Drug Safety (2024) 47:271–284
https://doi.org/10.1007/s40264-023-01391-4

ORIGINAL RESEARCH ARTICLE

Enhancing Transparency in Defining Studied Drugs: The Open-Source Living DiAna Dictionary for Standardizing Drug Names in the FAERS

Michele Fusaroli[1] · Valentina Giunchi[1] · Vera Battini[2] · Stefano Puligheddu[1] · Charles Khouri[3,4] · Carla Carnovale[2] · Emanuel Raschi[1] · Elisabetta Poluzzi[1]

Accepted: 11 December 2023 / Published online: 4 January 2024

Abstract

Introduction In refining drug safety signals, defining the object of study is crucial. While research has explored the effect of different event definitions, drug definition is often overlooked. The US FDA Adverse Event Reporting System (FAERS) records drug names as free text, necessitating mapping to active ingredients. Although pre-mapped databases exist, the subjectivity and lack of transparency of the mapping process lead to a loss of control over the object of study.

Objective We implemented the DiAna dictionary, systematically mapping individual free-text instances to their corresponding active ingredients and linking them to the World Health Organization Anatomical Therapeutic Chemical (WHO-ATC) classification.

Methods We retrieved all drug names reported to the FAERS (2004–December 2022). Using existing vocabularies and string editing, we automatically mapped free text to ingredients. We manually revised the mapping and linked it to the ATC classification.

Results We retrieved 18,151,842 reports, with 74,143,411 drug entries. We manually checked the first 14,832 terms, up to terms occurring over 200 times (96.88% of total drug entries), to 6282 unique active ingredients. Automatic unchecked translations extend the standardization to 346,854 terms (98.94%). The DiAna dictionary showed a higher sensitivity compared with RxNorm alone, particularly for specific drugs (e.g., rimegepant, adapalene, drospirenone, umeclidinium). The most prominent drug classes in the FAERS were immunomodulating (37.40%) and neurologic drugs (29.19%).

Conclusion The DiAna dictionary, as a dynamic open-source tool, provides transparency and flexibility, enabling researchers to actively shape drug definitions during the mapping phase. This empowerment enhances accuracy, reproducibility, and interpretability of results.

1 Introduction

1.1 The Need for Transparency in Data Pre-Processing

Spontaneous reporting systems (SRSs) are public and private services collecting individual case safety reports (ICSRs) of suspected adverse drug reactions to timely detect potential drug safety issues [1, 2]. These ICSRs come from various regional, national, and manufacturer databases, each with different languages, rules, and forms for data storage. Additionally, SRSs collect ICSRs from different kinds of reporters (e.g., manufacturers, healthcare professionals, lawyers, consumers) and use both paper and electronic forms. Consequently, ICSRs vary greatly in quality and completeness and may include duplicates. To facilitate the exchange of ICSRs among different

Michele Fusaroli and Valentina Giunchi are equally contributed to this work.

✉ Michele Fusaroli
michele.fusaroli2@unibo.it

1 Unit of Pharmacology, Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

2 Department of Biomedical and Clinical Sciences, Pharmacovigilance and Clinical Research, International Centre for Pesticides and Health Risk Prevention, ASST Fatebenefratelli-Sacco, Università degli Studi di Milano, Milan, Italy

3 Pharmacovigilance Unit, Grenoble Alpes University Hospital, Grenoble, France

4 HP2 Laboratory, Inserm U1300, University of Grenoble Alpes, Grenoble, France

Key Points

An exhaustive definition of the object of study should consider drug names-to-ingredient mapping.

DiAna dictionary provides a transparent and modifiable mapping, improving reproducibility and interpretability of results.

stakeholders, the International Council for Harmonization (ICH) has established standards for data storage and transmission formats (E2B-R3, last updated 17 January 2023) [3], providing leads to uniform the collection of data and simplify pharmacovigilance activities; however, transitioning to the new standard is expected to be challenging. Moreover, free text fields (with misspellings, out-of-context information, and the use of different languages and lexicons) remain significant challenges. Before conducting any statistical analysis, it is crucial to address this heterogeneity by coding the data into standardized lexicons, deduplicating entries, and potentially filling missing information.

In a recent meta-epidemiological study, the United States Food and Drug Administration (US FDA) Adverse Event Reporting System (FAERS) and the World Health Organization (WHO) VigiBase emerged as the two SRSs most commonly used in disproportionality analysis—a quantitative method to identify potential unknown adverse drug reactions—in 40% and 20% of the studies, respectively [4]. The wide use of the FAERS is due to its free access and its large catchment area mirroring the entire world (even if with a large representativity for the US). There are two ways the FDA provides free access to the FAERS data, with notable differences. First, the FDA provides pre-processed data that are available through an online public dashboard [5]. The dashboard was developed for transparency reasons and to promote higher-quality reporting. However, this tool utilizes an undisclosed and partial cleaning procedure, lacking duplicate detection and providing only partial access to ICSR information. The public dashboard is therefore unsuitable for complex analyses. Second, the FDA provides raw quarterly data (both in ASCII and XML format) [6] that allows to knowingly perform and document the entire pre-processing procedure. This cleaning and normalization procedure requires the researchers a conspicuous effort and multiple operative choices; for example, how to deal with duplicates, how to deal with dates that up to 2012 were completed automatically when partial, how to deal with unclear entries (e.g., in 2019 'RN' was often recorded as a reporter type; while this entry may refer to registered nurses, it is not documented in the 'readme' file of the FAERS). For this reason, multiple tools have been made available to access already cleaned versions of the FAERS [7, 8], where the choices necessary in the pre-processing have already been implemented. Nonetheless, these choices are seldom driven by objectivity alone and must be considered in both the design and interpretation [9], because the same database cleaned with different procedures may give different results [10]. The lack of transparency in the pre-processing, and subsequently the lack of replicability, heavily impacts the credibility of SRSs, already diminished due to their inherent bias (e.g., underreporting, notoriety bias, channeling bias [2]), which hinders any interpretation of disproportionality analysis beyond the generation of hypotheses [11].

Over time, researchers have developed various tools for accessing, standardizing, and deduplicating data from SRSs. These efforts have predominantly focused on the WHO VigiBase, which is accessible only via subscription or by accredited centers. However, there has been less cooperation and consensus regarding similar initiatives for the most commonly used SRS in research, namely the FAERS. Given the extreme rawness and heterogeneity of FAERS data, a meticulous cleaning process guided by clinical and pharmacological reasoning is necessary before conducting any analysis [12]. Throughout each stage of this process, it is crucial to uphold collaboration among multiple professionals and maintain an understanding of the data collection features and the relevant underlying theory for the phenomenon under investigation [13]. Additionally, it is of utmost importance for researchers to not only be in control and knowledgeable of the pre-processing procedure but also to be transparent about their operational choices, allowing for external assessment and interpretation [9].

1.2 Drug Nomenclature Issues

Each study centers on defining its object of study. In pharmacovigilance, this involves precisely defining the drugs and the adverse events under investigation. Varying definitions can lead to different cases being identified and yield different results in disproportionality analysis. This discrepancy affects the study's sensitivity, specificity, and accuracy; it is essential for result interpretation that the definitions adopted are transparent. It is therefore not surprising that the first aspect addressed in IMI PROTECT, a project aiming to establish best practices for analyzing SRSs, focused on the effect of adopting different definitions of the adverse events being studied [14]. These events are usually automatically coded in SRSs using the Medical Dictionary for Regulatory

Activities (MedDRA). IMI PROTECT therefore focused on the benefits and costs of grouping different terms denoting the same phenomenon (e.g., 'hepatitis', 'liver injury', 'transaminases increased').

