Adverse Event Open Learning through Universal Standardization (AEOLUS)

Lee D. Evans[1,3] , Richard D. Boyce[2,3], Erica A. Voss[4,3], Nicholas P. Tatonetti[5,3], Patrick B. Ryan[4,3], Juan M. Banda[6,3], Nigam H. Shah[6,3]

[1] LTS Computing LLC, West Chester, PA; [2] University of Pittsburgh, Pittsburgh, PA; [3] Observational Health Data Sciences and Informatics (OHDSI); [4] Janssen Research and Development, Raritan, NJ; [5] Columbia University, New York, NY; [6] Stanford University,Stanford, CA

Abstract

The US FDA Adverse Event Reporting System Database (FAERS) is an invaluable resource which has been used to generate pharmacovigilence safety signals for further investigation. However, it is challenging to take full advantage of this resource due to the use of 'free text' data entry of drug names. We created an ETL process to map FAERS drug names into RxNorm code ingredients & clinical drug forms and to map FAERS drug reactions & indications from MedDRA preferred terms into MedDRA codes & SNOMED-CT codes. Our ETL process standardized the FAERS adverse event reports, using the OHDSI Athena vocabularies, so they can now be analyzed using multiple vocabularies (including ATC, RxNorm & SNOMED-CT). The standardized version of the FAERS database that we created is called AEOLUS. It will be a useful evidence source for scientific research in:

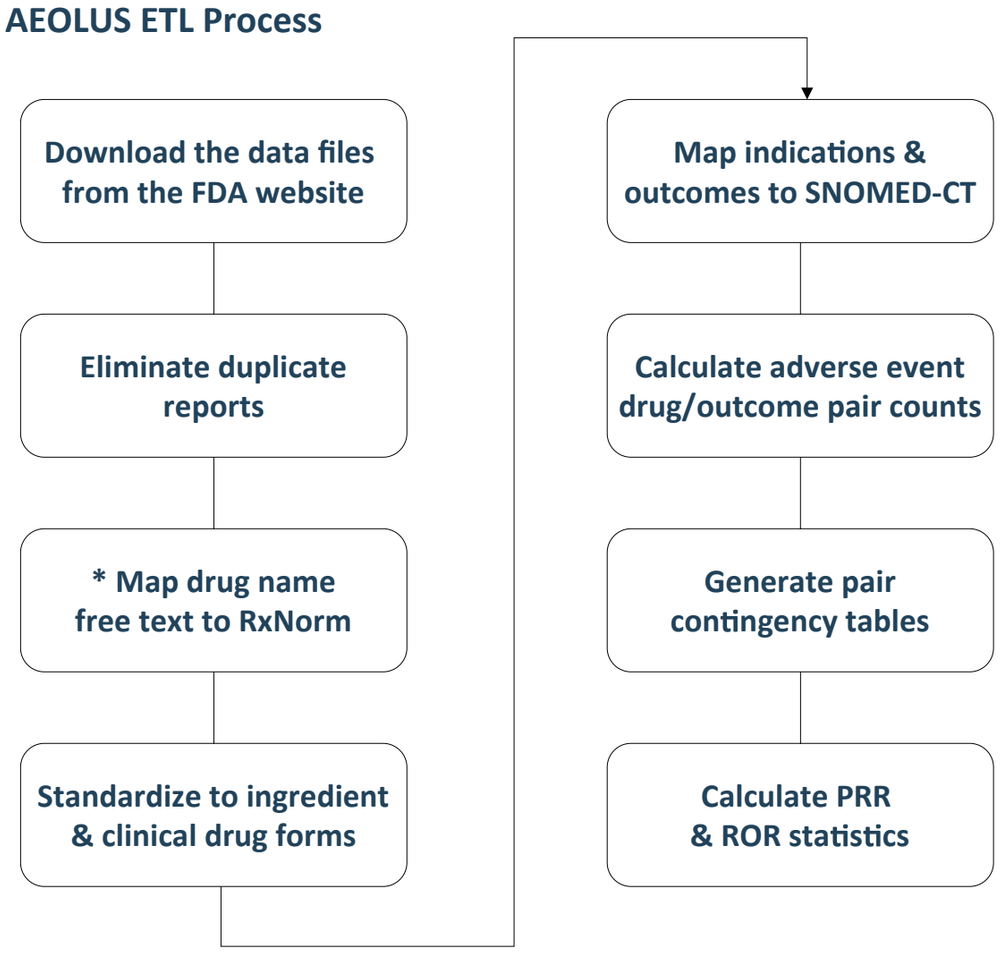
- Systematic discovery of adverse drug events from clinical notes
- Prediction of drug/drug interactions
- Pharmacovigilence safety signal detection and analysis

