Mining ClinicalTrials.gov via CTTI-AACT for Drug interventions

Jeremiah Abok Chemistry & Chemical Biology University of New Mexico

Introduction

Drug named entity recognition (D-NER) is a critical step for complex biomedical NLP tasks (8). D-NER is the task of locating drug entity mentioned in unstructured medical texts and classifying their predefined categories (9). NER in biomedicine is focused proteins, genes, diseases, drugs, organs, DNA sequences, RNA sequences and others (10). By mining ClinicalTrials.gov using the Clinical Trial Transformation Initiative – Aggregate Analysis of ClinicalTrials.gov database (ctti-aact db) from Duke University, new and distinctive drug target knowledge can be inferred from drugs and diseases. Extraction of chemical named entities (CNEs)is an essential task since it allows using the obtained data for building chemical-target associations and other relationships relevant for drug discovery.

Each table includes the unique identifier nct_id assigned by ClinicalTrials.gov

To find all data about a particular study & to join related information across multiple tables, the key is nct_id

In most cases, you need to use the nct_id to join tables.

Static copies of the DB as postgres_data.dmp - these are not text files. They are binary that come from Heroku's backup mechanism

Do you prefer to access the data via simple flat files? Static copies are also available as a zipped package of pipe-delimited files

Downloaded pipe delimited files dated 31October2022

There are 51 tables in this folder

Question1: How many drugs?

The first goal is to identify the types of interventions and their counts in interventions.txt

The second goal is to extract names of drugs and their unique ids from interventions.txt

Input file = interventions.txt

Data attributes: id (#), nct_id(e.g., NCT03309709), intervention_type (Behavioral, Biological, Combination Product, Device, Diagnostic Test, Dietary Supplement, Drug, Genetic, Other, Procedure or Radiation), name (e.g., Tramadol), description (e.g., IV 4 mg/ mL tramadol solution into 100 mL normal saline)

Question2: What is the start year of these studies?

Determine when these studies started

There are 3 different start dates format - start_month_year, start_date_type, & start_date in columns #19, #20, & #21.

Interest is in column #19 (start_month_year)

Input = studies_drug.txt

Data – 69 attributes

Question3: How many studies? How many references?

Determine the number of references using study_references table & nct_ids

I have obtained the nct ids for the intervention_type, drugs

Input files = study_references.txt & nct_ids_drugs.txt

Data attributes: [id|nct_id|pmid|reference_type|citation]

Question4: Drug trials by phases

In what phases are these drugs?

Input = studies_drug.txt

Data – 69 attributes one of which is the phase column #38)

Question5: Drug trials by status

What is the overall status of these drugs?

Input = studies_drug.txt

Data – 69 attributes one of which is the overall_status column #36

Identifying drug names

Using the drug names list generated then applying Chemical Identifier Resolver to resolve names to identify and find the chemical structure

The R package webchem "Chemical Information from the Web" interacts with a suite of web services for chemical information such as Chemical Identifier Resolver, ChEBI, ChemIDplus, ChemSpider, PubChem,

Wikidata etc. Current version is Version 1.2.0. (12). I used webchem to query chemical identifier resolver.

Results & Conclusions

Chemical NER (Named Entity Recognition)

Intervention names of drug trials: 185,768; drug names: 21,667; SMILES: 5,329 with many non-structure terms. The top 5 drug mentions are - 3348 paclitaxel, 3070 cyclophosphamide, 2985 cisplatin, 2620 carboplatin, 2384 gemcitabine. Other drug names mentioned in thousands are - 2224 metformin, 2074 docetaxel, 2041 bevacizumab, 1763 rituximab, 1592 doxorubicin, 1579 capecitabine, 1556 pembrolizumab, 1538 bupivacaine, 1408 oxaliplatin, 1296 fludarabine, 1281 etoposide, 1221 dexmedetomidine, 1209 azd, 1194 nivolumab, 1160 propofol, 1121 irinotecan, 1064 ketamine, and 1042 cytarabine.

MeSH (Medical Subject Heading) terms

The top MeSH terms mentioned connected to these drugs are neoplasms 47766, neoplasms by site 28070, pathologic processes 26197, neoplasms by histologic type 24750, immune system diseases 20105, digestive system diseases 16956, nervous system diseases 16667, cardiovascular diseases 16586, infections 16417, respiratory tract diseases 16038, lung diseases 13275, vascular diseases 3086, mental disorders 11992, metabolic diseases 11603, endocrine system disease 11522, skin diseases 11333 and neoplasms, glandular and epithelial 11296.

SMILES

Resolving the 21, 667 drug names was done using webChem. 5,329 SMILES were generated from these names.

Drug Target identification

One of the failures of this approach is due to limitations posed by the tools deployed. It was not possible to find a way around mapping the chemical identification id(cid) generated from PubChem via webChem to ChEMBL.

Challenges

Drugs are known by any of these three names - chemical name, generic name, or brand name. Chemical name is usually complex; a brand name may not identify a drug (especially once the relevant patents expire) and; generic name is negotiated when the drug is approved for use, but this is regulated by the

World Health Organization (WHO) and local agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency

Identifying chemical names is especially challenging for the reasons stated below:

variety of language expressions

variants can include an innumerable mix of typographical variants, alternating uses of hyphens, brackets, spacing, and word order

have many synonyms e.g., triphenyl phosphate, TPP, TPHP, OP(OC6H5)3, phosphoric acid, triphenyl ester, CAS: 115-86-6, ((PhO)3PO)) refer to the same chemical

mentions can refer to mixtures of chemicals

often ambiguous, especially when abbreviated

the complexity makes it difficult to identify exact boundaries for entity mentions and word tokens

Limitations of ChEMBL - ChEMBL records only link from PubChem via PubChem Bioassay (AIDs) but lacks direct links to PubChem substances (SIDs) (4).

Deploying Statistical learning models for NER is one of the common approaches in overcoming some of these challenges

This approach defines drug NER as a classification task that can classify an input expression either as an entity or not; and requires annotated data for training supervised models with human curated data, the gold-standard. But biomedical gold or silver standard corpus (3) is required to make this approach a success.

Future steps ...

Chemical NER with NextMove LeadMine

Compound-target mapping via PubChem & ChEMBL ; and targets via ChEMBL bioassays.

Targets mapping to IDG-TCRD/Pharos