NeuroTrialNER - Data Annotation Guidelines

Simona Doneva   Benjamin Victor Ineichen   Amelia Cannon

November 7, 2023

Contents

1 Task Description

The objective of the work is to create an annotated dataset for the text-mining challenge of information extraction from clinical trial registries. Specifically, we are interested in developing an automated solution for: (1) Intervention Named Entity Recognition and (2) Condition Named Entity Recognition.

This document contains the guidelines for the corpus construction process. To maximize data interoperability among the BioNLP community, our conventions are aligned closely with previous works [6], [5]. Our annotation rules are also guided closely from the PICO framework. Annotating Condition/Intervention is guided by the Population question: ”Who is the group of people being studied?” and the Intervention question ”What is the intervention being investigated?” respectively. The Control label is associated with the question: ”To what is the intervention being compared?” [3].

2 Task Data

The data source is the Aggregate Analysis of ClinicalTrials.gov (AACT) database of registered and publicly available studies in ClinicalTrials.gov. We first obtain a random sample of 1500 data rows from the *studies* table with the following query:

```
CREATE MATERIALIZED VIEW ctgov.random_sample_view AS
SELECT nct_id, start_date, completion_date, phase, official_title
FROM ctgov.studies order by RANDOM()
WHERE official_title IS NOT NULL
LIMIT 1500;
```

Afterwards, we obtain the description of the studies from the *brief_summaries* table by joining on the *nct_id* field:

```
CREATE OR REPLACE VIEW ctgov.random_sample_view_with_summaries AS
SELECT s.nct_id , start_date , completion_date , phase,

official_title, description
FROM ctgov.random_sample_view s
LEFT JOIN ctgov.brief_summaries bs ON bs.nct_id = s.nct_id
WHERE description IS NOT NULL;
```

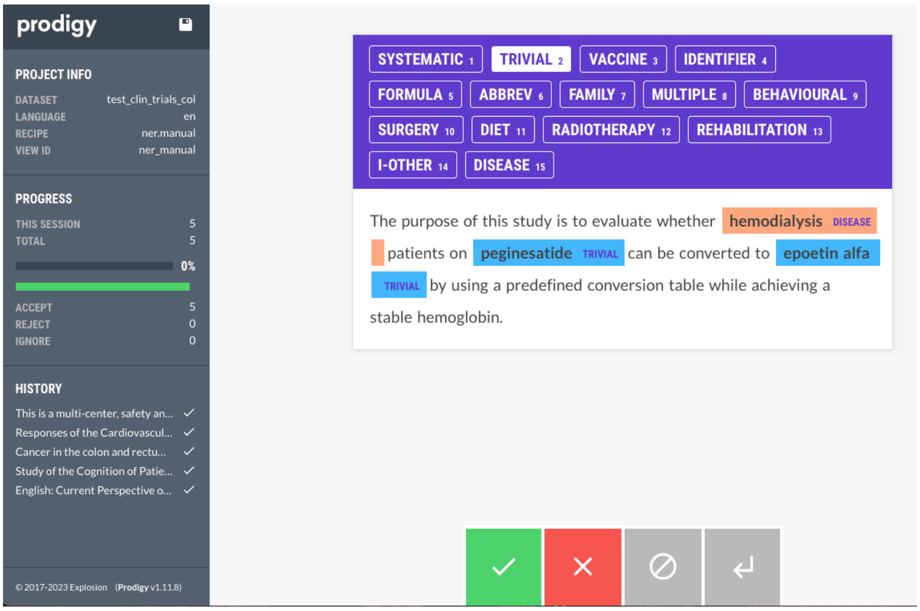
This results in a table of 1500 rows that has as columns the values for the *study ID*, *official title*, *description* and some additional metadata like the study begin and end dates. We convert the rows of the table into a JSON Lines file with the following schema example:

```
{"nct_id": "NCT00741221",
"source": "OfficialTitle",
"text": "Pemetrexed in Non Small Cell Lung Cancer (NSCLC)".}
```

During the generation of the file we concatenate the *official title* and *brief summary* text. This .jsonl file is the input to the Prodigy annotation tool [7]. We further generate an initial random sample for annotation from this file of 1500 lines, which we can increase depending on our needs.

