040	180533591	18053359	Haemodialysis
7041	180533591	18053359	Paralysis
7042	180533591	18053359	Renal failure
7043	180533591	18053359	Sepsis
7044	180533591	18053359	Therapy cessation

(D)

1	primaryid	caseid	drug_s	role_cc	drugname	prod_ai	outc_ci	pt	dsg_dr	start_dt	end_dt	indi_pt	event_dt	mfr_dt	fda_dt	rept_c	age	age_co	sex	wt	wt_coc	occp_c	rep
4748	182817431	18281743	1	PS	REMDESIVIR.	REMDESIVIR	OT	Bundle branch block right	1	20200711	20200716	COVID-19	20200712		20200917	DIR	31	YR	M			PH	US
4749	182817431	18281743	1	PS	REMDESIVIR.	REMDESIVIR	OT	Bradycardia	1	20200711	20200716	COVID-19	20200712		20200917	DIR	31	YR	M			PH	US
5226	182326271	18232627	1	PS	REMDESIVIR (EUA) (REMDESIVIR	REMDESIVIR	DE	Respiratory depression	1	20200706	20200712	COVID-19	20200712		20200904	DIR	78	YR	M	100	KG	MD	US
5331	182217851	18221785	1	PS	REMDESIVIR (EUA) (REMDESIVIR	REMDESIVIR	HO	Acute kidney injury	1	20200710	20200712	Upper res	20200712		20200902	DIR	84	YR	M	114	KG	PH	US
5332	182217851	18221785	1	PS	REMDESIVIR (EUA) (REMDESIVIR	REMDESIVIR	HO	Blood creatinine increased	1	20200710	20200712	Upper res	20200712		20200902	DIR	84	YR	M	114	KG	PH	US
6048	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Urine output decreased	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6049	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Therapy non-responder	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6050	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Respiratory failure	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6051	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Renal impairment	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6052	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Protein urine present	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6053	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	General physical health deterior	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6054	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Disease progression	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6055	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Acute respiratory distress syndro	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6056	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Acute kidney injury	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6057	181448571	18144857	1	PS	REMDESIVIR 100MG	REMDESIVIR	DE	Renal failure	1	20200711	20200711	COVID-19	20200712		20200812	DIR	15	YR	F	43.7	KG	MD	US
6320	181179291	18117929	1	PS	REMDESIVIR.	REMDESIVIR	DE	Liver function test increased	1	20200709	20200712	COVID-19	20200712		20200805	DIR	61	YR	M	81.3	KG	PH	US