Contrary to the events, coded using MedDRA, the FAERS raw data provide drug information only in free text format. In fact, drugs can be recorded in the FAERS using their brand names, active ingredients, international nonproprietary names (INNs, as defined by the WHO), United States adopted names (USANs, defined by the USAN council), terms from different languages or abbreviations. Misspellings might easily occur, and the drug name is sometimes followed by dose, route, and formulation details. Therefore, defining the drug of interest should not only involve grouping together various active ingredients but also assessing whether the mapping of raw drug names to active ingredients aligns appropriately with the specific inquiry of the study.

An exhaustive objective standardization of these entries is unattainable; for example, the same brand name may refer to different compositions in different countries, or two brand names may be just one letter apart, thus being extremely susceptible to misspellings. The inconsistencies that derive from the multiple operative options and the researcher's personal choices can affect case retrieval and impair replicability among studies. Nonetheless, no comprehensive drug name dictionary is routinely used for published FAERS analyses, and no consensus on the various dictionaries produced for FAERS has been achieved. Already-published analyses using FAERS data are rarely transparent on the cleaning choices adopted, lacking documentation on whether and how the FAERS was prepared for statistical analyses. The FAERS system itself, which does have a formal dictionary through which it cleans the data for the public dashboard, does not make it publicly available.

To better contextualize this work, we performed a small meta-research study to identify the practices of drug names standardization in published FAERS analyses. Among the 17 studies conducted using the FAERS quarterly data and published in February 2023 (accessed on PubMed on 15 March 2023), 10 did not state any drug standardization process, six used an automatic translation via dictionaries, and one performed a manual translation but did not make it publicly available (Table S1 in the electronic supplementary material [ESM]). This lack of transparency in drug names-to-ingredient mapping also affects some free ready-to-use pharmacovigilance databases that provide already pre-processed FAERS data [7, 8]. Moreover, many ready-to-use databases standardize drugs only according to US dictionaries of drug names (e.g., RxNorm, orange book), thus failing to identify foreign drug names, misspellings, and other free text issues that were described above.

Other tools have been developed to standardize FAERS drug names by automatically detecting potential misspellings. While this approach saves time, it can lead to mistranslations when similar drug names refer to formulations with different active ingredients [16]. Another attempt to standardize FAERS drug names was made by Wong et al. [17], who produced a manually revised translation of the LAERS drug names (the FAERS system up to 2012), with transparent explanations of their choices. Nonetheless, this dictionary is not publicly available and has not been adopted for use.

Pre-mapped databases prove highly valuable for signal detection, especially when achieving precise control over the definition of the study object is challenging due to the extensive number of drugs and events under investigation. However, the inherent subjectivity and insufficient transparency in the mapping process impact the level of control over defining the study object, which is a critical aspect in signal refinement.

1.3 Aim of the Study

In this work, we follow the efforts of Wong et al. [17] and extend their work to consider previously unattended nomenclature issues. We propose a collaborative and open-source drug name-to-ingredient dictionary for standardizing the FAERS updated to December 2022, together with a transparent report of the data cleaning protocol to identify and resolve drug nomenclature issues. This pharmacovigilance tool enriches DiAna, an R package for Disproportionality Analysis, with a pre-mapped dataset for signal detection. Unlike conventional pre-mapped datasets, the DiAna dictionary, providing access to drug mapping and linked to the ATC, empowers researchers with increased control over the definition of the object of study, a crucial element in signal refinement. The DiAna dictionary can thus enhance replicability and accuracy of disproportionality analyses, and a more appropriate interpretation of their results.

2 Methods

2.1 The US FDA Adverse Event Reporting System (FAERS) Database

We downloaded FAERS Quarterly Data (trimestral) Extract Files [6] in ASCII format from 04Q1 to 22Q4. These files are composed of five tables linked through a primary key ('primaryid') identifying a specific version of a report: DEMO (demographic and administrative information), DRUG (information on reported medications), REAC (adverse events), OUTC (outcomes), and RPSR (report sources).

DRUG is also linked through 'primaryid' and a secondary key ('drug_seq'), identifying a specific medication within a report, to another two tables: THER (start dates and end dates for the reported medications) and INDI (indications for using the reported medications).

2.2 Automatic Setup of the Dictionary

We combined all DRUG quarters into one database. We focused on three columns used to identify the medicinal product:

- Drugname, recording the name of the medicinal product.
- Prod_ai, recording the product's active ingredients, when available.
- Val_vbm, recording whether the source of drug name was a validated trade name (value = 1) or a verbatim name (value = 2).

Since our aim was the translation to active ingredients, we did not consider the column 'val_vbm'. We instead retrieved all the unique terms from the other two columns (i.e., Prod_ai and Drugname), lowered upper cases, and removed multiple spaces, leading and trailing spaces and punctuation, and spaces between parentheses. We merged these pre-formatted unique terms with brand names and ingredients recorded in the RxNorm (https://www.nlm.nih.gov/research/umls/rxnorm/) and WHO-ATC substances [18] to create a starting dictionary with automatic translations of medications sold in the US market to their active ingredients. The merging process was also repeated after several rounds of text editing, during which we removed leading or trailing spaces and specific terms or symbols such as chirality indicators (e.g., '+', '−', 'd', 's') and text between brackets or caret symbols (see Fig. 1).

2.3 Manual Revision

We manually revised all the automatic translations starting with the most frequently reported ones up to the terms recorded in a minimum of 200 reports (and beyond, ongoing). After an initial pilot study, seven operators independently revised translations, requiring a consensus with the two leading authors in cases of doubt or difficult translations. The operators consisted of the two leading authors, MF and VG, along with an interdisciplinary group of trained operators experienced in pharmacovigilance, including VB, SP, CB, MB, and LP.

When automatic translation failed to translate drug names, we attempted to code them to active ingredients that were already part of the initial set of active ingredients. If new ingredients were encountered, we expanded the set accordingly. We conducted manual searches for foreign drug names using online databases (e.g., DrugBank.com [19] and Drugs.com [20]), manufacturer websites, and websites storing information from foreign package labels (e.g., Kusuri-no-Shiori –drug information sheets– from the Japanese regulatory agency, accessed at https://www.rad-ar.or.jp/siori/english/).

2.4 Nomenclature Issues

Multiple issues were identified in the process of translating drug names, including brand names and abbreviations) to active ingredients (e.g., 'Zantac' was translated into 'ranitidine').

