TITLE PAGE

Title:

A Systematic Approach to “Cleaning” of Drug Name Records Data in the FAERS Database: A Case Report

Michael A. Veronin, B.S.Pharm., M.S., Ph.D.* * Corresponding author:

Associate Professor

The University of Texas at Tyler

Ben and Maytee Fisch College of Pharmacy

Department of Pharmaceutical Sciences

3900 University Blvd.

Tyler, Texas 75799

Tel: (903) 566-6148

Fax: (903) 565-5598

E-mail: mveronin@uttyler.edu

Robert P. Schumaker, Ph.D.

Professor and Director of the Data Analytics Laboratory

The University of Texas at Tyler

Soules College of Business

Department of Computer Science

3900 University Blvd.

Tyler, Texas 75799

Tel: (903) 565-5578

Fax: (903) 565-5598

E-mail: rschumaker@uttyler.edu

Rohit R. Dixit, M.S.

The University of Texas at Tyler

Soules College of Business

Department of Computer Science

3900 University Blvd.

Tyler, Texas 75799

Tel: (903) 566-7403

E-mail: rohitdixit188@live.in

Pooja Dhake, M.S.
The University of Texas at Tyler
Soules College of Business
Department of Computer Science
3900 University Blvd.
Tyler, Texas 75799
Tel: (903) 566-7403
E-mail: pdhake@patriots.uttyler.edu

Morgan Ogwo, Pharm.D.
The University of Texas at Tyler
Ben and Maytee Fisch College of Pharmacy
Department of Pharmaceutical Sciences
3900 University Blvd.
Tyler, Texas 75799
Tel: (903) 566-6148
E-mail: mogwo@patriots.uttyler.edu

Abstract

Introduction: Data "cleaning," also known as data "cleansing," or data "curation" is about identifying and rectifying errors in data. The objective of this report is to present a data cleaning and standardization process for the drug name files in the U.S. Food and Drug Administration Adverse Event Reporting System database, FAERS. *Methods:* Drug name data was cleaned and standardized using a combination of data cleaning tools and manual correction techniques. Data files were organized into frequency intervals and a strategy of cleaning using iteration and programming scripts in the MySQL Workbench was employed. *Results:* The download of the FAERS quarterly reports for the time periods ranging from Q1 2004 to Q3 2016 resulted in 32,736,657 DRUG file records. Records contained a variety of errors, such as misspellings, abbreviations, and nondescript or ambiguous names. Upon completion of the process, standardization of greater than 95% of the drug name data in the FAERS database was achieved. *Conclusions:* With large datasets such as FAERS, a cleaning process is necessary to rectify data that may be incomplete or inaccurate due to input errors, in order to improve the quality and validity of information.

Keywords: FAERS; U.S. Food and Drug Administration; FDA; adverse drug event; drug safety; data cleaning

Introduction

Data "cleaning," also known as data "cleansing," or data "curation" is about identifying and rectifying errors in the data in order to improve the data quality and validity of information (Rahm and Do, 2000; Poluzzi et al., 2012). Problems in data quality may arise in data files in databases for a number of reasons, often as a result of human error. Not uncommonly, spelling errors may occur during data entry, or simply information may be omitted, or inaccurately transcribed. Whenever a system increases in complexity, such as when several sources of data need to be accessed and integrated in data repositories or warehouses, an increase in input errors may arise, and the need for data cleaning is also likely to increase. For researchers and clinicians to have access to accurate data, a consistent strategy for correction of data errors or aberrations is warranted.

The U.S. Food and Drug Administration maintains one of the largest government databases in the country, known as the FDA Adverse Event Reporting System. Abbreviated as FAERS, it is comprised of reports on adverse event and medication error reports that have been submitted to the FDA (U.S. Food and Drug Administration, 2018a). The FAERS database may be viewed as a "large dataset," in that it is essentially "a large collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity" (TechTarget Network, 2016).

Specifically, FAERS is an electronic repository of information that supports the FDA's post-marketing safety surveillance program for all approved drug and therapeutic biologic products. Records are created in the FAERS database from reports of adverse events and medication errors by healthcare professionals, consumers, and drug manufacturers through the "MedWatch" reporting program (U.S. Food and Drug Administration, 2018a). The majority of

Medwatch reports (between 70% and 75%) are submitted by healthcare professionals (Ahmad et al., 2005; Toki & Ono, 2018).

Although the precise number of entries in the FAERS database is difficult to determine at any given time, it is estimated that 700,000 adverse event reports across a variety of therapeutic categories are documented in FAERS annually (Hoffman et al., 2013; U.S. Food and Drug Administration, 2018b; U.S. Food and Drug Administration, 2018c).