6826	180826103	18082610	1	PS	REMDESIVIR.	REMDESIVIR	DE	Maternal exposure during pregna	1	20200712	20200716	COVID-19	20200712	20201106	20201118	EXP	31	YR	F	92	KG	PH	US
6827	180826103	18082610	1	PS	REMDESIVIR.	REMDESIVIR	DE	Foetal death	1	20200712	20200716	COVID-19	20200712	20201106	20201118	EXP	31	YR	F	92	KG	PH	US
6828	180826103	18082610	1	PS	REMDESIVIR.	REMDESIVIR	DE	Endotracheal intubation	1	20200712	20200716	COVID-19	20200712	20201106	20201118	EXP	31	YR	F	92	KG	PH	US
7046	180616131	18061613	1	PS	REMDESIVIR.	REMDESIVIR	OT	Infusion site extravasation	1	20200711	20200713	COVID-19	20200712	20200712	20200723	EXP	89	YR	F	80	KG	PH	US
7223	180509803	18050980	1	PS	REMDESIVIR.	REMDESIVIR	DE	Proteinuria	1	20200711	20200711	COVID-19	20200712	20200812	20200820	EXP	15	YR	F	43.7	KG	MD	US
7224	180509803	18050980	1	PS	REMDESIVIR.	REMDESIVIR	DE	Haematuria	1	20200711	20200711	COVID-19	20200712	20200812	20200820	EXP	15	YR	F	43.7	KG	MD	US
7225	180509803	18050980	1	PS	REMDESIVIR.	REMDESIVIR	DE	COVID-19	1	20200711	20200711	COVID-19	20200712	20200812	20200820	EXP	15	YR	F	43.7	KG	MD	US
7275	180477391	18047739	1	PS	REMDESIVIR.	REMDESIVIR	RI	Intentional dose omission	1	20200709		COVID-19	20200712		20200720	DIR	28	YR	M	88.5	KG	PH	US
7276	180477391	18047739	1	PS	REMDESIVIR.	REMDESIVIR	RI	Liver function test increased	1	20200709		COVID-19	20200712		20200720	DIR	28	YR	M	88.5	KG	PH	US
7287	180450954	18045095	1	PS	REMDESIVIR.	REMDESIVIR	DE	Bradycardia	1	20200711	20200711	COVID-19	20200712	20200911	20200914	EXP	77	YR	F	63.99	KG	PH	US
7288	180450954	18045095	1	PS	REMDESIVIR.	REMDESIVIR	DE	Cardio-respiratory arrest	1	20200711	20200711	COVID-19	20200712	20200911	20200914	EXP	77	YR	F	63.99	KG	PH	US
7289	180450954	18045095	1	PS	REMDESIVIR.	REMDESIVIR	DE	COVID-19 pneumonia	1	20200711	20200711	COVID-19	20200712	20200911	20200914	EXP	77	YR	F	63.99	KG	PH	US
7306	180443121	18044312	1	PS	REMDESIVIR (GS?574)	REMDESIVIR	DE	Acute kidney injury	1	20200709	20200712	COVID-19	20200712		20200720	DIR	85	YR	M	108.6	KG	MD	US
7307	180443121	18044312	1	PS	REMDESIVIR (GS?574)	REMDESIVIR	DE	Blood creatinine increased	1	20200709	20200712	COVID-19	20200712		20200720	DIR	85	YR	M	108.6	KG	MD	US
7308	180443121	18044312	1	PS	REMDESIVIR (GS?574)	REMDESIVIR	DE	Blood urea increased	1	20200709	20200712	COVID-19	20200712		20200720	DIR	85	YR	M	108.6	KG	MD	US
7409	180359331	18035933	1	PS	REMDESIVIR.	REMDESIVIR	OT	Infusion site swelling	1	20200711	20200713	COVID-19	20200712		20200717	DIR	89	YR	F	80	KG	PH	US
7410	180359331	18035933	1	PS	REMDESIVIR.	REMDESIVIR	OT	Infusion related reaction	1	20200711	20200713	COVID-19	20200712		20200717	DIR	89	YR	F	80	KG	PH	US
7454	180349321	18034932	1	PS	REMDESIVIR.	REMDESIVIR	OT	Therapy cessation	1	20200708	20200711	COVID-19	20200712		20200716	DIR	57	YR	M	204	KG	PH	US