- A drug may include multiple ingredients. We translated the drug to all its ingredients and ordered them alphabetically, separating them by a semicolon. For example, 'Entresto' was translated into 'sacubitril;valsartan'.
- The spelling of an active ingredient can be different between the USAN and the INN; for example, acetaminophen (USAN) = paracetamol (INN); amphetamine (USAN) = amfetamine (INN); dimethicone/simethicone (USAN) = simeticone (INN); cysteamine (USAN) = mercaptamine (INN). We gave preference to the INN.
- The active ingredient may be recorded in languages different from English (e.g., acide folique). We translated everything into the English INN.
- Typing mistakes can occur (e.g., 'zoplicone' instead of 'zopiclone'; 'Diavan®' instead of 'Diovan®'). We manually fixed the mistakes taking into account the INN.
- The same drug name may contain different ingredients in different countries (e.g., Gaster® contains famotidine in the US, omeprazole in Japan, cromoglicic acid in Italy; Previscan® contains fluindione in the US, pentoxifylline in Italy and Spain; Furix® contains furosemide in the US, cefuroxime in India). In these cases, we translated the brand name to the active ingredient contained in the US packaging, assuming that the US is more represented in the FAERS than other countries. When the brand name of interest was not sold in the US, we checked the most reported country in the FAERS for that specific brand name.
- The drug name may be missing or underspecified.
    - When there was no medication, we translated the drug name field to 'no medication'.
    - When the medication was unspecified, we translated the drug name field as 'unspecified'.
    - Unspecified drug-class terms were translated to the most specific term possible (e.g., 'water pills' as 'diuretics, unspecified', 'antihypertensives' as 'antihypertensives, unspecified').

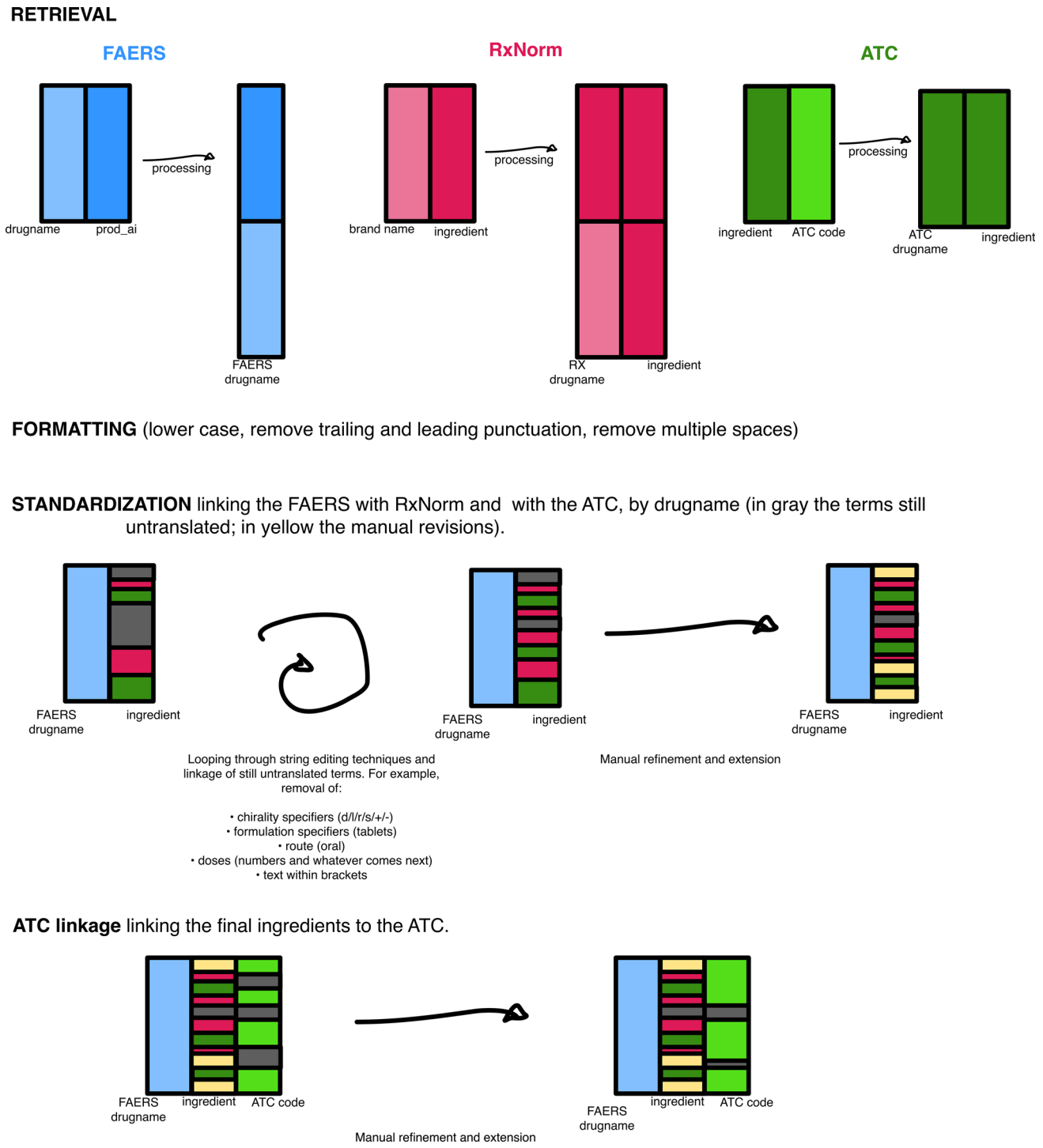

Fig. 1 Translation pipeline method. Flowchart showing the procedure to translate drug names to active ingredients. *FAERS* US FDA adverse event reporting system, *ATC* anatomical therapeutic chemical

- We specified when drug (or placebo) consumption occurred in a clinical trial (such as when blinding was specified, or when the investigational name was used, e.g., cc-223 for onatasertib) translating the drug name as '[active ingredient], trial', 'placebo, trial', or 'unspecified, trial'.
- Additionally, we decided to standardize terms other than drugs to broader categories, since specific details are seldom provided: minerals (e.g., calcium), vitamins (e.g.,

vitamin B5, independent of the route of administration; vitamin B12, independent of the form, e.g., cyanocobalamin, mecobalamin–), devices (e.g., intrauterine contraceptive device), vaccines (e.g., coronavirus disease 2019 [COVID-19] vaccine), and phytotherapies (e.g., Plantago spp).

- If a drug name was reported followed by a non-coherent active ingredient in square brackets, we assumed that an error was made during the compilation by the pharmacovigilance expert. In this case, we translated the entry based on the drug name alone, considering incorrect the active ingredient listed in square brackets.

The whole list of standardized names is available in the open-source repository https://osf.io/zqu89/, together with the pre-mapped dataset.

2.5 Linkage to the Anatomical Therapeutic Chemical (ATC) Classification

Furthermore, we performed data linkage between the DiAna dictionary and the hierarchical ATC classification, which was downloaded from the WHO Collaborating Centre for Drug Statistics and Methodology website [18] using the R package 'rvest'. Since this classification is mainly a tool for drug utilization research, the same active ingredient may be given more than one ATC code if it is available in multiple strengths or routes of administration with clearly different therapeutic uses [21]. We linked the final list of individual active ingredients from our translation with the ATC classification, manually integrating different choices in the nomenclature (e.g., 'vitamin B9' [folic acid] to B03BB01), for classes of drugs (e.g., 'antihypertensives, unspecified' to C02), and drugs recorded in the ATC only in combination (glecaprevir and pibrentasvir both to J05AP57). We have here linked each active ingredient to all its ATC codes (most of the time it is not possible to discriminate between the different ATC codes based only on the drug name), but since sometimes it is important to count each ingredient only once, we also proposed a unique primary ATC code for each ingredient. To this end, we prioritized the first level in the following order ('H', 'J', 'P', 'L', 'M', 'N', 'C', 'G', 'R', 'B', 'D', 'A', 'S', 'V'). Some exceptions did not fit the selected prioritization order and we manually corrected them, i.e., vitamin C was automatically classified as G and we moved it to A, sex hormones having both a genitourinal and an immunomodulating code were classified in H and we moved them to G, and sodium and calcium chloride were classified in B and we moved them to A. Importantly, linking the DiAna dictionary with WHO-ATC is useful to define or visualize drugs of interest as grouped in ATC classes, but it may not be suitable for identifying specific drug formulations due to the often insufficient information provided by SRS data.