The FAERS database is structured as a computerized, relational database consisting of seven file "packages" organized in tables (Poluzzi et al., 2012). Specific items of information are contained within the file packages and are linkable to other data files within the database.

The files by name are: the DEMO file (demographic characteristics), which includes information on "event date", patient "age" and "gender", "reporter country" and "reporter's type of occupation"; REACTION file, including all adverse drug reactions; OUTCOME file (type of outcome, such as death, life-threatening, hospitalization); RPSR file, with information on the source of the reports (i.e. company, literature); THERAPY file, containing drug therapy start dates and end dates for the reported drugs; INDICATIONS file, containing terms coded for the indications of use (diagnoses) for the reported drugs; and the DRUG file contains the DrugName data set, consisting of the names of the reported medications, which are suspect for Adverse Drug Events (ADEs).

In a discussion on large databases, Garcia et al (2016) asserts that "In most of current real-life problems, there is a potential for incomplete data. Because of either human or machine failure, input data can present some gaps or errors" (García et al., 2016, p15). As healthcare professionals, consumers, and manufacturers can each report an adverse event using the

MedWatch system, the FDA does not provide any corrections or checks for the input data (Marks, 2017).

Although researchers have expressed the need to have clean data in order to proceed with research efforts, often there is not an emphasis on data cleaning prior to analysis for Adverse Drug Events (ADEs). As asserted by Banda et al (2016), "The publicly available US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) data requires substantial curation before they can be used appropriately, and applying different strategies for data cleaning and normalization can have material impact on analysis results" (Banda et al., 2016, p1).

Researchers on drug safety may come to realize that data cleaning prior to using the data from FAERS for further analysis can involve a major commitment of time and effort. Various approaches of standardizing the publicly available FAERS database via cleaning have been described by several authors (Wang et al., 2014; Böhm et al., 2016; Grigoriev et al., 2014; Wong et al., 2015). Descriptions vary among authors, yet a common thread among these reports is that when erroneous and invalid types of errors are identified, remediation involves a combination of automated and manual approaches to cleaning.

The focus of this report is on specific files within the FAERS database involving errors and inconsistencies in reporting of drug names. We present a labor-intensive, yet thorough process that results in a significant number of drug name records cleaned in the DrugName file in the FAERS database, from 2004 to 2016. Our objective was to achieve the maximum number of cleaned drug name records that can serve as a preliminary data set for future studies in adverse drug event research, and perhaps also serve as a model of data cleaning for other investigators.

Methods

A framework for data cleaning has been described by Maletic and Marcus (2000), which involves the following steps:

- Define and determine error types;
- Search and identify error instances;
- Correct the errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

This framework provided guidance for the cleaning process, yet, to meet the unique challenges of this project, modifications of the initial steps were necessary (Maletic and Marcus, 2000).

With such a large, complex dataset, it was necessary to devise a process of cleaning that would provide a balance of efficiency and effectiveness along with practicality in terms of resources, particularly time and labor. As the case with projects where accurate data is required, the cleaning process should not consume inordinate amounts of time and effort, which would detract from the larger goal of research productivity for which the cleaned data would be utilized. To accommodate such a large dataset of entries in the FAERS database, we determined that the first two steps would have to be modified. Rather than review each data entry as suggested in the first two steps, consistent formatting was applied to all drug name entries. Our assumption was that all entries required standardization of some sort. This not only allowed us to arrive at a consistent format for drug names, but also streamlined the overall cleaning process. As illustrated in Figure 1, the FAERS data was cleaned and standardized using a combination of data cleaning tools and manual correction techniques.

Facilities

Analysis of the FAERS database was conducted in the Data Analytics Laboratory in the Computer Science Department of our academic institution. The laboratory is designed for big data analysis with several high-end servers and software that includes the MySQL Workbench for database work, Eclipse Neon for programming, R Studio for machine learning, and Tableau for data visualization.

Original Download

Since 2004, raw data extracted from the FAERS database can be downloaded from the FDA website in ASCII format (U.S. Food and Drug Administration (2018d). All available raw data for a specified time period was targeted, and consequently no inclusion or exclusion criteria for the data was specified.

The FAERS DRUG data files in ASCII format were accessed on December 4, 2016 at https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/ucm082193.htm for Q4 2012 through Q3 2016 and https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm083765.htm for Q1 2004 through Q3 2012. The maximum number of available records containing drug names in the files were obtained.