(E)

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	pt	A	B	C	D	PRR	ROR	X2	LowerCIROR	UpperCIROR	LowerCIPRR	UpperCIPRR	LowerCIRORGreaterThan1		
2	Acute kidney injury	418	12358	1256181	206749523	5.41756839	5.566989303	1514.347	5.049877805	6.137053429	4.930032269	5.953317486	TRUE		
3	Acute respiratory distress syndrome	80	12696	61364	233599101	23.8433163	23.9872565	1749.001	19.25087167	29.88895693	19.16169488	29.66876018	TRUE		
4	Acute respiratory failure	97	12679	101131	233559334	17.5419459	17.66849911	1512.304	14.46772925	21.57739168	14.3858954	21.39038661	TRUE		
5	Alanine aminotransferase increased	890	11886	292204	233368261	55.7050006	59.80120208	47728.1	55.85809931	64.0226541	52.27910282	59.35540066	TRUE		
6	Aspartate aminotransferase increas	594	12182	243842	233416623	44.5521089	46.67573007	25254.25	42.98157551	50.6873876	41.18323764	48.19656064	TRUE		
7	Atrial fibrillation	74	12702	419650	233240815	3.22503792	3.238000669	113.7826	2.5764897	4.069353871	2.569573773	4.047702273	TRUE		
8	Blood creatinine increased	316	12460	322352	233338113	17.9286276	18.35795718	5053.103	16.41772723	20.5274815	16.0780919	19.99215395	TRUE		
9	Bradycardia	369	12407	190966	233469499	35.3395202	36.36082127	12299.05	32.78119299	40.33133643	31.95576932	39.08157153	TRUE		
10	COVID-19 pneumonia	103	12673	26250	71034089	21.8242983	21.99354809	2039.379	18.10999953	26.70989344	17.99870044	26.46302144	TRUE		
11	Cardiac arrest	149	12627	233624	234426841	11.6643125	11.79015257	1453.285	10.03119424	13.85754222	9.942834405	13.68384312	TRUE		
12	General physical health deterioratio	88	12688	515446	233145019	3.1224093	3.137129668	127.2144	2.543727287	3.868961349	2.535450084	3.845250161	TRUE		
13	Glomerular filtration rate decreased	111	12665	58564	233601901	34.6643116	34.95935613	3622.985	28.99612541	42.14896176	28.79807883	41.72550911	TRUE		
14	Hepatic enzyme increased	180	12596	350759	233309706	9.38542655	9.505256401	1349.891	8.204458931	11.01229222	8.117834479	10.85095191	TRUE		
15	Hepatic function abnormal	45	12731	151098	233509367	5.4468339	5.462552027	163.4263	4.076248801	7.320327119	4.068711709	7.291742843	TRUE		
16	Hypotension	168	12608	806431	232854034	3.81006514	3.847508904	349.3179	3.304181461	4.480179113	3.278581047	4.427707042	TRUE		
17	Hypoxia	149	12627	172759	233487706	15.7737851	15.94811743	2061.489	13.56859418	18.74493746	13.44556643	18.50515547	TRUE		
18	Infusion related reaction	82	12694	410122	233250343	3.65671505	3.673876752	158.5218	2.956724527	4.564973932	2.947017025	4.537321925	TRUE		
19	Infusion site extravasation	118	12658	31300	233629165	68.949005	69.58243702	7873.075	58.02664143	83.43952748	57.59457625	82.54189195	TRUE		
20	Liver function test increased	403	12373	93813	187566911	63.0987137	65.12132601	24536.14	58.95851042	71.92832842	57.30599733	69.47698077	TRUE		
21	Multiple organ dysfunction syndrom	63	12713	129617	187531107	7.13932362	7.169747386	332.67	5.59717853	9.184141135	5.580234106	9.13401494	TRUE		
22	Oxygen saturation decreased	112	12664	283641	233376824	7.22169842	7.276722918	600.8317	6.041331383	8.764739603	6.005432428	8.684291885	TRUE		
23	Renal impairment	156	12620	396530	233263935	7.19513392	7.271714025	833.2094	6.209428022	8.515731992	6.155890106	8.409823969	TRUE		
24	Respiratory failure	222	12554	300684	233359781	13.5030846	13.72418424	2571.541	12.01801483	15.67257453	11.85169093	15.38458056	TRUE		
25	Sepsis	56	12720	536211	233124254	1.91004079	1.914047257	24.33434	1.472135955	2.488613155	1.470745702	2.480548341	TRUE		
26	Septic shock	80	12696	212749	233447716	6.87721803	6.91425154	402.0252	5.549565487	8.614525671	5.527443561	8.556600769	TRUE		
27	Sinus bradycardia	82	12694	38613	233621852	38.8392326	39.08366443	3017.022	31.44788803	48.57346299	31.2947519	48.20252281	TRUE		
28	Transaminases increased	161	12615	89561	233570904	32.8773856	33.28422343	4969.135	28.48829576	38.88753258	28.19521229	38.33709333	TRUE		

(F)

Figure 5. The manual validation of the algorithmic report of data extraction, matching deduplication process.