3 Results

We downloaded the FAERS quarterly data up to 22Q4 and retrieved 18,151,842 ICSRs, for a total of 74,143,411 drug entries (92.81% allegedly recorded using a validated trade name) and 955,778 unique Drugname and Prod_ai terms (see Fig. 2). After the initial formatting, we reduced them to 793,274 unique entries. The automatic procedure involved the translation of 346,854 terms (98.94% of total drug entries) and the manual validation covered the first 14,832 terms (96.88%) up to 174 occurrences (<0.00015%, ongoing).

A total of 6282 unique ingredients were included in the DiAna dictionary, of which 3209 were linked to the ATC classification. The most common primary ATC classes in the FAERS, after translation with the DiAna dictionary, were antineoplastic and immunomodulating (reported in 37.40% of FAERS reports), nervous system (29.19%), alimentary tract (25.18%), and cardiovascular agents (20.17%) [see Fig. 3]. The most frequently reported medicinal products were paracetamol (5.45%), acetylsalicylic acid (4.62%), adalimumab (3.81%), etanercept (3.35%), levothyroxine (3.17%), and ranitidine (3.13%) [see Table 1]. When compared with the untranslated formatted FAERS and with the FAERS translated according to RxNorm, the translation based on the DiAna dictionary showed clear advantages in case retrieval (98.94% of total drug entries against 76.32% by RxNorm). Among the most reported medicinal products, DiAna allowed retrieving more cases than RxNorm, from a ratio of 1.01 for etanercept (638,427 vs. 632,130) to a ratio of 8.55 for ranitidine (69,883 vs. 597,604). Due to differences in nomenclature, some ratios were not calculated (e.g., paracetamol is translated to acetaminophen and acetylsalicylic acid to aspirin by RxNorm). For some drugs, the added value of DiAna translation for case retrieval was extremely high; for example, rimegepant (ratio = 277.91; 6392 vs. 23), adapalene (122.60; 174,711 vs. 1425), drospirenone (108.49; 86,356 vs. 796), and umeclidinium (105.66; 45,751 vs. 433; not shown in the table). For example, products in RxNorm that contain adapalene include adapalene, Differin, and Epiduo. However, 96% of the cases of adapalene identified in the FAERS using the DiAna dictionary were related to the drug name Proactiv MD. The translation also took into account information about placebo and experiments, thus identifying 50,967 reports as generated within trials (0.28%).

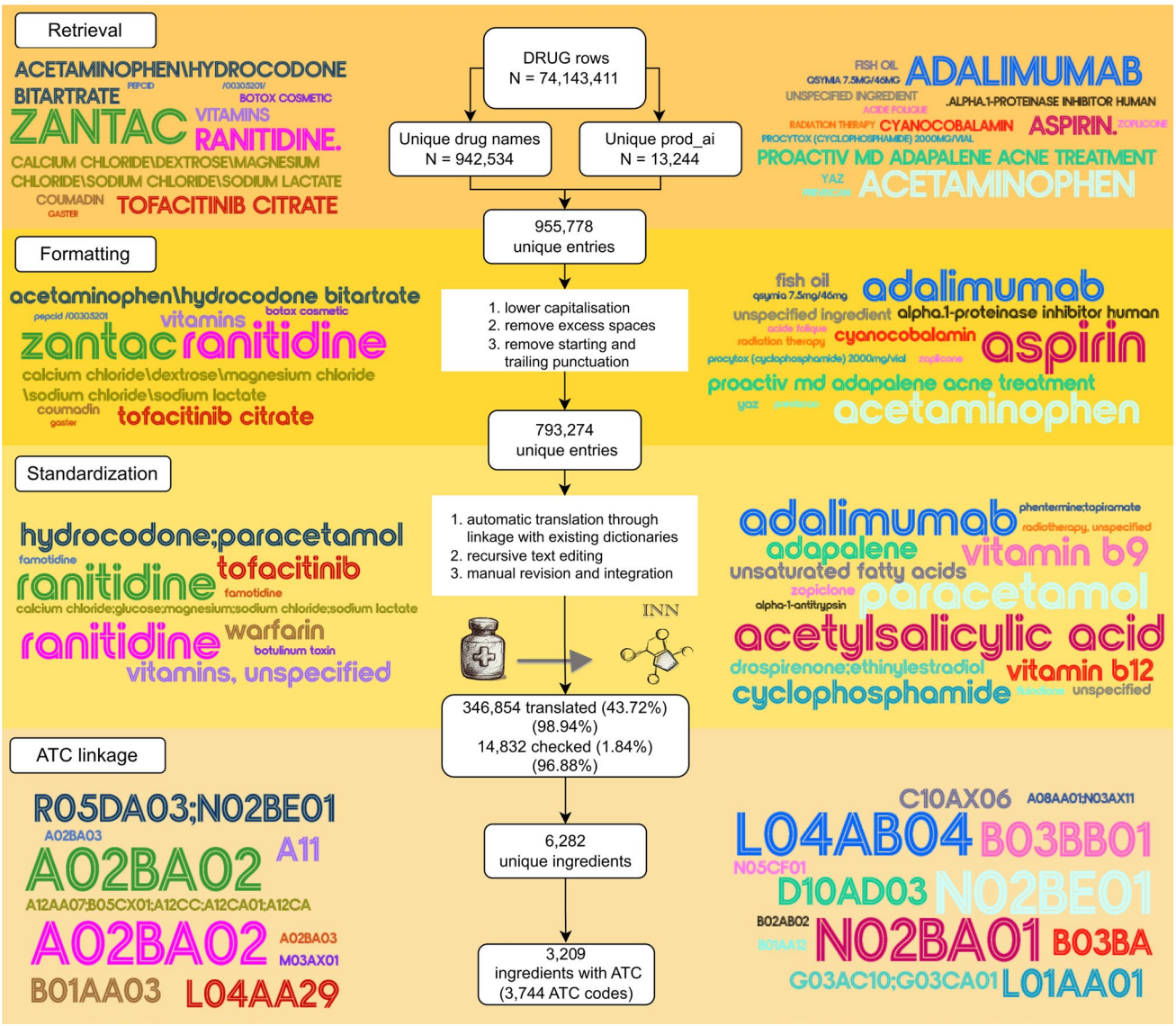

Fig. 2 Translation pipeline results. Flowchart showing the procedure to translate drug names to active ingredients. Some examples of the processing of entries are provided in the background. The color remains constant across the steps, and within each step, the dimension is proportional to the number of occurrences

4 Discussion

4.1 The DiAna Dictionary

The sensitivity of case retrieval and the relevant disproportionality analysis results may vary depending on the drug cleaning procedures used in SRSs. Disproportionality analysis is mostly performed on public dashboards or other analytical tools with no access to underlying data, ready-to-use databases with partial or non-transparent translation, or individually cured undisclosed databases. While these tools provide easy access to disproportionality analysis, they also pose a risk of inappropriate analyses and interpretation due to users' unawareness of the nature of data [9]. Common drug translation procedures involve automatic linkage to existing dictionaries (offering only partial translation) and automatic algorithms dealing with misspellings (potentially introducing errors). While the resulting pre-mapped datasets prove highly valuable for signal detection, for effective signal refinement it is recommended a higher control over the definition of the study object already at the drug name-to-ingredient mapping stage.