The downloaded DRUG file raw data was imported into the relational database management system, MySQL, version 6.0 (Oracle Corporation, Redwood Shores, CA). From this, a custom DrugName table was created, (the data labelled as Drugs.Drugname), which was an exact replica of the FAERS DRUG file. All procedures and processes were performed on this custom table.

Frequency Intervals

To organize the large set of drug name data, frequency intervals of drug name counts were created using the “Group by” reserved keyword in MySQL. With the grouped frequency intervals, it was determined how often each drug name value occurred within each group interval relative to the entire data set (Ali, 2017). From the initial count of drug names, twelve frequency intervals were formed (Table 1), and the frequency of the number of drug name data entries within each interval was observed.

Iteration and Drug Name Identification

To initiate the process of drug name cleaning, a strategy involving iteration was employed. Beginning with the interval with the highest drug name frequency count, each frequency interval was analyzed sequentially, and iterations of the cleaning process were conducted within each interval. In this way, manageable “chunks” of data were able to be cleaned and standardized.

To begin the iterative process, we first targeted only those drugs whose grouped count was equal to or greater than 1000. Consequently, the frequency of drug names that occurred 1000 times or more in the FAERS database was determined. As presented in Table 1, the cleaning process began with interval “>=1000” occurred up to the interval “0 to 29.”

Inspection, Programming Scripts, and Manual Correction

As drug names are often expressed in different ways, including brand name, generic, or combination of both, it was necessary to identify a standard drug name format for consistency. Standard journal style is to use non-branded generic names for medications (Citrome, 2016). For this study, the adopted standard drug name format included the generic name in lower case and not capitalized. All drug names found were transformed and standardized into this format.

Initial observations of the raw drug name data revealed that drug names existed in a variety of forms, which were potentially problematic for data cleaning. For instance, the drug name data records contained null values, ambiguous or nonspecific terms, misspellings, upper and lowercase letters or both, leading and trailing whitespace, new-line and tab characters, leading numbers, special characters, null values, drug name combinations with no delineation of entities, abbreviations, and nonspecific or ambiguous names. This information is illustrated in Table 2.

In order to help rectify variations or errors in drug names and assist in the correction process, programming scripts were created in the MySQL Workbench. Drug names from the original data files (i.e. "raw" data) were manually imported into the scripts. The scripts were designed to identify aberrations in drug name entries and allow for manually renaming them to the correct name.

Drug names encountered in the cleaning process included generic and brand names or combination product names. Infrequently, a drug's scientific or chemical, or investigational drug name was encountered. When this occurred, the scientific/chemical/investigational drug nomenclature was converted to the generic drug name.

In addition, the script allowed for identification of combination drugs, and separation into their individual components. Each component of a combination drug name was counted as individual drug name entities.

References were used to ensure correct identification of drug names, which included Drugs.com (Drugs.com, 2018a), Micromedex (IBM Micromedex Clinical Knowledge Suite, 2018), and the Drug Information Portal from the U.S. National Library of Medicine. (U.S. National Library of Medicine, 2018). Where needed, an automatic check for correct names of drugs was employed: a proprietary spell-checker is provided by Drugs.com, and Micromedex

Solutions intelligent searching provides in-line spelling suggestions in real time while conducting searches (Drugs.com, 2018b; Searching Micromedex Solutions, 2018).

In addition to standard drug references, drug names were verified and classified in therapeutic categories by their active ingredient according to the Anatomical Therapeutic Chemical (ATC) Classification System. In the ATC system, drugs are divided into groups according to the target organ, chemical and therapeutic characteristics, and mechanism of action (WHO Collaborating Centre for Drugs Statistics Methodology, 2016).

The overall data management process was coordinated by a data scientist (RS), manual and automated corrections to drug names were verified by graduate assistants (RD, PD, MO) and accuracy of drug name information verified by a consultant pharmacist (MV). All phases of the cleaning/standardization process were reviewed by all investigators, and any discrepancies were resolved through discussion until consensus on findings was reached.

Results

The download of the FAERS quarterly reports for the time periods ranging from Q1 2004 to Q3 2016 resulted in 32,736,657 DRUG file records from which drug names were imported into the DrugName table in MySQL. This is presented in Table 1.

Nondescript Drug Names

The initial stages of cleaning involved removal of “nondescript” drug names. In this case, the drug name file contained records of drug names that could not be corrected because a specific drug was not identified. In several instances, it was observed throughout the entire dataset that the keyword “NULL” appeared, indicating missing or unknown values. In addition, several entries were represented as nonspecific or ambiguous drug names, such as “antibiotic,” “analgesic,” or “painkiller.” As indicated in Table 1, these types made up 2.49% of the original data set and were removed [32736657 – 31921755 = 814, 902 entries], leaving only legitimate drug name entries for cleaning.