Table 4. 2 × 2 Contingency table for statistical analysis.

	Specific adverse event(adverse event of interest)	All other events(all other adverse events)	Total
Selected drug of interest	**A**	**B**	**A + B**
All other drugs in the database	C	D	C + D
Total	A + C	B + D	N = A+B+C+D

Ref: Poluzzi *et al*. [10].

- The value A is the count of the number of individual cases with Remdesivir involving a specific adverse event E (e.g., Bacterial infection).
- The value B is the count of the number of individual cases with Remdesivir, involving any other adverse events but E (e.g., Bacterial infection).
- The value C is the count of the number of individual cases involving event E (e.g., Bacterial infection) reported to any other medicinal products but Remdesivir in the database.
- The value D is the count of the number of individual cases with any other adverse events but E (e.g., Bacterial infection) reported to any other medicinal products but Remdesivir in the whole database.

3.2.2. Calculation of ROR and POR

The automated tool directly calculated the ROR and PRR, with the results extracted into Excel files. A total of 12,777 adverse events (PTs) were reported for Remdesivir, and 256 signals were identified. The demographic characteristics of adverse event reports, categorized by age group, gender, and reporter country, are provided in Supplementary File S1.

The output generated using Python and the corresponding downloaded Excel format are shown in Figure 5F. The results produced by the automated tool successfully identified true signals, as detailed in Supplementary File S2 and S3. Due to the large volume of data, we did not manually validate the calculated values of ROR and PRR, as it would be highly time-consuming and require extensive processing of large datasets.

3.3. Regular updating of the data

A code has been generated to regularly update the Megadrive data based on the quarterly update in the FAERS database. The code will have to run in every quarter to update the data, and this is an additional feature of this algorithm.

3.4. Presentation of findings

The findings from this algorithm can be presented in multiple ways, such as tables, pie charts, bar diagram, or in the form of forest plots based on the available results. A model forest plot depiction of ADRs of remdesivir with ROR of >100 is presented in Figure 6.

4. DISCUSSION

Spontaneous ADR and PMS reporting play a critical role in identifying new safety signals in a timely manner, providing early warnings of potential risks associated with drug products. FAERS is one of the most widely used PV databases, consisting of adverse event reports, medication error reports, and product quality complaints related to pharmaceutical products [6]. Disproportionality analysis has become a standard approach for signal detection in PV, with FAERS being commonly used for this purpose [16–19]. The generated signals serve as the foundation for further investigation, allowing regulators to assess the severity and clinical significance of potential safety concerns before making regulatory decisions [4]. However, analysing FAERS data is challenging due to its sheer volume, presence of duplicate cases, and inconsistencies in data recording**,** all of which can introduce bias in signal detection. It is, therefore, crucial to systematically clean, standardize, and deduplicate the data while minimizing human errors and resource requirements to ensure optimal risk estimation.

To facilitate the practical utility of our automated disproportionality analysis tool, we incorporated forest plots as a visual representation of statistically significant ADRs. The integration of forest plots into our automated disproportionality analysis tool represents a significant advancement in the practical application of PV data for policymaking. Unlike traditional tabular outputs, forest plots offer a visually intuitive summary of ADRs, highlighting both the strength of association (ROR) and the precision of estimates (95% CI). This format enables rapid identification of statistically significant safety signals, even by stakeholders without formal training in biostatistics. The application of disproportionality analysis metrics like ROR and PRR is crucial for the early detection of adverse events in PV perspective, which reports the events at a higher rate than expected.

