Dryad Blog post

As I sit in a room full of over one hundred bio-hackers at the 2016 Biohackathon in Tsuruoka, Yamagata, Japan, the need to have publicly available and accessible data for research use is acutely evident. Organized by Japan’s National Biosciences Database Center (NBDC) and Databases Center for Life Science (DBLS), this yearly hackathon gathers people from all over the world including the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), as well as universities from around the world; with purpose of extending and interlinking resources like PubChem, PhenomeCentral, Bio2RDF, PubAnnotation. The end goal being better ways to access data that will allow researchers to focus on analysis of the data rather than preparation.

In the same spirit, our publication “A curated and standardized adverse drug event resource to accelerate drug safety research”(​doi:10.1038/sdata.2016.26) helps researchers in the drug safety domain with the standardization and curation of the freely available data from the Federal Food and Drug Administration (FDA) adverse events reporting system (FAERS). FAERS collects information on adverse events and medication errors reported to the FDA and is comprised of over 10 million records collected between 1969 to the present. As one of the most important resources for drug safety efforts, the FAERS database has been used in at least 750 publications as reported by PubMed and was probably manipulated, mapped and cleaned independently by the vast majority of the authors of said publications. This cleaning and mapping process takes a considerably amount of time; hours that could have been spent analyzing the data further. Our publication hopes to eliminate this needless work and allow researchers to focus their efforts in developing methods to analyze this information.

As part of the Observational Health Sciences Initiative (OHDSI), whose mission is to “Improve health, by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care”, we decided to tackle the task of cleaning and curating the FAERS database for our community, and the wider drug safety community. By providing a general common data model (CDM) and a general vocabulary to standardize how electronic patient data is stored, OHDSI allows its participants to join a research network with over 655 million patients. With a significant fraction of the community’s research being focused on drug safety, it was a natural decision to standardize the FAERS database with the OMOP vocabulary, to allow all researchers on our network access to FARES. Since the OMOP vocabulary incorporates general vocabularies such as: SNOMED, MeSH, and RxNORM among others, the usability of this resource is not limited to participants of this community.

In order to curate this dataset, we use the source FAERS data in CSV format and de-duplicate cases reports which are featured more than once. We then perform value imputation for certain fields that are missing. Drug names have been standardized to RxNorm ingredients and standard clinical names (for multi-ingredient drugs), this mapping is tricky because in some drug names have spelling errors, some are non-prescription drugs, as well as international brand names. We achieved coverage of 93% of the drug names, which in turn cover 95% of the case reports in FARES. For the first time, the indication and reactions have been mapped to SNOMED-CT from their original MedRA format. Coverage for indications and reactions is around 64% and 80% respectively. The OMOP vocabulary allows RxNorm drug codes as well as SNOMED-CT codes to reside in the same unified vocabulary space, simplifying use of this resource. We also provide the complete source code we developed in order to allow researchers to refresh the dataset with the new quarterly FAERS data releases and improve the mappings if needed. We encourage users to contribute the results of their efforts back to the OHDSI community.

With a firm commitment of making open data easier to use, this resource allows researchers to utilize a professionally curated (and refreshable) version of the FAERS data, enabling them to focus on improving drug safety analyses and finding more potentially harmful drugs, as a part of OHDSI’s core mission.

The data:

http://dx.doi.org/10.5061/dryad.8q0s4

A full description of the dataset in Scientific Data:

http://www.nature.com/articles/sdata201626