To address these concerns, a dictionary for drug name-to-ingredient mapping was developed through an automatic procedure that was manually checked and extended. This dictionary, called the DiAna dictionary, has required a

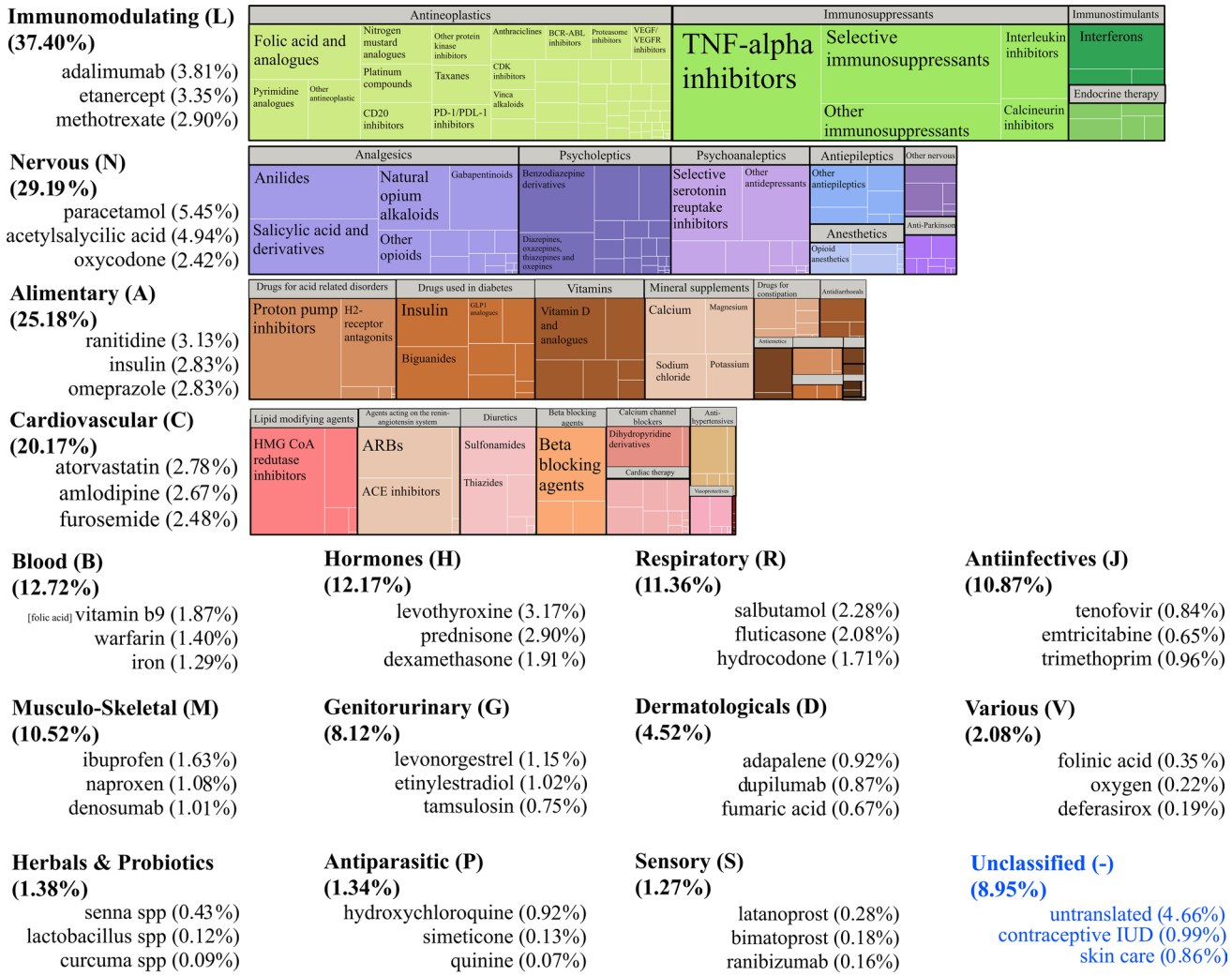

Fig. 3 Distribution of medicinal products in the FAERS. Drugs most frequently reported in the FAERS, after translation, according to ATC class. Each step is a first level, starting from the most reported one. Within each level, a tree map shows how ATC levels 2 and 4 are reported in FAERS reports. The three most reported active ingredients of each 1st level are also shown. *FAERS* US FDA adverse event reporting system, *ATC* anatomical therapeutic chemical, *IUD* intrauterine device, *PD-1* programmed death-1, *PD-L1* programmed death-ligand 1, *VEGF* vascular endothelial growth factor, *VEGFR* vascular endothelial growth factor receptor, *TNF* tumor necrosis factor, *GLP1* glucagon-like peptide 1, *ARBs* angiotensin receptor blockers, *ACE* angiotensin-converting enzymes

time-consuming effort and is made available open source for everyone to use it and propose changes. The use of the DiAna dictionary will allow authors to better define studied drugs, and the pharmacovigilance community to propose more appropriate definitions, contributing to the achievement of an agreement on the best possible drug names-to-ingredient mapping.

The DiAna dictionary is already implemented in a pre-mapped dataset for signal detection accessible through the R DiAna package. The innovative feature and added value of this pre-mapped dataset, compared with previously published attempts of drug name standardizations [7, 16, 17, 17], is its ability to translate almost 99% of drug names reported to the FAERS. The only other dictionary with the same translation percentage was developed by Wong et al. [17], but it was not publicly available. Additionally, the mapping of free text to active ingredient is freely accessible for easy inspection, update, and modification according to the specific research question in signal refinement activities (see Table 2). A greater control on data cleaning and focusing on the definition of the studied drugs, and not only of the studied events, will result in improved replicability and accuracy of signals and more conscious and appropriate interpretation of results, with relevant benefit for the scientific community.

Table 1 Performance of DiAna translation.