Frequency Counts

Results from the drug name frequency intervals and counts are presented in Table 1. The Group Query function in MySQL for Drug names ≥1000 produced an individual count for cleaned drug names equal to 1444. This means that 1444 drug names have a frequency count in the FAERS database of 1000 times or more. This first frequency interval produced the sum of 30,849,955 drug name entries, or 96.64% of the total amount of data records.

The frequency data for drug names in each interval was ranked from highest (i.e., most frequent in number) to lowest (i.e. least frequent in number.) For illustrative purposes, the top ten drugs ranked from this subset are shown in Table 3.

As shown in Table 3, the most frequently appearing drug name was aspirin, which occurred 493,379 times, followed by Adalimumab (Humira®), which occurred 492,173 times. The top ten drugs by frequency made up over 10 percent of the cleaned drug name records in the FAERS database. As the case with all drug names, the initial count of the top ten drugs by frequency included both valid and invalid (pre-cleaned) drug names.

Data Cleaning

Of the 1444 drug name subset with 30,849,955 entries, each drug name entry passed through a series of automated/manual cleaning steps. The process involved standardizing the data by converting all drug names to uppercase, removing punctuation/spaces, acronym expansion, correlating drug name information against Drugs.com and Micromedex Solutions (databases of current drug names), and then removing drugs in the corpus that were nondescript. The steps employed are summarized in Figure 2.

When the interval of ≥1000 was completed, the next interval, 900-999, and all subsequent intervals were processed in this manner. Upon completion of this iterative process, standardization of greater than 95% of the drug name data in the FAERS database was achieved. The interval of 0 to 29 containing 533235 records (approximately 1.67%) was not included in the data cleaning process due to manpower and time constraints.

Discussion

A recent survey of data scientists ascertained that data preparation accounts for about 80% of the work of data scientists, and cleaning data is the least enjoyable and most time-consuming data science task (L.V., 2006). In this paper, we describe a pragmatic approach to cleaning one of the U.S government's largest healthcare databases, FAERS.

It is reasonable to assume that no single method will be sufficient to identify all errors in a dataset, and so we felt it important to use a combination of methods that essentially best restore the data to its originally intended format, free from errors, or other ambiguous formats. Our expectation was somewhat met, in that with limited resources, greater than 95% cleaning and standardization of the retrieved drug names in the Drug File of the FAERS database was achieved. *To date, this yielded the largest number of drug name records in the FAERS database to be amended to a standardized, practicable format.*

Limitations

The process to rectify errors in the DrugName subset of the FAERS database was fairly prescriptive, yet there was a degree of subjectivity by data extractors in the standardization process, but in our assessment, this had little impact on the overall results.

The drug name variants that did not identify a specific drug were not subjected to the cleaning and standardization process. For instance, terms such as "pain killer," "analgesic," or "antibiotic" may have been entered, but are not recognized as specific drugs. When this occurred, these drug names were not included in the cleaning and standardization process.

Issues with Quantification of Drug Name Variations

In this report, it is tempting to quantify the erroneous drug name entries and report as a proportion of the database as a whole, (i.e. distinguish between entries that are "correct" and not needed to be cleaned, and entries considered "erroneous" that needed to be modified), but the conclusions drawn may be misleading.

For instance, consider the inflammatory-blocking agent adalimumab (brand name Humira®). An actual entry into the database was expressed as "Humira (Abbott)". For the purposes of communicating a potential problem to the FDA via the MedWatch program, this entry would not be incorrect. That is, there was no miscommunication or misinterpretation of the identity of the drug to the FDA.

However, in the context of data cleaning, the drug name in not expressed in a format that is consistent with other drug names, and is problematic for data analysis purposes. In our data cleaning process, "Humira (Abbott)" was converted to HUMIRA, then the generic name, and the subsequent steps of the cleaning and standardization process were followed.

In short, the communication via MedWatch was correct, but for the sake of data analysis, the format of the drug name was not correct and required correction. In our view, it would be undue to label the entry "Humira (Abbott)" and other similar entries as incorrect, and quantify these types of entries compared to the total number of entries in the database.

Based on the number of scripts used to correct drug names, we can estimate that the drug name variations as illustrated in Table 2 accounted for no more than than 10% of the total number of entries, which is somewhat consistent with previous reporting on MedWatch entry errors, in general (Getz et al., 2014).

Specific determination of the number of incorrect entries compared to correct entries in the database is beyond the scope of this study and would warrant further investigation.