In the current analysis, we have computed ROR values for individual ADRs. The interpretation of ROR mainly involves the point estimate and its CI. In general, an ROR greater than 1 indicates that the drug tested reports adverse events more frequently than the other drugs, which generate a signal. Moreover, this also depends on the precision and the statistical significance of the estimate, especially if clinical and regulatory relevance is considered. For example, our study resulted in a significant demonstration of extravasation (ROR: 120.54; 95% CI: 83.79–157.29), but the CI does not include 1. This is a strong indication to understand a robust association requiring clinical attention or expert validation through further research. In converse, ADRs such as abnormal nasogastric output (ROR: 241.32; 95% CI: −643.96 to 1126.60) and permanent deafness (ROR: 156.76; 95% CI: −395.73 to 709.25) show a wide CI, indicating low precision and high uncertainty resulting in sparse data or rare event reporting. Hence, if there is an

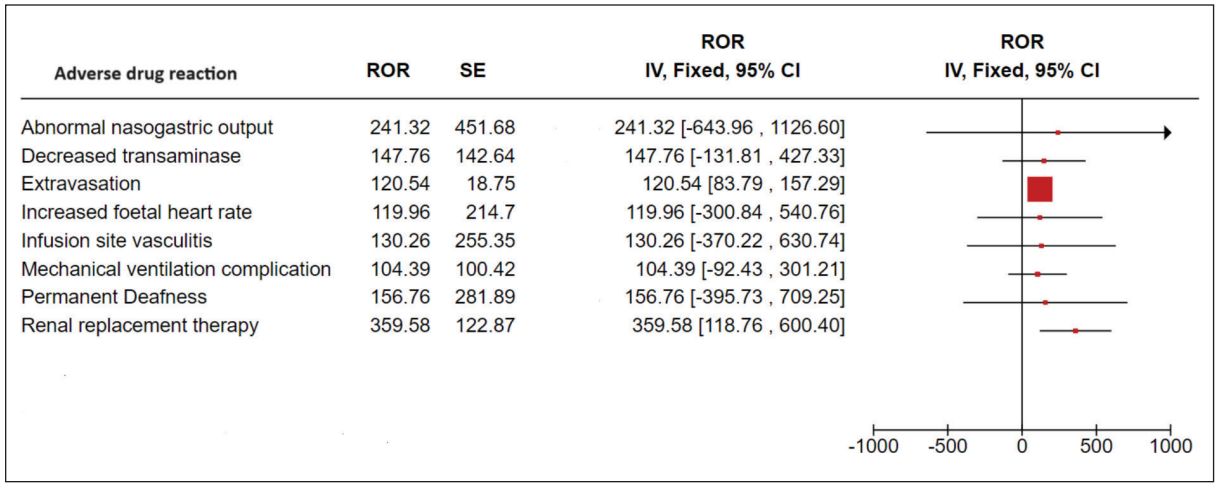


Figure 6. The forest plot presentation for ADRs with ROR > 100.

elevated point estimate representation, the evidence cannot be considered as a statistically robust and clinically not a definite signal unless further research exploration is done. Disproportionality metrics such as ROR and PRR highlight disproportionate reporting and is not establish the causality that may further need to be evaluated. Therefore, the clinical and regulatory authorities can utilise these reports for signal detection, which gives the basis for inferencing the causality and evaluating the risk-benefit ratio.

Compared to previous studies, which often relied on static tables or required external tools like OpenVigil for statistical analysis, our approach offers several key improvements, such as automated visualization where forest plots are generated directly within the tool, eliminating the need for external software or manual formatting. The tool highlights only those ADRs with statistically significant and clinically relevant signals, reducing noise and enhancing focus. The visualization is tailored to support non-technical users, including regulatory bodies and public health officials, by simplifying complex data into actionable insights. With quarterly FAERS data integration, the forest plots reflect the most current safety landscape, supporting timely policy interventions. This approach transforms PV outputs from technical datasets into strategic decision-support tools, bridging the gap between data science and public health policy. By enabling clear visualization of drug safety signals, our tool empowers regulators to act swiftly and confidently in safeguarding public health.