Active substance	Formatting (*n* occurrences)	RxNorm (*n* occurrences)	DiAna (*n* occurrences)	DiAna/RxNorm	DiAna/formatting
Paracetamol	112,939	(–)	1,040,051	(–)	9.21
Acetylsalicylic acid	61,652	(–)	942,051	(–)	15.28
Adalimumab	21,403	699,331	727,730	1.04	34.00
Etanercept	19,332	632,130	638,427	1.01	33.02
Levothyroxine	55,115	288,916	604,688	2.09	10.97
Ranitidine	69,874	69,883	597,604	8.55	8.55
Methotrexate	224,467	230,866	553,011	2.40	2.46
Prednisone	177,872	181,092	552,531	3.05	3.11
Omeprazole	132,069	273,775	539,760	1.97	4.09
Insulin	73,619	(–)	539,088	(–)	7.32
Metformin	279,461	320,567	534,234	1.67	1.91
Atorvastatin	210,528	421,702	529,453	1.26	2.51
Calcium	150,507	(–)	513,772	(–)	3.41
Amlodipine	250,232	347,454	508,501	1.46	2.03
Furosemide	99,789	269,463	472,999	1.76	4.74
Oxycodone	98,512	287,818	462,109	1.61	4.69
Salbutamol	40,326	(–)	435,816	(–)	10.81
Pantoprazole	182,322	288,211	421,710	1.46	2.31
Metoprolol	76,981	104,124	420,331	4.04	5.46
Magnesium	68,548	(–)	414,166	(–)	6.04
Fluticasone	12,549	91,204	397,614	4.36	31.68
Lenalidomide	27,201	370,133	380,151	1.03	13.98
Hydrochlorothiazide	72,069	(–)	380,068	(–)	5.27
Gabapentin	77,949	170,801	372,316	2.18	4.78
Dexamethasone	91,353	131,730	365,042	2.77	4.00
Lisinopril	125,376	150,113	361,929	2.41	2.89
Vitamin B9	100	(–)	356,564	(–)	3,565.64
Simvastatin	101,474	163,417	341,174	2.09	3.36
Vitamin D3	142,967	(–)	327,770	(–)	2.29
Hydrocodone	55,669	56,140	325,467	5.80	5.85

Drugs most frequently reported in the FAERS, after DiAna translation, relative to simple formatting and the merging with RxNorm. The number of occurrences in the three translations is reported together with the ratio of occurrences between DiAna and the others. In some cases, differences in the nomenclature resulted in empty cells

FAERS US FDA adverse event reporting system

4.2 Better Retrieval for Higher Sensitivity

We were able to translate 98.94% of total drug entries to 6282 unique active ingredients using the DiAna dictionary, compared with 76.32% using only RxNorm. When considering unique drug entries, we translated 346,854 terms over 793,274 (43.72%). We manually checked the first 14,832 terms (up to 174 occurrences), which were responsible for the translation of 96.88% of total drug entries. We believe that this is a good starting point to share our work with the pharmacovigilance community and enable more participative use and development of the DiAna dictionary. In contrast to the previous work by Wong et al. [17], made on the FAERS up to 2012, we made our dictionary (up to 2022) open source. We chose to design the translation so that a new column is produced with only active ingredients while keeping the original verbatim text in a separate column for more in-depth analyses. We have also decided not to translate to salts as this is rarely taken into account in disproportionality analysis and can lead to confusion about whether the same ingredient with unspecified salt should be considered among cases or non-cases. Instead, we have included the linkage to the ATC classification. In some cases, underspecified drug names were translated to higher ATC classes such as 'antihypertensives, unspecified', as this information can be important for adjusting

Table 2 Comparison between the standardization steps performed to create DiAna and other published versions of the FAERS.

Standardization procedure	Reference nomenclature	Percentage translated (declared)	Updated	Open access dictionary
Wong et al. (2015) [17]	INN	99% (limited to LAERS)	No (last known version, presented in the article, limited to the LAERS and not updated after 2012)	No: In the manuscript, it is specified the standardized database will be made available on request. Nothing is specified for the final drug name-to-ingredient mapping
Banda et al. (2016) [7]	USAN	93%	No (last activity on GitHub repository on January 2020)	No: The code to obtain the automatic mapping is available together with the translated database. The final drug name-to-ingredient mapping is not available.
Khaleel et al. (2022) [8]	USAN	97%	No (last updated September 2021 on the Mendeley Data repository)	No: The code to obtain the automatic mapping is available together with the translated database. The final drug name-to-ingredient mapping is not available.
DiAna	INN	99%	Yes (last update 2023 Q1, 2023 Q2 ongoing)	Yes: The final drug name-to-ingredient mapping and the linkage to the ATC are available. Also available are the database cleaned with its step-by-step documentation and an R package for performing disproportionality analysis and retrieving the drug names coded to the ingredient of interest

Steps followed are descripted together with the authors' declared percentage of drug names standardized. It is also reported whether the dictionary is currently updated and whether it is open access

FAERS US FDA adverse event reporting system, *INN* international nonproprietary name, *USAN* United States adopted name, *LAERS* legacy adverse event reporting system, *Qx* quarter *x, ATC* anatomical therapeutic chemical

the analysis and assessing individual cases. The most frequently observed higher classes to which unspecified drug names were mapped were vitamins, immunoglobulins, and estrogens, appearing in 1.7%, 0.8%, and 0.6% of the FAERS reports, respectively. Retrieving information about estrogen exposure is crucial due to its significance as a risk factor for conditions such as thrombosis. This information, not detected by standardization procedures focusing on active ingredients alone, is essential for conducting a more thorough evaluation of relevant cases.

The DiAna dictionary translates a higher proportion of the database, enabling a higher sensitivity in case retrieval, and a higher number of identified cases. This results in better specificity in the definition of non-cases and higher accuracy in signal detection, leading to earlier and clearer signals, as, in specific products, the number of retrieved reports significantly increased. For example, for rimegepant, the DiAna dictionary identifies 278 times more reports than RxNorm alone.

In addition to identifying active ingredients, the drug name information enabled us to identify reports derived from clinical trials (0.28% of total reports), as they recorded placebo, blinding, or drug codes. This information can help researchers exclude evidence already taken into account in other steps of drug safety characterization from the disproportionality analysis.

Finally, the linkage between the DiAna dictionary and the ATC classification can help in the retrieval of drug classes and visualization. The information on the distribution of drug classes in the database is particularly useful for the design of future disproportionality analyses, as it provides insight into the representativeness of the population chosen for comparison. Over one-third of the database consists of reports with anticancer and immunomodulating drugs. The large contribution from these agents is in line with recent global reporting patterns observed for serious and fatal events [23, 24]. Moreover, the remarkable number of reported cases for paracetamol and acetylsalicylic acid underscores once more the relationship between drug consumption and adverse event reporting [25]. Recent observations, specifically in the context of the extensive rollout of COVID-19 vaccines, have reignited attention to the possibility that this uneven distribution of drugs in the SRS should be considered during study design since it may lead to masking/cloaking bias, thus potentially hiding disproportionality signals [26].

The DiAna dictionary and its linkage to the ATC classification are freely available online for everyone to use

(https://osf.io/zqu89/) and can be corrected and expanded by experts in the field. Changes can be proposed in the GitHub repository (https://github.com/fusarolimichele/DiAna) under the issue DiAna dictionary, and will be periodically validated and integrated into the existing dictionary. This collaborative effort will improve the quality and reproducibility of pharmacovigilance research. The dictionary can be downloaded in Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) and csv formats and can be imported into any data management software, such as R, to automatically translate drug names to active ingredients before conducting analyses. Users can also easily modify the translation of specific terms for their analyses, which is not possible with ready-to-use FAERS databases.

