BioDEX (Biomedical Drug Event Extraction) Dataset Card

Karel D'Oosterlinck, April 2023

Disclaimer: work in progress, not yet published, everything subject to change.

Links

- HuggingFace Raw dataset
- HuggingFace ICSR dataset

Motivation

Medical experts often publish case studies that describe adverse drug events (ADEs). Minimally, these ADEs describe **a patient** (e.g. 'a 52-year old female') who took **a set of drugs** (e.g. 'BRENTUXIMAB VEDOTIN, CARMUSTINE, CISPLATIN, CYCLOSPORINE') and experienced **a set of adverse reactions** (e.g. 'Acute respiratory distress syndrome, COVID-19 pneumonia, Epstein-Barr viraemia').

Surfacing ADEs is an important step in the pharmacovigilance pipeline, which is absolutely essential for public safety. In the United States, this process is regulated by the FDA. Specifically, pharmaceutical companies are required to periodically survey the recent biomedical literature to catch any ADEs reported about the drugs they produce. The pharmaceutical companies are required to fill-in a report (called an Individual Case Safety Report or ICSR) for every relevant ADE surfaced. These reports get saved in the FAERS database (FDA Adverse Event Reporting System) and are publicly accessible.

To achieve this, pharmaceutical companies employ highly-skilled medical experts to triage and analyze recent biomedical literature. This is a challenging task, experts need to read entire papers and use their domain knowledge to fill in reports. These experts are placed under constant time pressure to keep up with all the recent literature, since this is important from both a regulatory and public safety perspective.

In this work, we introduce the first large-scale dataset for Individual Case Safety Reporting. Our dataset features abstracts as well as full-text papers from PubMed with their corresponding ICSR reports from the FAERS dataset. Our resource is rooted in the historical output of all regulated reporting in the U.S. and is thus the result of the work of countless medical experts.

As far as we know, our dataset is the first to introduce the concept of ICSR reporting to the NLP community. We hope that our resource enables the development of (semi-)automated ICSR reporting systems, which one day could aid humans perform this task. Additionally, we believe our task is a good resource to train and evaluate the biomedical capabilities of Large Language Models.

No extra annotations were needed in the creation of the dataset. Thus, we will be able to scale our dataset seamlessly when new articles and reports are published.

Data

Our dataset is created by merging PubMed MEDLINE articles (parsed using pubmed_parser) with reports from FAERS. Additionally, we add full-text papers where possible by downloading these from the PubMed Central Open Access Dataset.

Description of the data fields

Every row in our raw dataset consists of one article and a list of reports. We have implemented this interface in python to allow for easy manipulation of all the data.

Article Fields

Almost all fields are derived from the pubmed_parser package:

- pmid : PubMed ID
- pmc : PubMed Central ID
- doi : DOI
- other_id : Other IDs found, each separated by ;
- title : title of the article
- abstract : abstract of the article
- authors : authors, each separated by ;
- mesh_terms : list of MeSH terms with corresponding MeSH ID, each separated by ; e.g. 'D000161:Acoustic Stimulation; D000328:Adult; ...
- publication_types : list of publication type list each separated by ; e.g. 'D016428:Journal Article'
- keywords : list of keywords, each separated by ;
- chemical_list : list of chemical terms, each separated by ;
- pubdate : Publication date. Defaults to year information only.
- journal : journal of the given paper
- medline_ta : this is abbreviation of the journal name
- nlm_unique_id : NLM unique identification
- issn_linking : ISSN linkage, typically use to link with Web of Science dataset
- country : Country extracted from journal information field
- reference : string of PMID each separated by ; or list of references made to the article
- delete : boolean if False means paper got updated so you might have two
- languages : list of languages, separated by ;

Additionally, we add two fields concerning the full-text papers:

- fulltext: string containing the full-text of the paper if available
- fulltext_license: string containing the license of the full-text if available

Report fields

We parse this information from the FAERS dataset. The original descriptions of the fields can be queried using OpenFDA. A .yaml file of all fields, their format, description and

possible values can be found here. We don't include the fields under the 'openfda' object since they are not included in the original FAERS distribution.