Despite efforts to develop data mining tools, existing solutions have notable limitations, including treating multi-ingredient drugs as a single entity, inefficiencies in deduplication, and the inclusion of all suspects in the analysis, which can distort risk estimates [5]. Moreover, many of these tools lack transparency due to the unavailability of source code**,** limiting their adaptability for researchers [9]. Traditional disproportionality analysis, although widely used [16], is time-consuming, requires extensive human and technical resources, and often relies on external open-source software such as OpenVigil to perform statistical calculations [20]. Despite the utility, this tool is also susceptible to reporting bias, and hence, it is advised to use the tool in a context of clinical plausibility, biological mechanism, and with evidence that is supported from the real-life data.

To overcome these challenges, we developed an automated Python-based tool that mirrors traditional data mining principles while enhancing precision, efficiency, and user accessibility. Unlike existing methods, our tool focuses exclusively on primary suspect cases, providing a more precise estimation of ROR values and enabling the identification of the most accurate safety signals in under two hours**.** Additionally, our approach incorporates a rigorous deduplication process, including both basic matching and a six-step deduplication protocol, ensuring the elimination of redundant cases and yielding highly reliable results. The tool's performance has been rigorously validated against manual processing**,** confirming its accuracy in deduplication and data cleaning.

A key advantage of our tool is its open-source nature, making it easily accessible and modifiable to suit researchers' specific needs. We are currently developing a user-friendly graphical interface, enabling researchers with no programming expertise to efficiently conduct disproportionality analyses. The tool's customization features allow users to analyze multiple drugs and ADRs simultaneously, making it adaptable for various PV applications. Furthermore, the tabulated ROR output format simplifies the reporting process, ensuring ease of interpretation for both researchers and regulatory agencies. Given its robust deduplication**,** high efficiency, and ease of use, our tool represents a major step forward in signal detection and PV research. It has the potential to be widely adopted by researchers, regulatory agencies, and policymakers for refining ADR signal detection methodologies. Additionally, our algorithm is designed to automatically update with each quarterly FAERS database release, ensuring researchers always have access to the most up-to-date PV data. Future work will focus on enhancing the tool's functionalities by

integrating machine learning algorithms for more advanced pattern recognition and risk stratification.

Although the tool had proved reliability and utility on Remdesivir data, its broader adoption necessitates scalability and generalisability across different datasets. In the real world, the PV datasets are designed to handle a large volume of spontaneous reports of multiple drugs and patient data. It is also important to assess the computational efficiency, accuracy, and interpretability of the tool while handling complex and diverse datasets. However, the performance of the tool will be influenced by certain factors such as heterogeneity in the data, disparities in reporting volume, and signal dilution when dealing with polypharmacy or reporting rare drug-events. Scalability is also another challenge with such automated tools, where future iterations could help to improve the performance. Integration of these tools with EudraVigilance, FAERS, and Vigibase is providing the opportunity to evaluate the performance and robustness in the real-world conditions. Moreover, spontaneous reporting system, such as FAERS database is having certain inherent limitations, such as underreporting and selective reporting leading to false positives or false negatives. The true associations can also be distorted by the Weber effect and notoriety bias, and confounders. Therefore, the signals generated by our tool warrant cautious interpretation and imply hypothesis generation rather than confirmation. To strengthen the judgement, it is necessary to complement the findings with clinical judgement validated based on the real world evidence.