3 Annotation Tool

Human annotation of the texts for the target Intervention and Disease named entity types is performed using Prodigy [7].



DATASET test_clin_trials_col
LANGUAGE en
RECIPE ner.manual
VIEW ID ner_manual

THIS SESSION 5
TOTAL 5
0%
ACCEPT 5
REJECT 0
IGNORE 0

This is a multi-center, safety an... ✓
Responses of the Cardiovascul... ✓
Cancer in the colon and rectu... ✓
Study of the Cognition of Patie... ✓
English: Current Perspective o... ✓

The purpose of this study is to evaluate whether hemodialysis DISEASE patients on peginesatide TRIVIAL can be converted to epoetin alfa TRIVIAL by using a predefined conversion table while achieving a stable hemoglobin.

© 2017-2023 Explosion (Prodigy v1.11.8)

Figure 1: A screenshot of Prodigy annotation page.

4 Guidelines

4.1 General Guidelines

1. The curators are encouraged to crosscheck information from reference sources such as Wikipedia, and chemical databases (ChEBI, DrugBank, etc.) to facilitate the annotation process and ensure compliance with the guidelines.

2. Do not tag unclear cases. If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabelled.

3. Mentions should be annotated considering the context in which they are used and only if fulfill the definitions for Condition and Intervention described in 4.2 and 4.3. E.g. While the word *Immunotherapy* is a valid Intervention in some cases, it is not to be annotated in the sentence ”The Efficacy and Safety of the United Allergy Service (UAS) Immunotherapy Protocol”, as it has a different semantics in this context. If the text mentions the same intervention/condition in another context, e.g. existing research such as animal studies, it should be annotated. Example of the latter is the text: ”Different Efficacy Between Rehabilitation Therapy and Umbilical Cord Derived Mesenchymal Stem Cells Transplantation in Patients With Chronic Spinal Cord Injury in China — [...] However, it can not repair the damaged nerve function. Studies show that mesenchymal stem cell transplantation can remarkably improve the neurological function of SCI in animals without any severe side effect.” Here the tokens ”mesenchymal stem cell transplantation” and ”SCI” should be labeled in the last sentence.

4. Conditions are more reliably maintained in AACT than Interventions. Therefore we have a more broad inclusion criteria for Interventions than Conditions, which need to be more specific to be annotated. If there is an overlap in the phrase, we prefer annotating for the intervention rather than the condition, e.g. in ”Clinical Assessment of Perfusion Techniques During Surgical Repair of Coarctation of Aorta With Aortic Arch Hypoplasia in Infants” the phrase ”Surgical Repair of Coarctation of Aorta With Aortic Arch Hypoplasia” should be annotated as INTERVENTION.

5. Conditions and Interventions should be annotated only if they appear in relation to the target study population or intervention. E.g. in ”Pain is a common symptom of Multiple Sclerosis. In the present study we assess whether aspirin relieves headache.” the words ”Pain” and ”Multiple Sclerosis” should not be annotated, while ”aspirin” (DRUG) and ”headache” (CONDITION) should be annotated.

6. Interventions or Conditions mentioned within the context of the study name, should not be annotated. E.g. ”Nova Scotia Chronic Pain Collaborative Care Network: A Pilot Study” should result in no annotations.

7. If there are multiple CONDITION or INTERVENTION mentioned which are separated with ”versus”, ”vs”, ”and”, ”or”, ”/” or similar, annotate preferably as separate entities. A positive example is ”Rehabilitation program by rhythmic auditory cueing” - here ”Rehabilitation program” and ”rhythmic auditory cueing” should be annotated separately. However, if the words can’t stand by themselves, the whole phrase should be annotated as one entity. E.g. ”Moderate and Severe Dementia”, ”early versus standard AR therapy” should be annotated together. In ”Multimodal Opiate-sparing Analgesia Versus Traditional Opiate Based Analgesia”, the two INTERVENTIONs can be clearly separated in two entities: ”Multimodal Opiate-sparing Analgesia” and ”Traditional Opiate Based Analgesia”.