4.3 Higher Control on the Mapping for Signal Refinement

The DiAna dictionary has also already been implemented in the DiAna open-source R package [22], which, together with other functions for disproportionality analysis, allows to import a cleaned and documented version of the FAERS preserving the possibility of adjusting drug-names translation. In particular, the drug names coded to a specific active ingredient of interest can be retrieved with the function 'get_drugnames()' (see ESM Table S2 for an example), and, if deemed necessary, they can be modified.

For example, when investigating systemic reactions to ingredients that can be administered topically or systemically, we may want to exclude the topical formulations from the drug definition. One approach is to consider the variable storing information about route of administration, but these data are often unavailable. For instance, gentamicin might be treated differently depending on whether it is administered systemically, topically as a cream (e.g., rinderon-vg), or in eye/ear drops (e.g., garasone). Similarly, when studying aripiprazole, the long-acting injectable form (e.g., Aristada) might be handled differently. This flexibility is lost in databases that have already mapped ingredients without preserving drug name information.

4.4 Identification of Medicinal Products, Towards Higher Standardization in the Collection and Management of Drug Information

In the future, a higher standardization of drug information could already be achieved in the collection, management, and storage of spontaneous reports, following E2B-R3 recommendations to use the Identification of Medicinal Products (IDMP) system developed by the International Organization for Standardization (ISO)—a set of codes unambiguously identifying not only the active ingredients but also the strength and the route of administration of the product. Nonetheless, the E2B-R3 still allows for a free text field for the name of drugs as reported.

The WHO Vigibase, following E2B-R3 recommendations, is embedded with a tool for drug name standardization, i.e., the WHODrug Dictionary. This dictionary compiles extensive drug information, including information about herbal medicines, links drug names to the Anatomical Therapeutic Chemical (ATC) classification, and automatically deals with misspellings and new entries [15]. Nonetheless, this dictionary is only available upon subscription and therefore it cannot be used for, and linked to, an open-source database aiming for complete transparency. Additionally, even if employing a database where drug names are pre-standardized to active ingredients simplifies the process of defining the object of study, as it only requires grouping the active ingredients of interest, it makes it challenging to recognize that different raw drug names translated to the same active ingredient might vary in their suitability for inclusion in the definition.

4.5 Limitations, Strengths, and Further Goals

The DiAna dictionary is not designed as a static dictionary but as a living one—it will require ongoing efforts to keep up with new drugs and terms. We are recursively extending our translation to reach and maintain a fully checked translation of any entry with over 100 co-occurrences. Users of the DiAna dictionary should be aware of this limitation (which is even more impairing in other pre-mapped datasets), especially with less frequent terms that may not be included in the dictionary. It is recommended that before any signal refinement activity concerning a specific drug, inherent terms are checked in the dictionary and any new translations are shared to integrate into the DiAna dictionary for everyone to benefit in their signal detection and refinement activities. The translation will plausibly never be complete, since some terms are not easily translated (e.g., 'chinese food') and many choices are partly subjective. However, these choices can be defined in agreement with the entire pharmacovigilance community.

The translation of ambiguous terms was also noted as a challenge, especially with over-the-counter cold, cough, and flu agents (multiple ingredients changing over the years). When we were not certain, we used the higher-level term (e.g., 'cough preparations, unspecified'). The lack of expertise in supplements and phytotherapies may have resulted in the dictionary being excessively generic (for example, referring to Plantago spp instead of individual species, and COVID 19 vaccines instead of specific types), and it could benefit from refinement by experts for higher specificity and coverage of entries provided to other spontaneous report databases (CAERS and VAERS are more appropriate to investigate the safety profile of these medicinal products).

For example, mapping herbals using the Medicinal Plant Names Service would significantly improve their standardization (cfr. Medicinal Plant Names Services Portal, Royal Botanic Gardens, Kew; https://mpns.science.kew.org/mpns-portal/).

Since lack of completeness is a known problem in spontaneous reports, and other information is not always available, we implemented sharp-cut operative choices to retrieve active ingredients based only on the drug name. The use of additional columns such as country, year of occurrence, dose, indication, and route of administration could help discriminate between mistranslations when the same drug name may be translated to multiple active ingredients. Moreover, information from the drug name column could be used to impute information into other columns. For example, 'nizoral a-d' is translated to ketoconazole and refers specifically to an anti-dandruff shampoo (i.e., the indication, formulation, and route of administration could be imputed if missing), while 'hypersal' refers to a sodium chloride nebulizer solution, and 'jinarc' refers to a formulation of tolvaptan specifically indicated for autosomal-dominant polycystic kidney disease. By incorporating a drug name-to-product translation feature, for example, referring to the WHO Drug Global or to the IDMP, we could streamline the process of the imputation of structured fields using free text, thereby enhancing the value of the DiAna dictionary.

Highlighting the importance of transparency in drug standardization and drug definition and increasing the sensitivity of case retrieval were our two main goals. Nonetheless, it would be interesting to compare accuracy of disproportionality analysis using different drug name standardization strategies.

Linking INN names to ATC codes was a complex task due to the existence of combination products (e.g., glecaprevir and pibrentasvir), medicinal products with ingredients that do not have an ATC code yet, and experimental substances that are missing even the INN. The linkage will be annually updated according to changes in the ATC classification to preserve its utility.

With the recent advent of Natural Language Processing (NLP) techniques, tailored tools have also been implemented to extract information from free text sources such as medical records, as exemplified by Apache cTAKES [27]. Nonetheless, named entity recognition techniques' accuracy decreases in the lack of context and when having to deal with many possible entities. Given that the FAERS drug name fields often do not provide more than one word and given the high number of active ingredients (i.e., entities), we have chosen to employ existing drug name dictionaries and manual revision as a more sensitive method for translation. Nevertheless, the use of NLP techniques, particularly taking into account multiple variables of the FAERS (e.g., route of administration, dose, indication, country) is a promising endeavor to further improve drug names standardization, particularly for instances in which the ambiguity of a drug name may be solved taking into account additional fields, and to extend the automatic translation to drug names with few occurrences. The subscription based UMC WHODrug Koda service (https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-koda/), for example, is an AI tool that takes into account multiple information and helps drug mapping, even if requiring a manual validation.

5 Conclusion

We offer the DiAna dictionary as an open-source tool for the pharmacovigilance community to standardize drug names in the FAERS database and as a means to improve awareness into the importance of the definition of studied drugs. Its public accessibility, transparency, and flexibility provide a foundation for ongoing improvement and refinement through input from experts in the field. With periodic updates, this living project can drive a common effort toward a more transparent and cleaner shared pre-mapped FAERS database, leading to more replicable and reliable research in signal detection. Moreover, it allows higher control on drug definition for signal refinement activities. These functionalities are already implemented and made available in the DiAna R package for disproportionality analysis, through which we aim to promote collaboration and develop a common pharmacovigilance toolbox, sharing not only pre-processing procedures and cleaned data but also consolidated and innovative analyses functions, pipelines, and knowledge, thus promoting drug safety and improving the accuracy, replicability, reliability, and interpretability of pharmacovigilance studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-023-01391-4.

Acknowledgments Our work would not have been possible without public access to the WHO-ATC code. Nonetheless, the results, discussion, and conclusions of the study are those of the authors alone and in no way represent the position of the WHO Collaborating Centre for Drug Statistics Methodology on this subject.