Each report object contains nesting since a given report can contain a variable amount of reported drugs and reactions.

TODO: we highlight some of the most important fields here.

Basic statistics

We refer to our analysis notebook that will be published soon for more detailed results.

Size

Our dataset features 65.6k unique PubMed articles with one or more associated ICSR reports. For all of these articles, the abstract is available. In total, 19.3k articles also have the full text of the article available.

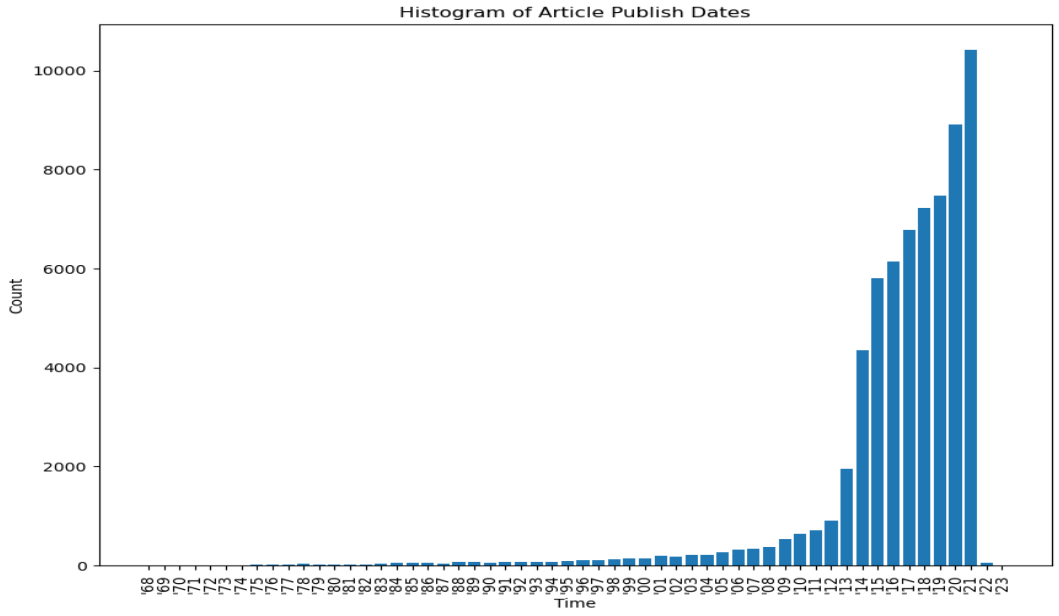
Our dataset features a total of 256k ICSR reports. Thus, there are about 4 reports per article on average.

Abstract and full-text length

The average abstract is 1,277 characters long (~319 tokens assuming 4 chars per token). The average full text is 23,767 characters long (~5,942 tokens assuming 4 chars per token).