8. If possible, the labeled word string should not be a combination of terms with and without brackets. E.g. ”oral appliance (OA) device” should result in two labeled words ”oral appliance” and ”OA”.

9. Typing errors or formatting errors should be labelled, unless they have impact on the tokenization provided by Prodigy and would result in wrong entity span.

4.2 Condition Mention Annotation

Our working definition for a **Condition** is any ”state labeled as diseases by virtue of consensus on prevalent sociocultural and medical values”. It has to have ”clearly identifiable diagnostic features and disease progression, and response to specific treatment.” [2] In contrast, we do not label the symptomatic manifestation of a disease, that is the ”self-conscious sensation of dysfunction and/or distress that is felt to be limitless, menacing and aid-requiring.” [4]

Whenever possible we will follow closely the annotations presented in [6].

What to annotate?

1. As a general guideline, annotated should be conditions that have an ICD-11 code [1].

2. We annotate conditions even in the absences of an intervention or if a diagnostic/explorative method was investigated in the trial.

3. Further defining characteristics should be included: Acute/Chronic; Active/Inactive; Mild/-Moderate/Severe; End Stage/Early Stage; Drug-resistant; Total/Partial; Intermittent/Relapsing and others. Similarly, ”Post-stroke” should be annotated instead of only ”stroke” because it refers to the phase after the acute stroke. This includes genotypes further specifying diseases, e.g. ”GBA-associated Parkinson’s Disease.”

4. Annotate deficiencies of one or more essential vitamins, e.g. ”Vitamin B deficiency”, ”Zinc deficiency”.

5. Annotate words like ”pain” and ”cognitive dysfunction”, only if is a clear target for the intervention. It should not be annotated if its role is an OUTCOME, e.g. In the case of ”Test if [...] offer a better pain relief.”, the word ”pain” should not be annotated.

6. Compound strings like ”PwMS” (Person with Multiple Sclerosis) should not be annotated.

7. Symptoms should be annotated only if they are a clear target of the Intervention, e.g. in ”depressive symptoms after stroke” both ”depressive symptoms” and ”stroke” should be annotated separately.

8. Annotate the most specific disease mentions. For instance, the complete phrase “partial seizures” should be preferred over “seizures” as it is more specific.

9. Annotate minimum necessary text spans for a disease. For example, select “hypertension” instead of “sustained hypertension.”

10. Annotate all mentions of a disease entity in an abstract. All occurrences of the same disease mention should be marked, including duplicates within the same sentence.

11. Annotate abbreviations.Abbreviations should be annotated separately. For instance,“Huntington disease (HD)”should be separated into two annotations: “Huntington disease” and “HD”.

12. Annotate mentions with morphological variations such as adjectives. Only when the adjective describes a specific disease. For instance, “hypertensive” should be annotated as it comes from “hypertension.”

[1] `https://icd.who.int/browse11/l-m/en`

13. Annotate all words from a composite disease mention should be annotated. For example in ”ovarian and peritoneal cancer”, ”ovarian and peritoneal cancer” should be annotated as one entity.

What not to annotate?

1. Do NOT annotate words that define *how* a disease is expressed, e.g. plaque in ”plaque psoriasis”.

2. Do NOT annotate patient demographics, e.g. ”elderly people”.

3. Do NOT annotate the word ”patient”, e.g. ”knee surgery patients”.

4. Do NOT include species names as part of a disease. Organism names such as “human” are generally excluded from the preferred mention unless they are critical part of a disease name. Viruses, bacteria, and other organism names are not annotated unless it is clear from the context that the disease is caused by these organisms. e.g. “HIV-1-infected” means the disease caused by the organism “HIV”.Thus, “HIV” should be included.

5. Do NOT annotate symptoms, e.g. stomach ache, headache, arm weakness. Unless it’s a clear target of the Intervention, e.g. in ”depressive symptoms after stroke” both ”depressive symptoms” and ”stroke” should be annotated separately.

6. Do NOT annotate general terms that occur individually and are not specific, such as: disease, syndrome, deficiency, complications, etc.

7. Do NOT annotate references to biological processes such as “tumorigenesis” or “cancerogenesis”.

8. Do not annotate the condition if it is within another linguistic expression. For example, in ”Total Tic Severity Index”, ”Tic” should not be annotated.

