Drug-Induced Liver Injury (DILI) Classification using US Food and Drug Administration (FDA)-Approved Drug Labeling and FDA Adverse Event Reporting System (FAERS) data

Qais Hatim1, Liliian Rosario1, Eileen Navarro Almaro1, Kendra Worthy1, Tom Sabo2, Emily McRae2
1 U.S. Food and Drug Administration, 2 SAS Institute Inc.

Research Problems

1. Why is it valuable to define DILI?
   - Defining DILI is challenging as:
     - It is a diagnosis of exclusion, the incidence and severity may be drug specific
   - Biomarkers and methodologies are being developed to assess hepatotoxicity but:
     - Require a list of drugs with well-annotated DILI potential
   - What do we get from assessing hepatotoxicity?
   - A drug classification scheme is essential to evaluate the performance of existing DILI biomarkers and discover novel DILI biomarkers but:
     - No adopted practice can classify a drug’s DILI potential in humans.
   - Do we label properly to save lives?
     - Drug labels used to develop a systematic and objective classification scheme Rule-of-two (RO2). However:
       - High context specific
       - Rarity of DILI in the premarket experience
       - Complex phenotypes of DILI
     - Drugs are often used in combination with other medications.

Research Solution

This research aimed to enrich the RO2 model based on machine learning and data-mining modeling of premarket and post-market DILI narrative reports. We utilized the FDA FAERS database that provides post-marketing surveillance data in order to improve the DILI classification by integrating the post-marketing data into the drug-label based approach to further improve the accuracy of DILI classifications. This research will develop a statistical prediction model for better predicting DILI in humans. Therefore, by Applying computational approaches in the analysis of FAERS, we advance the understanding of the feasibility of applying these advance techniques to uncover relationships between drugs and hepatic failure. Outputs offer information on:
   - The drug combinations related to reported hepatic failure
   - Clarify factors that can predict a greater degree of the hepatic failure from serious (death) to minor events (treatable)
   - Rate the serious impact of using these drugs on patients.

Data Extraction & Preprocessing

Imprecise Signal

1. Empirica Signal served as the source of data retrieval based on the preferred term (PT) or standard MedDRA query (SMQ) equal to Drug related hepatic disorders - severe events only (SMQ [narrow]). Prioritizing investigations might be based on statistics calculated as either the proportion of DILI or association, to avoid following up potential associations that could have arisen merely by chance.
   - However, unnecessary focus on drugs and events that are common overall in the database can be an outcome of using a PRR or ROR p-value rank associations.
2. Therefore, in this research, we prioritize investigations based on both significance and association scores, rather than relying on only one score. A threshold of EBGM>2 and EB50>1 are used and therefore the number of cases reduced to only 14,436 cases from the initial retrieved 171,890 cases.

Rule of two dataset

1036 FDA-approved drugs were classified into 192 vMost-DILI concern, 278 vLess-DILI concern, 312 vNo-DILI concern and 254 Ambiguous DILI concern drugs.

Model Building & Domain Experts

Analytics Application/Association Analysis

Association analysis is a popular technique that is used to identify and visualize relationships (associations) between different objects. Query could be nontrivial to be answered manually with big dataset. For example, what linkage of DILI preferred terms (events) can be observed from post-market data? Association analysis can address such relationship by defining association rules and calculating the support for the combination of the PTs.

Three scenarios are developed for the subset data from Empirica Signal (i.e., the 14,436 cases after utilizing both significance and association scores). Respective to each scenario, association model is built based on different settings for minimum support, minimum confidence, minimum lift, maximum antecedents, and maximum rule size. Selecting these values allows us to cover more association rules as well as understand the optimal setting that provides more informative rules.

Methodology

Research Questions

1. Why is it valuable to define DILI?
2. What do we get from assessing hepatotoxicity?
3. Do we label properly to save lives?

Data Sets Aggregation

This research dataset different two different domains, i.e., pre-marketing and post-marketing reports.

Drug exposure to RO2 classifying DILI risk based drug labeling

Imprecise Signal and Drug labeling analysis are performed on the FAERS data which is post-marketing data.

Evaluation

Numerically-annotated SQL was developed and used to perform SQL queries to query the drug data with 14,436 DILI drugs from FAERS.

Text Analytics

To meet the challenges posed by unstructured text, text mining, which employs a wide range of statistical, machine learning, and linguistic techniques for natural language processing (NLP), is utilized in this paper. Moreover, while some of the FAERS information is structured, a significant portion of it remains in narrative formats. Much of the information that is critical to risk assessments, such as signs and symptoms, disease status and severity, and medical history, are typically imbedded in narrative text. Accordingly, we explored both the structured and unstructured FAERS data items and developed different analytical models.