In conclusion, our automated tool significantly improves the speed, accuracy, and reliability of signal detection in PV. The findings from the disproportionality analysis signify the mixture of potential and significant results with inconclusive associations. Inclusion of forest plots in PV helps in understanding the signal strength and precision, along with facilitating transparent communication among policymakers. This offers a pragmatic approach for prioritizing the ADRs for drug safety monitoring and opens the door for further research. By addressing the limitations of existing methods and ensuring accessibility through an intuitive interface, this tool empowers researchers and regulatory bodies to make faster, evidence-based decisions that enhance drug safety monitoring and public health outcomes. Moreover, in the future, the tool targets non-technical users to enhance the applicability and usability of the tool; development of a graphical user interface (GUI) is required. The features of the GUI involve data uploading, forest plot interpretation, automated flagging of potential signals, and final exporting of the data. The developed GUI focuses on an agile, user-centric approach with a functional prototype expected to launch in the future. Pilot testing involves the subject and methodological experts to iterate on the improvements. This ensures that the barrier to adoption is minimised, enabling a broader application across the stakeholders in regulatory bodies, PV units at hospitals, and research institutions. In addition, the current study focuses on the implementation of ROR/PRR as the disproportionality measures, as they are widely used and well recognised. Acknowledging that approaches such as MGPS, Empirical Bayes Geometric Mean, and log-likelihood ratio tests offer additional advantages in data handling and accounting for variability. Although we could not incorporate this into the current version, its utility will be recognised in the future development of the model.

5. CONCLUSION

To our knowledge, this is one of the novel approaches to perform the disproportionality analysis of FAERS databases using Python language to perform the signal detection in a lesser time span. This tool ensures more accuracy through adopting an extreme de-duplication process with regular updates. This can be translated into practice through developing suitable interfaces with desktop tools.

5.1. Code availability

All the data related to this work is available in the manuscript and was not deposited in any repository. Any additional data can be made available from the corresponding author upon appropriate request.

6. ACKNOWLEDGMENT

We sincerely acknowledge Manipal Academy of Higher Education (MAHE) for their continuous support and guidance throughout this study. We extend our gratitude to the institution for providing the necessary resources, infrastructure, and an enabling research environment that facilitated the successful completion of this work. We also appreciate the encouragement and insights from faculty members and colleagues, which greatly contributed to the development and validation of this automated tool.

7. AUTHOR CONTRIBUTIONS

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work. All the authors are eligible to be author as per the International Committee of Medical Journal Editors (ICMJE) requirements/guidelines.

8. FINANCIAL SUPPORT

There is no funding to report.

9. CONFLICTS OF INTEREST

The authors report no financial or any other conflicts of interest in this work.

10. ETHICAL APPROVALS

This study does not involve experiments on animals or human subjects.

11. DATA AVAILABILITY

All data generated and analyzed are included in this research article.

12. PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of the

publisher, the editors and the reviewers. This journal remains neutral with regard to jurisdictional claims in published institutional affiliation.

13. USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that they have not used artificial intelligence (AI)-tools for writing and editing of the manuscript, and no images were manipulated using AI.

14. SUPPLEMENTARY MATERIAL

The supplementary material can be accessed at the journal's website: Link here [THIS IS A DUMMY LINK THAT WILL BE CHANGED AFTER PUBLICATION].