4.3 Intervention Mention Annotation

Our working definition of **Intervention** includes any ”treatment, procedure, or other action taken to prevent or treat disease, or improve health in other ways.” [1]

For the annotation on Drug/Chemical-based therapies, we follow closely the guidelines of constructing CHEMDNER corpus for annotating chemical mentions [5], as well as [6]. The basic rule for chemical entity annotation is that the chemical should have a specific structure.

General guidelines:

1. Annotate both the tested intervention and its control intervention, e.g. ”home visits (OTHER) vs out-patient visits (CONTROL)” results in two annotations. A special label for CONTROL is provided.

2. In the case of a non-drug intervention, annotate all further specifying terms. E.g. in the sentence ”[...] a single injection Transmuscular Quadratus Lumborum (TQL) block, when compared to [...]”, the whole phrase ”single injection Transmuscular Quadratus Lumborum (TQL) block” should be annotated. Words in parenthesis that give further details about the intervention should not be annotated, e.g. in ”remote visit (via phone or videochat)” only ”remote visit” is to be annotated. An exception are abbreviations or a clear synonym of the intervention. E.g. in ”Brindley technique (anterior sacral root stimulation with posterior

rhizotomy) is the only technique" both "Brindley technique" and the defintion in the brackets should be annotated.

3. Prophylaxis and prevention related Interventions should be annotated as "OTHER". E.g. in "safe and efficacious ischemic stroke prophylaxis for [...]." the phrase "ischemic stroke prophylaxis" is to be annotated. This holds only if there is no other more specific intervention stated. E.g in "Migrane prevention using Short Pulswave Therapy", "migrane" should be annotated as CONDITION while the INTERVENTION is "Short Pulswave Therapy".

4. Monitoring and diagnostic procedures should not be annotated as interventions, e.g. in "The aim of this study is to evaluate nocturnal hypertension with 24-hour ambulatory blood pressure [...]" the phrase "24-hour ambulatory blood pressure" is not an intervention.

5. We annotate any interventions that aim at improving the health quality outcomes, even if the population/condition is not of immediate relevance. E.g. in "Evaluation of Computer-based Training to Educate Japanese Physicians in the Methods of Interpreting PET Scans." the terms "Computer-based Training" should be labeled.

6. Words that can not stand alone as a specific intervention outside of the study context should not be annotated, e.g. "stimulation", "rehabilitation" alone should not be included. At the same time "rehabilitation treatment" should be annotated. An exception should be made if the generic word is the only mention of the tested intervention in the text.

7. Both umbrella terms, and more specific annotations (if eligible) should be annotated, e.g. If those two terms appear in different positions of the sentence, "rehabilitation treatment [...] yoga exercise", both need to be annotated. Equally valid in "Mitoxantrone (MITO, Novantronae), a synthetic anthracenedione approved for [...]", both "Mitoxantrone" and "anthracenedione" should be annotated.

8. If the intervention is part of an accepted therapeutic regiment, e.g. "radio-chemotherapy", all involved interventions need to be annotated as such. E.g. In "study will evaluate whether the dosage of 1500 mg/m2 of capecitabine is tolerable after radiation" both "capecitabine" (DRUG) and "radiation" (RADIOTHERAPY) should be annotated.

What to annotate?

I. DRUG

1. Below are general guidelines for Chemical annotation that should help identify entities for annotation. Chemicals' fine-granular types are described in Table 1 and Fig. 2. They are to be annotated with the single label **DRUG**. :

   (a) Chemical Nouns convertible to:

   -A single chemical structure diagram: single atoms, ions, isotopes, pure elements and molecules such as: Calcium(Ca), Iron(Fe), Lithium (Li),Potassium(K), Oxygen(O2),

   -A general Markush diagram with R groups such as: Amino acids

   (b) General class names where the definition of the class includes information on some structural or elemental composition such as: steroids, sugars, fatty acids, saturated fatty acids

(c) Small Biochemicals

- Monosaccharides, disaccharides and trisaccharides: Glucose, Sucrose...

- Peptides and proteins with less than 15 aminoacids: Angiotensin II...