Background

The FAERS database files are publicly available on the FDA web site. There are quarterly files available for download, starting from Q1 2004. The database changed in Q4 2012 (especially the approach to managing unique ids for report case/version) so the AEOLUS ETL accommodates both the older legacy data (LAERS) format & the current FAERS format. We created the ETL process as a set of Linux shell scripts and SQL code that downloads & standardizes the data. The ETL could be run quarterly or biannually to maintain AEOLUS as a standardized, publicly available version of the FAERS database for scientific research. We ran the ETL using a 4 CPU 15GB memory PostgreSQL server hosted on the Google cloud. The ETL process is automated except for some manual code mapping using the OHDSI USAGI code mapping tool.

Methods

The AEOLUS ETL Process



* Map drug name using NDA number look-up in FDA Orange Book, regular expression matching to RxNorm drug name and manual mapping with the OHDSI Usagi mapping tool

Contact: levans@ohdsi.org

Results

Performance of the AEOLUS ETL process

We processed FAERS data from Q1 2004 through Q4 2014. The data contained over 6 million adverse event case reports. It took one and a half days to run the entire ETL process. The calculation of the contingency table counts for all the drug/outcome pairs was the longest running step in the ETL. It completed in 8 hours when the four contingency table count columns were generated in parallel.

Information loss

50% of all free text drug names could not be mapped to RxNorm codes. The unmapped drug names included spelling errors, non-specific names (e.g. "blood pressure medication" or "steroid") and unrecognized drugs from countries outside the US or EU. The occurrence of the unmapped drug names on FAERS adverse event reports was low (6.2% of total occurrences). 0.6% of the mapped drug name adverse event report occurrences could not be mapped from brand names & non-standard/deprecated vocabulary codes to standard ingredient or clinical drug form RxNorm codes. Around a third of drug/outcome pair outcome MedDRA codes and indication MedDRA codes could not be mapped to SNOMED-CT using the available MedDRA to SNOMED-CT mappings in the Athena vocabularies.

Vocabulary mapping measures

Measure	Value
Drug name adverse event report occurrences mapped to RxNorm	93.2%
Indication MedDRA codes mapped to SNOMED-CT codes	64.5%
Outcome (FAERS reaction) MedDRA codes mapped to SNOMED-CT	66.5%

Limitations of the current ETL process

- The ETL only recognizes US and EU drug names. The drug name identification code could be enhanced to identify drug names from additional countries.
- The matching algorithm that identifies duplicate adverse event reports could be improved by incorporating more sophisticated matching methods. e.g. probabilistic matching.

Conclusions

The standardized FAERS adverse event reports in the AEOLUS database will be a useful evidence source for the scientific community. The ETL process removes duplicate adverse event reports and standardizes the FAERS data within AEOLUS. By standardizing the data to the Athena vocabularies, the adverse event reports in AEOLUS can be analyzed & combined with clinical data using a variety of vocabularies (including RxNorm, SNOMED-CT & ATC). The AEOLUS data will be made openly available in the public domain. The ETL code we created is modular open-source code that is publicly available on GitHub. We welcome code improvements contributed from the OHDSI community or other interested parties.

Conflicts of Interest

LTS Computing LLC is a commercial IT projects & services company focused on Life Sciences. Stanford University provided funding to LTS Computing LLC for this project.

LTS Computing LLC