Conclusions

As often the case with large datasets, a cleaning process is necessary to detect incomplete, inaccurate, or inappropriate data and then improving the quality through correction. We describe a novel method of cleaning and standardizing drug names and correction of drug name deviations. Even though this method is time-consuming - which may limit practicality - it is detailed and complete. Moreover, this method does not result in the loss of original downloaded data from the FDA's website. In addition to methods for improving error detection and cleaning, future research should include error prevention, as it is more efficient to prevent errors than to attempt to identify and correct them after the fact.

Implications

The stepwise approach that was taken in this project for data cleansing is much more than just updating records with "good" data. Our view of data cleaning is best expressed by Maletic and Marcus (2000), where data cleaning involved "decomposing and reassembling" the data and adds value to the resultant information.

In this project, we have sought a cleansed version of the DrugName subset of the FAERS database, using common, yet unique methods. As a result of this project, perhaps we have increased awareness of the potential to improve research for data mining and statistical methods researchers.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interests with respect to the research, authorship, and/or publication of this article.

Funding

The authors have no support or funding to report.

References

Ahmad S, Goetch R, Marks N. (2005). Spontaneous reporting in the United States. In B.L Strom (Ed.), Pharmacoepidemiology. 4th ed. (p 135). West Sussex, England: John Wiley & Sons.

Ali L. (2017). How to create a grouped frequency table. Sciencing. April 24, 2017. Available at: http://sciencing.com/create-grouped-frequency-table-5531910 html. (accessed 12 August 2019).

Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. Scientific Data 3:160026.

Böhm R, von Hehn L, Herdegen T, Klein HJ, Bruhn O, Petri H, Höcker J. (2016). OpenVigil FDA–Inspection of US American adverse drug events pharmacovigilance data and novel clinical applications. PloS one 11(6): e0157753.

Citrome L. (2016). What's in a name? Use of brand vs. generic drug names. International Journal of Clinical Practice 70(1):3-4.

Drugs.com. (2018a). Available at: www.drugs.com/. Updated November 1, 2018. (accessed 12 August 2019).

Drugs.com. (2018b). Contact Drugs.com. Corporate inquiry 181200-9009012003. Available at: www.drugs.com/support/contact.html (accessed 20 August 2018).

García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics (1)1:9.

Getz, K. A., Stergiopoulos, S., Kaitin, K. I. (2014), Evaluating the completeness and accuracy of MedWatch data. American Journal of Therapeutics 21(6): 442-446.

Grigoriev I, zu Castell W, Tsvetkov P, Antonov AV. (2014). AERS spider: an online interactive tool to mine statistical associations in adverse event reporting system. Pharmacoepidemiology and Drug Safety 23(8): 795-801.

Hoffman KB, Overstreet BM, P. Doraiswamy PM. (2013). Development of a drug safety ePlatform for physicians, pharmacists, and consumers based on post-marketing adverse events. Drugs and Therapy Studies (3)1:4.

L.V. (2016). Least enjoyable and most time-consuming data science tasks. In: Data Science Central. Available at: https://www.datasciencecentral.com/profiles/blogs/bubu. (accessed 12 August 2019).

Maletic JI, Marcus A. (2000). Data cleansing: beyond integrity analysis. Proceedings of the 2000 Conference on Information Quality: 200-209.

Marks, N. S. (2017), MedWatch: safety information and adverse event reporting. 11 July 2017. Available at: www.medscape.org/viewarticle/588757. (accessed 12 August 2019).

Micromedex Clinical Knowledge Suite. Available at: http://truvenhealth.com/Products/Micromedex/Product-Suites/Clinical-Knowledge (accessed 20 October 2018).

Poluzzi, E., Raschi, E., Piccinni, C., and de Ponti, F. (2012). Data mining techniques in pharmacovigilance: analysis of the publicly accessible FDA Adverse Event Reporting System (AERS). In Karahoca, A., (Ed.), Data mining applications in engineering and medicine (pp267-301). Croatia: InTech. Available at: https://mts.intechopen.com/books/data-mining-applications-in-engineering-and-medicine/data-mining-techniques-in-pharmacovigilance-analysis-of-the-publicly-accessible-fda-adverse-event-re (accessed 12 August 2019).

Rahm, E., Do, H. D. (2000). Data cleaning: problems and current approaches. *IEEE Data Engineering Bulletin*. 23(4), 3-13. Available at: www.betterevaluation.org/sites/default/files/data_cleaning.pdf (accessed 12 August 2019).

Searching Micromedex Solutions. (2018). Available at: www.micromedexsolutions.com/micromedex2/4.85.0/WebHelp/Home_Page/Searching/Searching.htm (accessed 12 August 2019).