- Monomers, dimmers, trimmers of nucleotides: e.g. ATP, cAMP...

- Fatty acids and their derivatives excluding polymeric structures. e.g. Cholesterol, glycerol, prostaglandin E1

(d) Synthetic Polymers such as: Polyethylene glycol

(e) Special chemicals having well-defined chemical compositions. E.g. “ethanolic extract of Daucus carota seeds (DCE)”; “grape seed proanthocyanidin extract”

(f) Other substances, that cannot be associated to a clear molecular structure, such as Olive Oil, Herbal Extracts, Cannabis, Tea, are to be annotated as **OTHER**.

2. For combined drugs, mark them separately, e.g. ”levodopa/carbidopa” should be two entities ”levodopa” and ”carbidopa”.

3. Chemicals that are compared in a study and separated with a ”vs” should be annotated separately, e.g. ”GLP-1 analogues vs DPP4 inhibitors for the treatment of type 2 diabetes mellitus”.

4. Annotate all mentions of a chemical entity in an abstract.

5. Annotate the word ”Vaccine” together with the immunogenic component.

6. Annotate abbreviations. Some abbreviations are ambiguous by convention. Take “Nitric Oxide (NO)” as an example, “NO” could also be interpreted as a negative response. Ambiguity should be avoided using context, i.e. in this case ”NO” should not be annotated.

7. If a DRUG mention is present that is already part of the patient treatment (but is not the primary target of investigation), it should still be label as DRUG, as it is part of the overall treatment.

II. Other interventions

The below mentions represent individual labels.

1. **BEHAVIOURAL**, e.g. meditation, cognitive behavioural therapy, or other education related interventions.

2. **SURGICAL** (incl. tissue-based therapy), e.g. organ transplantation, stem cell transplantation. Injections and transfusions do not fall into this category and should be annotated as ”OTHER” instead.

3. **RADIOTHERAPY**, e.g. proton beam therapy, radioactive iodine.

4. **PHYSICAL**, interventions requiring active participation from the study population e.g. cardiovascular strengthening. In case the intervention does not clearly state that active participation is required, but it could involve it based on the intervention description, the label PHYSICAL should be used, e.g. ”Kinesiology”.

5. **OTHER**, other types of interventions that should be annotated in a more inclusive/broad way e.g. gluten-free diet, clear liquid diets, gene therapy, Virtual Reality, medical massage. An example for a broad inclusion is ”Ultrasound-guided Erector Spinae Plane Block”.

6. **CONTROL**, The most specific mention of the control interventions should be annotated, e.g. in ”sham product (vitamins)” the word ”vitamins” should be annotated. However if there is no specific mention, general words such as ”placebo”, ”sham product” should be labeled. Drugs should be annotated as drugs even if they are a control intervention. If in doubt about whether something is a control intervention, annotate as ”Other” (or the respective intervention class). e.g., ”Test catheters compared to SL catheters”.

NER Tag	Description	Examples
SYSTEMATIC	Systematic names of chemical mentions, e.g. IUPAC and IUPAC-like names.	2-Acetoxybenzoic acid 2-Acetoxybenzenecarboxylic acid 2-Acetoxybenzoic acid N-(4-hydroxyphenyl)acetamide
IDENTIFIER	Database identifiers of chemicals: CAS numbers, PubChem identifiers, registry numbers and ChEBI and CHEMBL ids	CAS Registry Number: 501-36-0445154 CID 445154 CHEBI:28262 CHEMBL504
VACCINE	Vaccine trade names or abbreviations.	BioThrax (AVA), SPIKEVAX (1vCOV-mRNA), Gardasil (9vHPV)
FORMULA	Mentions of molecular formula, SMILES, InChI, InChIKey	CC(=O)Oc1ccccc1C(=O)O InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12) C9H8O4
TRIVIAL	Trivial, trade (brand), common or generic names of compounds.	Aspirin Acylpyrin paracetamol acetaminophen Tylenol Panadol resveratrol
ABBREV	Mentions of abbreviations and acronyms of chemicals compounds and drugs	DMSO GABA
FAMILY	Chemical families that can be associated to some chemical structure are also included	Iodopyridazines (FAMILY- SYSTEMATIC) diphenols (FAMILY- SYSTEMATIC) quinolines (FAMILY- SYSTEMATIC) terpenoids (FAMILY- TRIVIAL) ROH (FAMILY- FORMULA)
MULTIPLE	Mentions that do correspond to chemicals that are not described in a continuous string of characters. This is often the case of mentions of multiple chemicals joined by coordinated clauses.	thieno2,3-d and thieno3,2-d fused oxazin-4-ones
BIOMOLECULE	Large oligomeric and polymeric or established DNA/RNA/protein sequences.	Proteins, polypeptides, nucleic acid polymers, polysaccharides, oligosaccharides, e.g. Insulin, DNA, mRNA, collagen, starch, cellulose, glycogen, lipopolysaccharide, glucocorticoid, glucagon, monoclonal antibodies

