Supporting Online Material

A Process of Cleaning, Rectifying and Standardizing Drug Name Data from the U.S. Food and Drug Administration Adverse Event Database (FAERS)

Facilities

The Data Analytics Laboratory in the Computer Science Department of our academic institution 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.

Data Download

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 download of the FAERS quarterly reports for the specified time period resulted in 32,736,657 DRUG file records from which drug names were evaluated for errors.

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.

Data Organization

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.[1] Twelve frequency intervals were formed from the initial count of drug names, and the frequency of the number of drug name data entries within each interval was determined.

The process of drug name standardization involved assessing data in manageable "chunks" by frequency interval. Each frequency interval was analyzed beginning with the interval with the highest drug name frequency count (>=1000), down to the lowest drug name frequency count (0 to 29). That is, when the interval of ≥1000 was completed, the next interval, 900-999, and all subsequent intervals of drug names were processed as described below.

Rectifying and Standardizing Drug Names

Each drug name entry passed through a series of automated/manual cleaning steps. To assist in rectifying variations or errors in drug names, 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 are expressed in different ways (e.g., brand name, generic, or combination of both).[2] For consistency, a standard drug name format was identified and implemented, which involved expressing the generic drug name in lower case (i.e., no capitalized letters). All drug names identified were transformed into this format.

In addition to correct expression of drug names, the inputted 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.

Drug name references that were used to ensure correct identification of drug names included Drugs.com, Micromedex, and the Drug Information Portal from the U.S. National Library of Medicine.[3,4,5] An automatic check for correct names of drugs was employed with a proprietary spell-checker provided by Drugs.com,[6] and Micromedex Solutions intelligent searching, which provides in-line **s**pelling suggestions in real time while conducting searches.[4]

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.[7]

References

S1. Ali L. How to create a grouped frequency table. *Sciencing.* April 24, 2017. Available from: http://sciencing.com/create-grouped-frequency-table-5531910.html. Accessed September 30, 2019.

S2. Citrome L. (2016). What's in a name? Use of brand vs. generic drug names. *Int J Clin Pract.* 70(1): 3-4.

S3. Drugs.com. Available from: www.drugs.com/. Updated September 30, 2019. Accessed September 30, 2019.

S4. Micromedex Clinical Knowledge Suite. Available from: http://truvenhealth.com/Products/Micromedex/Product-Suites/Clinical-Knowledge. Accessed September 30, 2019.

S5. Searching Micromedex Solutions. Available from: www.micromedexsolutions.com/micromedex2/4.85.0/WebHelp/Home_Page/Searching/Searching.htm. Accessed September 30, 2019.

S6. Drugs.com. Contact Drugs.com. Corporate inquiry 181200-9009012003. Available at: www.drugs.com/support/contact.html. Accessed August 20, 2019.

S7. WHO Collaborating Centre for Drugs Statistics Methodology. Anatomical Therapeutic Chemical Classification - ATC Code. Oslo. Updated December 13, 2018. Available from: https://www.whocc.no/atc_ddd_index/. Accessed September 30, 2019.