REFERENCES

1. Jeetu G, Anusha G. Pharmacovigilance: a worldwide master key for drug safety monitoring. J Young Pharm. 2010;2(3):315–20. https://doi.org/10.4103%2F0975-1483.66802
2. Kengar MD, Patole KK, Ade AK, Kumbhar SM, Patil CD, Ganjave AR. Introduction to pharmacovigilance and monitoring. Asian J Pharm Res. 2019;9(2):116–22. doi: https://doi.org/10.5958/2231-5691.2019.00019.4
3. Alomar M, Tawfiq AM, Hassan N, Palaian S. Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future. Ther Adv Drug Saf. 2020;11:2042098620938595. doi: https://doi.org/10.1177/2042098620938595
4. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6. doi: https://doi.org/10.1002/pds.677
5. Khaleel MA, Khan AH, Ghadzi SMS, Adnan AS, Abdallah QM. A standardized dataset of a spontaneous adverse event reporting system. Healthcare (Basel). 2022;10(3): 420. doi: https://doi.org/10.3390/healthcare10030420
6. FDA. The FAERS public dashboard and its value to the pharmaceutical industry. United States: CDER SBIA Chronicles; 2018. Available from: https://www.fda.gov/media/114303/download?attachment
7. Fukazawa C, Hinomura Y, Kaneko M, Narukawa M. Significance of data mining in routine signal detection: analysis based on the safety signals identified by the FDA. Pharmacoepidemiol Drug Saf. 2018;27(12):1402–8. doi: https://doi.org/10.1002/pds.4672
8. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016;3:160026. doi: https://doi.org/10.1038/sdata.2016.26
9. Wang L, Jiang G, Li D, Liu H. Standardizing adverse drug event reporting data. J Biomed Semantics. 2014;5:36. https://doi.org/10.1186%2F2041-1480-5-36
10. Poluzzi E, Raschi E, Piccinni C, De Ponti F. Data mining techniques in pharmacovigilance: analysis of the publicly accessible FDA adverse event reporting system (AERS). Data mining applications in engineering and medicine. IntechOpen; 2012. doi: https://doi.org/10.5772/50095
11. Hauben M, Reich L, Demicco J, Kim K. 'Extreme duplication' in the US FDA adverse events reporting system database. Drug Saf. 2007;30(6):551–4. doi: https://doi.org/10.2165/00002018-200730060-00009
12. US Food and Drug Administration. Guidance for industry E2B (M): data elements for transmission of individual case safety reports. 2010. Available from: https://www.fda.gov/files/drugs/published/E2BM-Data-Elements-for-Transmission-Of-Individual-Case-Safety-Reports.pdf
13. US Food and Drug Administration. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. Rockville, MD: Food and Drug Administration. 2005. Available from: https://www.fda.gov/files/drugs/published/Good-Pharmacovigilance-Practices-and-Pharmacoepidemiologic-Assessment-March-2005.pdf
14. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18(6):427–36. doi: https://doi.org/10.1002/pds.1742
15. Van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(1):3–10. doi: https://doi.org/10.1002/pds.668
16. Palapra H, Viswam SK, Kalaiselvan V, Undela K. SGLT2 inhibitors associated pancreatitis: signal identification through disproportionality analysis of spontaneous reports and review of case reports. Int J Clin Pharm. 2022;44(6):1425–33. doi: https://doi.org/10.1007/s11096-022-01476-7
17. Subeesh V, Maheswari E, Singh H, Beulah TE, Swaroop AM. Novel adverse events of iloperidone: a disproportionality analysis in US Food and Drug Administration Adverse Event Reporting System (FAERS) database. Curr Drug Saf. 2019;14(1):21–6. doi: https://doi.org/10.2174/1574886313666181026100000
18. Zeba Z, Shettigar A, Lukose L, Kaur G, Nair G, Haider N, *et al.* Disproportionality analysis of tumour lysis syndrome (TLS) associated with Bruton's tyrosine kinase inhibitor (BTKi) using Food and Drug Administration Adverse Events Reporting System (FAERS) database. Research Square; 2022. https://doi.org/10.21203/rs.3.rs-1849066/v1
19. Undela K, Kalaiselvan V, Gudi SK, Viswam SK, Ali SK. Risk of serious skin and subcutaneous tissue disorders for nimesulide among the pediatric population: a jeopardy identified through the analysis of global individual case safety reports. Expert Opin Drug Saf. 2023;2023:1. doi: https://doi.org/10.1080/14740338.2023.2274416
20. Böhm R, Von Hehn L, Herdegen T, Klein HJ, Bruhn O, Petri H, *et al.* OpenVigil FDA - Inspection of U.S. American Adverse Drug Events Pharmacovigilance Data and Novel Clinical Applications. PLoS One. 2016;11(6):e0157753. doi: https://doi.org/10.1371/journal.pone.0157753

How to cite this article:
Gurram D, Laddha A, Rajendran R, Rashid M, Poojari PG, Nair S, Khan S, Subramonian K, Vardhan MC, Jackeray R, Thunga G. Development and evaluation of Python language processed automated disproportionality analysis system for FAERS database. J Appl Pharm Sci. 2026. Article in Press. http://doi.org/10.7324/JAPS.2026.245748