Table 1: Chemical Entity Mention (CEM) classes as defined for the CHEMDNER-patent task [5].

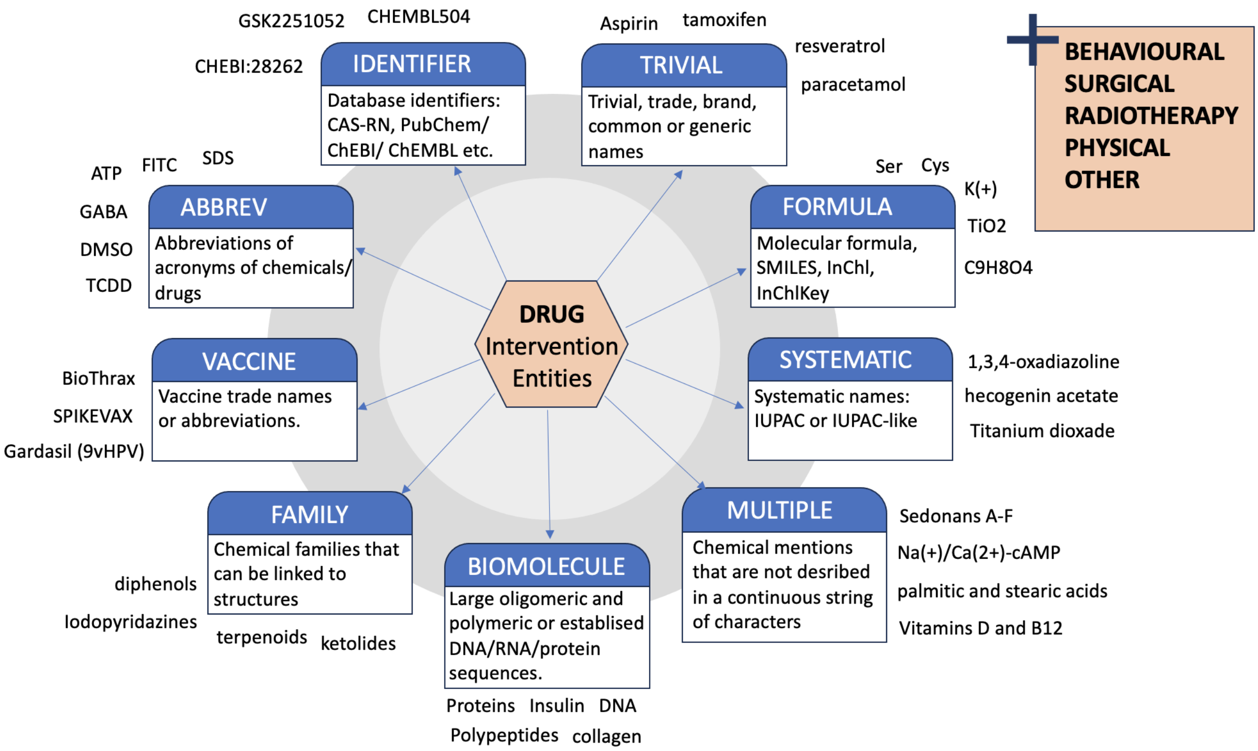

GSK2251052
CHEMBL504
CHEBI:28262
IDENTIFIER
Database identifiers: CAS-RN, PubChem/ ChEBI/ ChEMBL etc.
Aspirin
tamoxifen
resveratrol
paracetamol
TRIVIAL
Trivial, trade, brand, common or generic names
BEHAVIOURAL
SURGICAL
RADIOTHERAPY
PHYSICAL
OTHER
ATP
FITC
SDS
GABA
DMSO
TCDD
ABBREV
Abbreviations of acronyms of chemicals/ drugs
Ser
Cys
K(+)
TiO2
C9H8O4
FORMULA
Molecular formula, SMILES, InChI, InChIKey
DRUG Intervention Entities
BioThrax
SPIKEVAX
Gardasil (9vHPV)
VACCINE
Vaccine trade names or abbreviations.
SYSTEMATIC
Systematic names: IUPAC or IUPAC-like
1,3,4-oxadiazoline
hecogenin acetate
Titanium dioxade
FAMILY
Chemical families that can be linked to structures
diphenols
Iodopyridazines
terpenoids
ketolides
BIOMOLECULE
Large oligomeric and polymeric or establised DNA/RNA/protein sequences.
Proteins
Insulin
DNA
Polypeptides
collagen
MULTIPLE
Chemical mentions that are not desribed in a continuous string of characters
Sedonans A-F
Na(+)/Ca(2+)-cAMP
palmitic and stearic acids
Vitamins D and B12

Figure 2: Overview of chemical-based interventions, adapted from [5] and other types of interventions of interest.

What not to annotate?

1. Do NOT annotate words that describe *how* an intervention is delivered, unless it is an essential part of the intervention. For example *Household Water Treatment Device* in ”Trial of a Household Water Treatment Device as a Delivery System for Zinc in Zinc Defficient children.” should NOT be annotated, while *computer-guided interpositional sandwich osteotomy* should be annotated in ”The aim was to assess the efficiency of the computer-guided interpositional sandwich osteotomy [...].” Other examples include ”Vitamin B (DRUG) supplement (not annotated)”, ”THC (DRUG) infusion (not annotated)”

2. Do NOT annotate other terms different from chemical nouns. Adjective forms of chemical names are also excluded. For instance, muscarinic, adrenergic and purinergic.