TechTarget Network. (2016). Whatis.com. data set. Available at: http://whatis.techtarget.com/definition/data-set (accessed 12 August 2019.)

Toki T, Ono S. Spontaneous Reporting on Adverse Events by Consumers in the United States: An Analysis of the Food and Drug Administration Adverse Event Reporting System

Database. Drugs Real World Outcomes. 2018 Jun; 5(2): 117–128. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5984610/ (accessed 12 August 2019).

U.S. Food and Drug Administration. (2018a). Questions and answers on FDA's adverse event reporting system (FAERS). What is FAERS? Available at: https://www.fda.gov/drugs/surveillance/fda-adverse-event-reporting-system-faers (accessed 12 August 2019).

U.S. Food and Drug Administration. (2018b). Reports received and reports entered into AERS by year. Available at: www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm (accessed 12 August 2019).

U.S. Food and Drug Administration. (2018c). MedWatch: The FDA safety information and adverse event reporting program. Available at: www.fda.gov/Safety/MedWatch/default.htm. (accessed 12 August 2019).

U.S. Food and Drug Administration (2018d). FDA adverse event reporting system (FAERS): Latest quarterly data files. Available at: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm (accessed 12 August 2019).

U.S. National Library of Medicine. Drug information portal. Available at: https://druginfo.nlm.nih.gov/drugportal/. Updated November 2018. (accessed 12 August 2019).

Wang L, Jiang G, Li D, Liu H. (2014). Standardizing adverse drug event reporting data. Journal of Biomedical Semantics 5(1): 36.

World Health Organization. Essential medicines and health products. 2. Anatomical Therapeutic Chemical (ATC) Classification. 2019. Available at: https://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/ (accessed 12 August 2019).

Wong CK, Ho SS, Saini B, Hibbs DE, Fois RA. (2015). Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. Pharmacoepidemiology and Drug Safety 24(7): 731-737.