MF is enrolled in the PhD program in General Medical and Services Sciences, Alma Mater Studiorum—Università di Bologna, which supports his fellowship. VG is supported by EU funds (Programma Operativo Nazionale Italian funds for green and innovative research based on European Structural and Investment Funds). VB is enrolled in the PhD program in Experimental and Clinical Pharmacological Sciences, Università degli Studi di Milano, which supports her fellowship. The authors thank Chiara Ballarin, Margherita Bonaiuti, and Laura Pierantozzi (University of Bologna) for their help in the manual revision of drug name standardization.

Declarations

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Conflict of Interest Michele Fusaroli, Valentina Giunchi, Vera Battini, Stefano Puligheddu, Charles Khouri, Carla Carnovale, Emanuel Raschi, and Elisabetta Poluzzi declare no conflicts of interest specific to this research.

Ethical Approval Not applicable. FAERS spontaneous reports are anonymous and publicly available.

Consent to Participate Not applicable. FAERS spontaneous reports are anonymous and publicly available.

Consent for Publication Not applicable. FAERS spontaneous reports are anonymous and publicly available.

Availability of Data and Material The dictionary and the linkage to the ATC classification are available, together with the cleaned FAERS database, at https://osf.io/zqu89/, and the documentation is available at https://github.com/fusarolimichele/DiAna. The DiAna R package for disproportionality analysis, with the function to retrieve the drug name-to-ingredient mapping can be downloaded at https://github.com/fusarolimichele/DiAna_package.

Code Availability All the processing, analyses, and visualization were obtained through R-software (version 4.2.1). The code for using the dictionary in the cleaning of the FAERS database is available at https://github.com/fusarolimichele/DiAna.

Author Contributions MF, VG, and VB conceptualized and designed the study and developed the methodology. MF and VG implemented the automatic translation, and MF, VG, VB, and SP manually checked the automatic translation. MF and VG performed the visualization, and wrote the original draft of the manuscript. All authors contributed significantly to interpretation of the data, critically revised the work, and approved the final version.

References

1. Poluzzi E, Raschi E, Piccinni C, Ponti FD. Data mining techniques in pharmacovigilance: analysis of the publicly accessible FDA adverse event reporting system (AERS). In: Data mining applications in engineering and medicine. InTech; 2012.
2. Raschi E, Moretti U, Salvo F, Pariente A, Antonazzo IC, Ponti FD, et al. Evolving roles of spontaneous reporting systems to assess and monitor drug safety. pharmacovigilance. 2018. https://www.intechopen.com/online-first/evolving-roles-of-spontaneous-reporting-systems-to-assess-and-monitor-drug-safety. Cited 3 Feb 2019.
3. ICH Official web site : ICH [cited 2023 Oct 17]. Available at: https://ich.org/page/e2br3-individual-case-safety-report-icsr-specification-and-related-files. Cited 17 Oct 2019.
4. Fusaroli M, Salvo F, Bernardeau C, Idris M, Dolladille C, Pariente A, et al. Mapping strategies to assess and increase the validity of published disproportionality signals: a meta-research study. Drug Saf. 2023. https://doi.org/10.1007/s40264-023-01329-w.
5. FDA. FDA adverse event reporting system (FAERS) Public Dashboard | FDA [].https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard. cited 14 Dec 2022.
6. Center for Drug Evaluation and Research. FDA adverse event reporting system—latest quarterly data files. FDA. 2019 http://www.fda.gov/drugs/fda-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files. cited 28 Jul 2019.
7. 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: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4872271/. cited 17 Dec 2020.
8. Khaleel MA, Khan AH, Ghadzi SMS, Adnan AS, Abdallah QM. A Standardized dataset of a spontaneous adverse event reporting system. Healthcare. 2022;10:420.
9. Giunchi V, Fusaroli M, Hauben M, Raschi E, Poluzzi E. Challenges and opportunities in accessing and analysing FAERS data: a call towards a collaborative approach. Drug Saf. 2023;46:921–6.
10. Hauben M, Reich L, Gerrits CM, Younus M. Illusions of objectivity and a recommendation for reporting data mining results. Eur J Clin Pharmacol. 2007;63:517–21.
11. Mouffak A, Lepelley M, Revol B, Bernardeau C, Salvo F, Pariente A, et al. High prevalence of spin was found in pharmacovigilance studies using disproportionality analyses to detect safety signals: a meta-epidemiological study. J Clin Epidemiol. 2021;138:73–9.
12. Rocca E, Grundmark B. Monitoring the safety of medicines and vaccines in times of pandemic: practical, conceptual, and ethical challenges in pharmacovigilance [special issue]. Argumenta. 2021;7:127–46.
13. Leonelli S. The challenges of big data biology. Elife. 2019;8:e47381.
14. Wisniewski AFZ, Bate A, Bousquet C, Brueckner A, Candore G, Juhlin K, et al. Good signal detection practices: evidence from IMI PROTECT. Drug Saf. 2016;39:469–90.
15. Lagerlund O, Strese S, Fladvad M, Lindquist M. WHODrug: a global, validated and updated dictionary for medicinal information. Ther Innov Regul Sci. 2020;54:1116–22.
16. Stanford T. The fuzzyfaers package. 2022: https://github.com/tystan/fuzzyfaers. Cited 24 Dec 2022.
17. Wong CK, Ho SS, Saini B, Hibbs DE, Fois RA. Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. Pharmacoepidemiol Drug Saf. 2015;24:731–7.
18. WHOCC - ATC/DDD Index https://www.whocc.no/atc_ddd_index/. Cited 9 May 2023.
19. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.
20. Drugs.com|Prescription drug information, interactions & side effects. Drugs.com. https://www.drugs.com/. Cited 24 Dec 2022.
21. WHO Collaborating Centre for Drug Statistics Methodology. Guidelines for ATC classification and DDD assignment 2023 https://www.whocc.no/atc_ddd_index_and_guidelines/guidelines/. Cited 9 May 2023.
22. Fusaroli M, Giunchi V. DiAna R package: advanced disproportionality analysis in the FAERS for drug safety. 2023 https://github.com/fusarolimichele/DiAna_package. Cited 21 Oct 2023.
23. Sonawane KB, Cheng N, Hansen RA. Serious adverse drug events reported to the FDA: analysis of the FDA adverse event

reporting system 2006–2014 database. J Manag Care Spec Pharm. 2018;24:682–90.
24. Montastruc J-L, Lafaurie M, de Canecaude C, Durrieu G, Sommet A, Montastruc F, et al. Fatal adverse drug reactions: a worldwide perspective in the World Health Organization pharmacovigilance database. Br J Clin Pharmacol. 2021;87:4334–40.
25. Orhon P, Robert M, Morand T, Cracowski J-L, Khouri C. Investigating the link between drug consumption and adverse events reporting in France. Fundam Clin Pharmacol. 2023;37:879–82.
26. Harpaz R, DuMouchel W, Van Manen R, Nip A, Bright S, Szarfman A, et al. Signaling COVID-19 vaccine adverse events. Drug Saf. 2022;45:765–80.
27. Apache cTAKES™—Clinical text analysis knowledge extraction system https://ctakes.apache.org/. Cited 20 Oct 2023.