3. Do NOT annotate chemical nouns named for a role or similar, that is, nonstructural concepts (e.g. anti-HIV agents, anticonvulsants, anticholinesterase drug, antipsychotic, anticoagulant, etc).

4. Do NOT annotate very nonspecific structural concepts.e.g. Atom, Ion, Molecular, Lipid, Protein. Exception is when some of these workds are part of a longer specific chemical name, e.g. ”chloride ion”, ”thiol dimers”.

5. Do NOT annotate words that are not chemicals in context, even if they are co-incidentally the same set of characters (synonyms and metaphors). For instance,“Gold” should not be annotated if it appears in “gold standard.” This applies also to general drug names, e.g. cellulose, glucocorticoid.

6. Do NOT annotate general vague compositions. For instance, according to Wikipedia, the term opiate describes any of the narcotic opioid alkaloids found as natural products in the opium poppy plant, Papaver somniferum, and thus should be excluded.

7. Do NOT annotate special words not to be labeled by convention (e.g. Water, saline, juice, etc).

8. Do NOT tag acronyms that are of 1 letter in length.

9. Do NOT include trademark symbols, e.g. Mesupron®should result in the annotation ”Mesupron”.

References

[1] Nci dictionary of cancer terms. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/intervention`. Accessed: 2023-05-15.

[2] Calvo, F., Karras, B. T., Phillips, R., Kimball, A. M., and Wolf, F. Diagnoses, syndromes, and diseases: a knowledge representation problem. In *AMIA annual symposium proceedings* (2003), vol. 2003, American Medical Informatics Association, p. 802.

[3] Huang, X., Lin, J., and Demner-Fushman, D. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings* (2006), vol. 2006, American Medical Informatics Association, p. 359.

[4] Kottow, M. H. A medical definition of disease. *Medical Hypotheses 6*, 2 (1980), 209–213.

[5] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., and et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics 7* (1 2015), 1–17.

[6] Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation 2016* (2016), 68.

[7] Montani, I., and Honnibal, M. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.