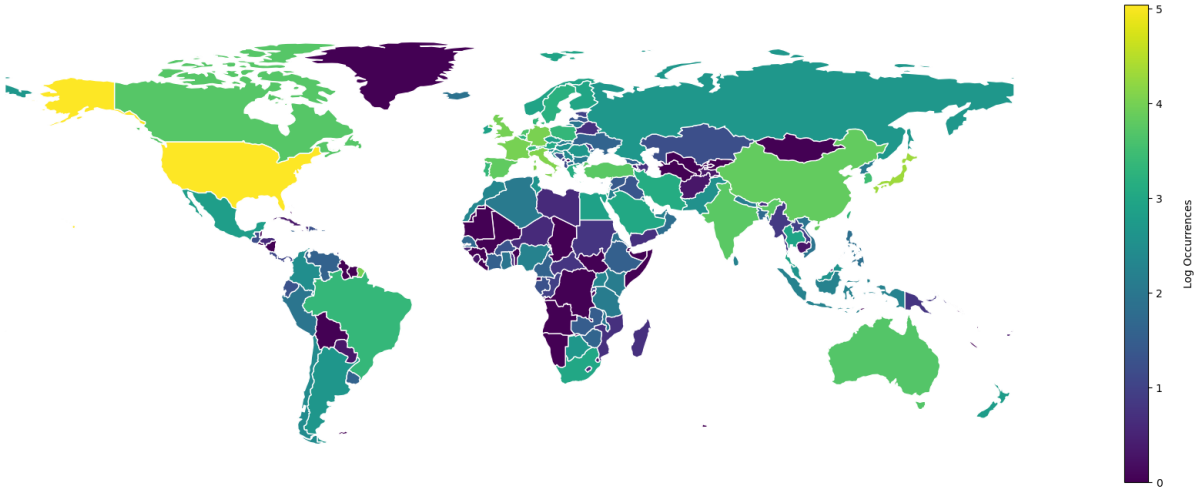
Time period

Our dataset features articles from 1968 till 2022. Most articles are published after 2013 and before 2022. See the distribution below.



Origin country of the report

While the report database is mainly focussed on the United States, reports can be submitted from all over the world. The heatmap below shows the occurrence country of reports in LOG scale. Most reports occur in the US, but a significant amount of reports also come from Japan, Canada, Western Europe, Australia, China and India. Some regions/continents, such as Africa, are underrepresented in the data.



Different countries can face different health issues. When we develop biomedical language systems, it is important they work for everyone. Subsequent data collection efforts could focus on underrepresented countries to alleviate this issue.

Tasks

Our raw resource can be used in multiple settings. We start by exploring document-level ICSR extraction.

ICSR Extraction

The goal of this task is to predict a coarse version of a report. The model is given as input either the abstract (abstract-level ICSR extraction) or the abstract and full-text (document-level ICSR extraction).

The coarse representation of the report contains 4 fields:

- serious: the report.serious field
- patientsex: the report.patient.patientsex field
- drugs: an alphabetized list of drug activesubstance names associated with the ICSR. Created by selecting drug.activesubstance.activesubstancename for every drug in report.patient.drugs.
- reactions: an alphabetized list of reaction names associated with the ICSR. Created by selecting reaction.reactionmeddrapt for every reaction in report.patient.reactions.

This coarse representation is then embedded in a string to allow for seq2seq modeling. Example:

```
"serious: 1
patientsex: 1
drugs: ACETAMINOPHEN, ASPIRIN DL-LYSINE, BROMAZEPAM, HALOPERIDOL, HEPARIN CALCIUM,
HYDROCORTISONE, HYDROXYCHLOROQUINE, LINEZOLID, MEROPENEM, METHYLPREDNISOLONE, MIDAZOLAM,
MORPHINE, OLMESARTAN MEDOXOMIL, OXYGEN, PREDNISOLONE, TOCILIZUMAB
reactions: Diarrhoea, Drug hypersensitivity, Rash"
```

Leaderboard

Model	F1	train loss	eval loss	Params	Remarks
stanford-crfm/BioMedLM	0.3644	0.7793	0.8546	2.7B	These results should be much better. Training dynamics are bad.
gpt2	0.4530	0.3677	0.6796	124M	
gpt2-fulltext (512 tokens)	**0.4787**	**0.3266**	**0.6473**	**124M**	
google/flan-t5-small	0.3598	1.3469	1.3288	60M	
“human performance” on raw resource	**0.7204**	/	/	/	
“random performance” on raw resource	0.2428	/	/	/	

Medical QA (work in progress)

Our raw resource contains many interesting fields that can be exploited in a QA-type setting.

For example, we can mine questions and answers following multiple formats:

- What was the indication for the drug X taken by patient Y
- What was the dosage for the drug X taken by patient Y
- What was the total dosage consumed for drug X taken by patient Y
- Which drug(s) led to reaction(s) X (for patient Y)
- Which reaction(s) was associated with drug(s) (for patient Y)
- What as the outcome of reaction X (for patient Y)
- etc…

We can also ask questions that are not related to the reports (we could do this on any pubmed resource):

- What would be possible keywords for this abstract?
- What are some MeSH headings for this article?
- What types of chemicals are discussed in this article?
- etc…