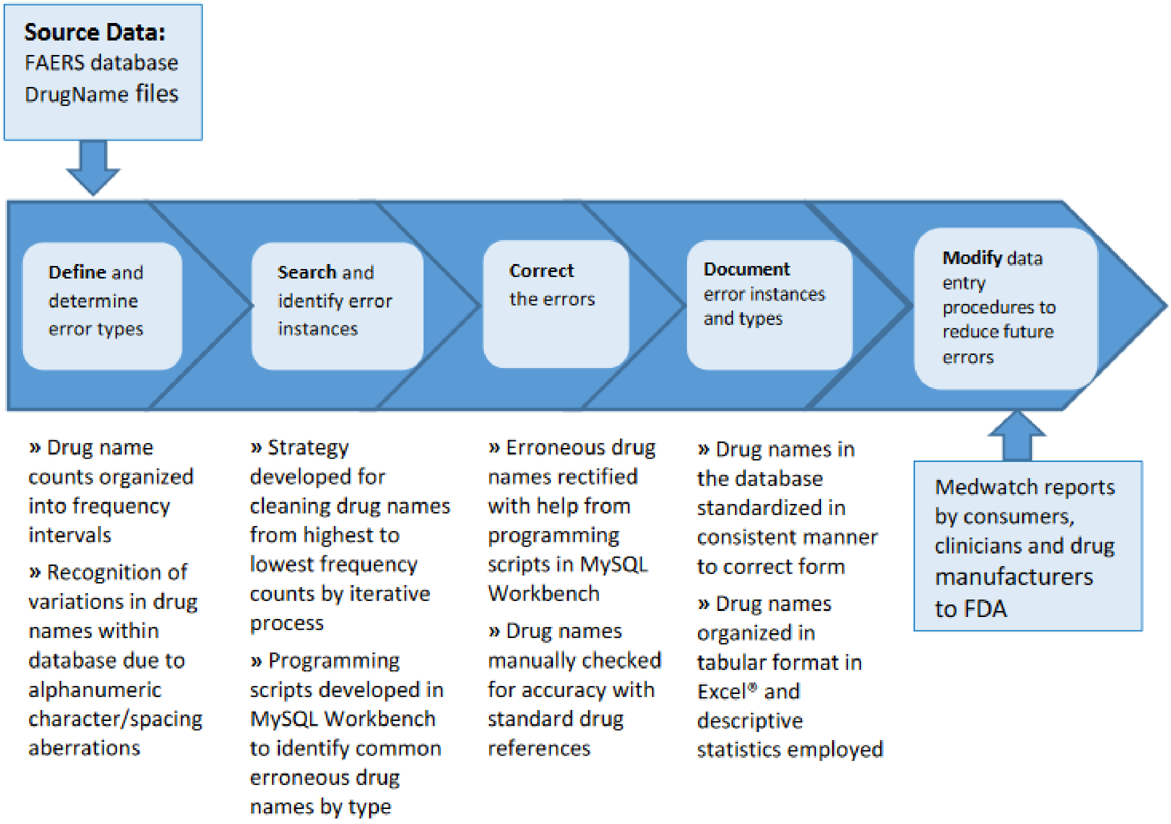

Source Data:
FAERS database
DrugName files
Define and determine error types
Search and identify error instances
Correct the errors
Document error instances and types
Modify data entry procedures to reduce future errors
Medwatch reports by consumers, clinicians and drug manufacturers to FDA
» Drug name counts organized into frequency intervals
» Recognition of variations in drug names within database due to alphanumeric character/spacing aberrations
» Strategy developed for cleaning drug names from highest to lowest frequency counts by iterative process
» Programming scripts developed in MySQL Workbench to identify common erroneous drug names by type
» Erroneous drug names rectified with help from programming scripts in MySQL Workbench
» Drug names manually checked for accuracy with standard drug references
» Drug names in the database standardized in consistent manner to correct form
» Drug names organized in tabular format in Excel® and descriptive statistics employed

Figure 1. Framework for Data Cleaning of the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) Database. (Adapted from Maletic and Marcus, 2000)

1. Obtain DRUG file records from FAERS database
    a. Import drug names into MySQL

2. Remove records with "nondescript" drug names (e.g., "painkiller", "antibiotic")

3. Develop frequency intervals to determine frequency counts

4. Data cleaning/standardization of drug names

    a. Convert drug names to generic form using Drugs.com, Micromedex (Truven Health Analytics), and the Drug Information Portal from the U.S. National Library of Medicine

    b. Convert drug names to uppercase

    c. Acronym expansion, remove punctuation/spaces, expand drug name abbreviations (e.g.; VIT B12 → VITAMIN B12)

    d. Remove leading and trailing whitespace, newline and tab characters, leading numbers, special characters and null values

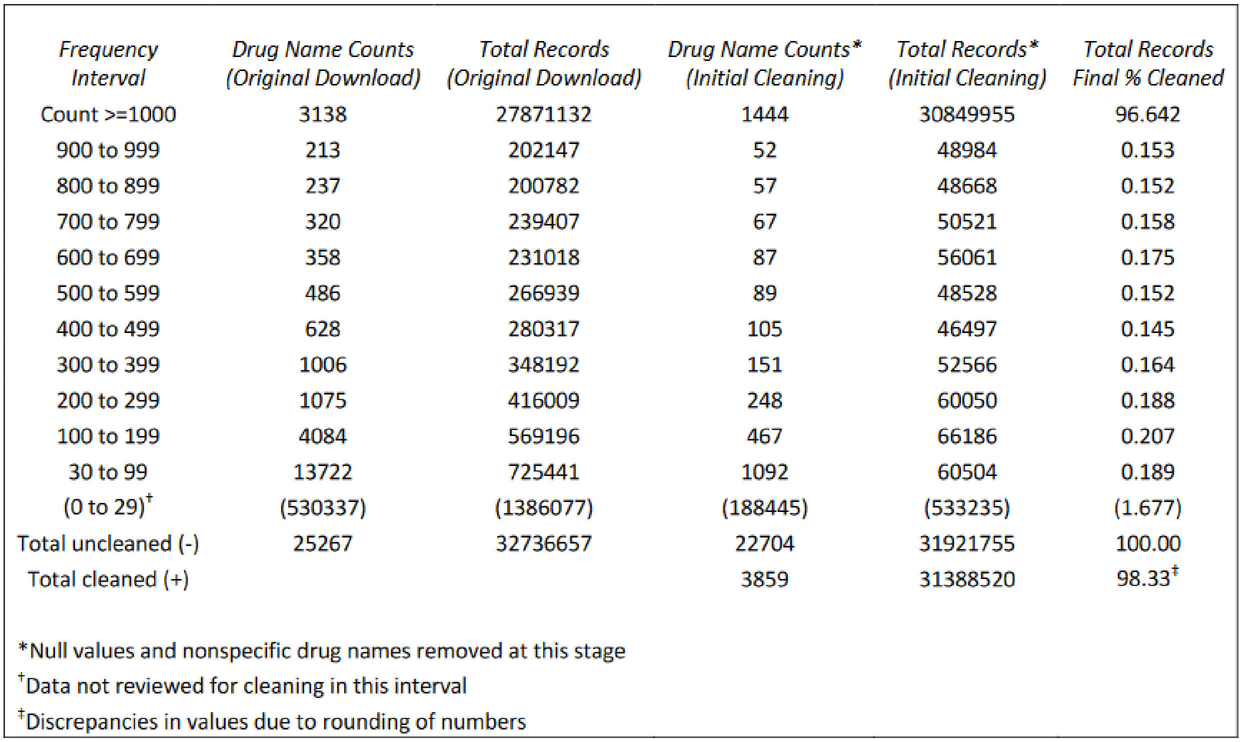
    e. Partition individual records with multiple drug names into separate records

