The Journal of Open Source Software

Assembling a Historical National Drug Code Directory from the Internet Archive

Mark Howison[1, 2], **Ted Lawless**[1], **and John Ucles**[1]

1 Research Improving People's Lives, Providence, RI, USA **2** Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA

DOI: 10.21105/joss.01128

Software

- Review
- Repository
- Archive

Submitted: 13 December 2018
Published: 14 December 2018

Summary

The `historical-ndc` package provides a data-processing pipeline implemented in Python to process historical data from the National Drug Code Directory, a legislatively mandated database of drug data maintained by the Food and Drug Administration (FDA). Every distinct preparation of a marketed prescription drug is assigned a unique National Drug Code (NDC) in this database, and these codes are also used in medical claims data to identify the dispensed prescription drug associated with a pharmacy claim. Researchers who study medical claims can join data on prescription drugs to pharmacy claims using NDCs. In particular, it is often useful to classify groups of NDCs for drugs that share a particular function or medical use. For example, in our own research we are interested in identifying claims for prescription opioid drugs to study the opioid epidemic.

The NDC Directory is a current snapshot and does not contain historical data on drugs that are discontinued or no longer marketed. In addition, the FDA transitioned the NDC Directory from paper-based to electronic submissions in 2011, and removed data from paper-based records. Although the FDA provides a classification of drugs in the NDC Directory, it has not been consistent over time. The Directory included an initial classification up until the beginning of 2005, removed the classification between 2005 and 2011 while developing a new classification, and published the new classification starting in mid-2011. Because of these inconsistencies and missing data, research with historical pharmacy claims data will have an increasing number of claims going back in time that cannot be matched on NDC to drug data in the current Directory. The `historical-ndc` pipeline builds a comprehensive list of historical NDCs linked to their last known drug data in prior snapshots of the Directory.

Data

We systematically obtained all available snapshots from the Internet Archive of the following files from the FDA's website:

- http://www.fda.gov/cder/ndc/drugclas.txt
- http://www.fda.gov/cder/ndc/formulat.txt
- http://www.fda.gov/cder/ndc/listings.txt
- http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/UCM070838.zip
- http://www.accessdata.fda.gov/cder/ndc.zip
- https://www.accessdata.fda.gov/cder/ndctext.zip

To create a full list of available snapshots for these URLs, we reverse engineered the React.js interface of the Internet Archive website and made HTTP requests to obtain the JSON data that populates the web interface (see `InternetArchive/timestamps.py`). The resulting list of unzipped filenames and timestamps is located in `InternetArchive/files.csv`. We downloaded, unzipped, checksummed, and packaged all of the files, as of April 11, 2018, into a single archive available from Figshare (Howison, Lawless, & Ucles, 2018).

The pipeline starts by downloading the prepackaged Internet Archive files from Figshare and binning them by year. Each year's files are then loaded, joined, and collated according to three different formats used by the Directory during the periods 2000-2005, 2006-2010, and 2011 to the present. The pipeline is automated using the SCons software construction tool (Knight, 2005). It outputs a list of all distinct combinations of NDC, drug name and DEA schedule (`output/ndc-drugs.csv`), and all distinct combinations of NDC with active ingredient name, amount, and unit (`output/ndc-ingredients.csv`). In total, there are 202,557 drugs and 417,272 ingredient records.

Application to Classifying Opioid Drugs

We provide an example of how the processed drug and ingredient files can be used to classify drugs based on the presence of certain active ingredients above a threshold amount. In particular, we use the opioid prescribing guidelines from the Washington State Agency Medical Directors' Group (2015) to define a minimum thresholds of active ingredients for prescription opioid drugs. That is, we classify a drug as a prescription opioid if it contains an opioid ingredient at or above the minimum amount recommended by Washington State as the starting dose for opioid therapy (see `lib/opioids.py` for the threshold amounts). Additionally, we classify a drug as a "recovery" prescription drug if it contains any amount of an ingredient used in medication-assisted treatment of opioid use disorder. The resulting classification in `output/ndc-opioids.csv` identifies 4,175 drugs as prescription opioids and 414 as recovery drugs. An additional 184 drugs contain an opioid ingredient, but not at or above the threshold. This example can be extended to any other class of drug that is defined by a table of threshold amounts for specific active ingredients.

References

Howison, M., Lawless, T., & Ucles, J. (2018). Historical data from the National Drug Code Directory. doi:10.6084/m9.figshare.6128225.v1

Knight, S. (2005). Building software with SCons. *Computing in Science Engineering*, *7*(1), 79–88. doi:10.1109/MCSE.2005.11

Washington State Agency Medical Directors' Group. (2015). *Interagency guideline on prescribing opioids for pain.* Retrieved from http://www.agencymeddirectors.wa.gov/Files/2015AMDGOpioidGuideline.pdf