5. Remove record of any drug that appears less than 30 times in frequency count

6. Continued check of drug name accuracy with standard drug references (Drugs.com, Micromedex, Drug Information Portal from the U.S. National Library of Medicine)

Figure 2. Summary of the Drug Name Cleaning/Standardization Process
The drug name data needs to pass through a series of automated/manual cleaning steps before it is ready for use. Using these techniques on 2004 Quarter 1 → 2016 Quarter 3 FAERS data netted >95% standardization of DrugName data.

Table 1. Grouped Frequency Intervals of Drug Name Counts.



Frequency Interval	*Drug Name Counts (Original Download)*	*Total Records (Original Download)*	*Drug Name Counts* (Initial Cleaning)*	*Total Records* (Initial Cleaning)*	*Total Records Final % Cleaned*
Count >=1000	3138	27871132	1444	30849955	96.642
900 to 999	213	202147	52	48984	0.153
800 to 899	237	200782	57	48668	0.152
700 to 799	320	239407	67	50521	0.158
600 to 699	358	231018	87	56061	0.175
500 to 599	486	266939	89	48528	0.152
400 to 499	628	280317	105	46497	0.145
300 to 399	1006	348192	151	52566	0.164
200 to 299	1075	416009	248	60050	0.188
100 to 199	4084	569196	467	66186	0.207
30 to 99	13722	725441	1092	60504	0.189
(0 to 29)†	(530337)	(1386077)	(188445)	(533235)	(1.677)
Total uncleaned (-)	25267	32736657	22704	31921755	100.00
Total cleaned (+)			3859	31388520	98.33‡

*Null values and nonspecific drug names removed at this stage

†Data not reviewed for cleaning in this interval

‡Discrepancies in values due to rounding of numbers

Table 2. Drug Name Variation Types in the FAERS Database with Corrective Actions for Data Cleaning

Drug Name Variation	Example	Correction
Nonspecific/ambiguous drug description	"Pain med", "antibiotic"	Not able to be clarified; removed and not counted in record totals
Null value	"NULL"	Not counted in record totals
*Drug name misspelled	"acetaminofen", "oxycodon"	Changed to correct spelling (i.e. "acetaminophen", "oxycodone")
Only brand name represented without generic name	"Humira"	Drug names converted to generic for consistency
Inconsistent Upper/Lower case drug name description	"Aspirin", "aspirin", "ASPIRIN"	Names converted to upper case for consistent representation
Leading/Trailing Whitespace	"␣aspirin", "aspirin␣", "␣aspirin␣"	Character spaces removed
Keyboard new line/carriage return inserted	"<CR >aspirin", "aspirin <CR >"	Spaces removed
Tab character inserted	"»acetaminophen"	Spaces removed
Leading numbers	"123Humira"	Numbers removed
Special characters	"@Tylenol"	Characters removed
Drug name combinations as single entity	"lisinopril hydrochlorothiazide"	Single entries created with names separated by "and"
Abbreviations	"Vit B12"	Names expanded to complete description (i.e. "Vitamin B12"

*Invalid entries included spelling errors due to extra letters, numbers, characters, hyphens, extra spaces, omitted letters, and truncations

Table 3. Frequency Count of Drug Names in the FAERS Database

Rank	Drug Name	ATC Code	Frequency Count	Percentage of Database (%)*
1.	Aspirin	A01AD05 B01AC06 N02BA01	493379	1.55
2.	Adalimumab (Humira®)	L04AB04	492173	1.54
3.	Etanercept (Enbrel®)	L04AB01	447063	1.40
4.	Levothyroxine	H03AA01	329402	1.03
5.	Acetaminophen	N02BE01	303893	0.95
6.	Atorvastatin	C10AA05	284672	0.89
7.	Furosemide	C03CA01	277207	0.87
8.	Omeprazole	A02BC01	272731	0.85
9.	Interferon Beta-1a	L03AB07	270810	0.85
10.	Prednisone	A07EA03 H02AB07	261323	0.82
Total			**3432653**	**10.75**

*Based on cleaned data records; approximations are due to rounding of values