The Annotation Guidelines for Concept in Eligibility Criteria according to OMOP CDM v5

Tian Kang, Gregory W. Hruby, Alexander Rusanov
April 2016

Overall Goal
To develop a machine learning-based parser for clinical trial eligibility criteria texts; use the parser to extract concepts and their clinical relations from the free-text eligibility criteria and re-represent the information in a structured format.

Subtasks
1. Named Entity Recognition (NER)
2. Syntactic relations identification
3. Negation/Hedge detection
4. Semantic relations identification

This guideline is designed for the first task: **Named Entity Recognition**, recognizing entities (here, means clinical meaningful concepts) in the eligibility criteria texts.

Introduction to this guideline
Named Entity Recognition (NER) refers to the task of information extraction that seeks to locate and classify concepts in the texts and builds towards the relation extraction task. This guideline describes the specific types of concepts that should be annotated for the NER task in clinical trial eligibility criteria texts.

In this annotation tasks, we focus on seven semantic entities:
- 4 entity classes: conditions, observations, procedure/device, and drug/substance,
- 3 concept attributes: qualifiers, measurement, and temporal constraints.

The definition of each semantic entity in this annotation schema follows the **OMOP Common Data Model Specifications Version 5.0**, in order to match the EHR data modeled using this standard.

Data
Annotators will follow this guideline to annotate clinical trial eligibility criteria texts retrieved randomly from ClinicalTrials.gov to prepare training data for machine learning-based NER task. (The number of trials is determined upon the complexity of the texts, e.g., for nuero-psychological disease, our plan is 250 trials)

Sections in this guideline:

table

*****Example eligibility criteria and concepts in this guideline are all *italic (all examples are from clinicaltrials.gov studies in on neuropsychological diseases).*

Overviews of the Concept categories:

For OMOP definition, see attached table 1

ENTITY categories:

- **Condition**

a disease or a medical condition determined by a provider or reported by a patient.

- **Observation**

any clinical fact about a patient obtained in the context of examination, questioning or a procedure.

- **Procedure/Device**

Procedure: an activity or process ordered by and/or carried out by a healthcare provider on the patient that has a diagnostic and/or therapeutic purpose.
Device: inferred exposure to a foreign physical object or instrument used for diagnostic or therapeutic purposes through a mechanism beyond chemical action. Devices include implantable objects (e.g. pacemakers, stents, artificial joints), durable medical equipment and supplies (e.g. bandages, crutches, syringes), and other instruments used in medical procedures (e.g. sutures, defibrillators).

- **Drug/Substance**

inferred utilization of a biochemical substance with a physiological therapeutic effect when ingested or otherwise introduced into the body. Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies (for ex., …) . Drug exposure is inferred from clinical events associated with orders, prescriptions written, pharmacy dispensing, procedural administrations, and other patient- reported information.

ATTRIBUTES categories:

- **Measurement**

structured value (here we focus on numerical values) obtained through systematic examination of a person or sample. The MEASUREMENT table captures measurement orders and measurement results. The measurement domain can contain laboratory results, vital signs, or quantitative findings from pathology reports.

- **Temporal coAnstraints**

Temporal records which uniquely define the spans of time for the concepts from entity classes (condition, observation, drug/substance, procedure/device)

- **Qualifiers/Qualifiers**

Pre-coordination and post-coordination of the concepts from entity classes

- **(Anatomic location,** not included in neuropsychological disease annotation**)**

Detailed description of each Concept category

Note: the example concepts are noted in italic. For example criteria, the concepts in each category is marked in red.

1. Condition:

1) Phrases that name a disease, syndrome, sign, or symptom
   - *Alzheimer's Disease; AD*
   - *medical or psychiatric illness*
   - *frontotemporal dementia*
   - *congenital long Q-T syndrome.*
   - *conduction defect abnormalities*
   - *Lesions*
2) Phrases that describe health status
   - *healthy/unhealthy*
   - *Pregnancy; pregnant*
   - *childbearing potential*
3) Condition that need to be specified by concepts from other categories
   - *Contraindication* (to Procedure concepts)
   - *Allergy/allergic* (to Drug/sub concepts)
   - Example criterion*:*
     - *Contraindication to undergoing MRI (e.g., pacemaker)*
4) Abnormal clinical findings:
   - *lacunes*
   - *abnormality*
   - Example criterion:
     - *Clinically significant abnormality on blood test.*

2. Observation:

1) Lab test variables:
   - *serum pregnancy test*
   - *HIV, HBV test*
   - *hepatitis C virus antibodies; hepatitis B surface antigen*
   - Criteria examples*:*
     - *hemoglobin (Hgb) < 10.0 g/dL*
2) Standard evaluation matrices for diagnostic purpose
   - *Clinical Dementia Rating (CDR) total score; CDR*
   - *Mini-Mental State Examination; MMSE*
   - *Folstein mini-mental status examination total score*
   - Criteria examples*:*
     - *MMSE score between 18 and 26 (inclusive)*
3) Physical exam
   - *systolic blood pressure*
   - *Visual and auditory acuity*
   - Criteria examples*:*
     - *Visual and auditory acuity adequate*
4) Standard criteria to classify the disease or degree of disease

- *DSM-IV criteria*
- *DSM-IV-R and NINCDS-ADRDA criteria*
- *NINCDS-ADRDA* 1984 criteria*
- Criteria examples*:*
  - *Possible AD by NINCDS-ADRDA criteria*
  - *Congestive heart failure (New York Heart Association Class III or IV)*

3. Procedure/Device:

1) Procedure: Process or activities carried out by healthcare providers for therapeutic purpose.
   - Surgery
   - Scanning: *MRI*, *CT*, *PET scan*
   - Examination*: ECG; EKG*
   - Treatment using machine or device: *Radiotherapy; lumbar puncture*
2) Health related activities:
   - *blood donation*
     - *Donation of blood in excess of 500ml within a 56-day period*

3) Drug dosage
   - *Started or changed within 60 days prior to the screening visit the dosage of any drug*

3) Device: Foreign physical object or instrument that is used for diagnostic or therapeutic purposes.
   - *Metal implants; pacemaker; cochlear implants*
   - *drug-eluting stents; coronary stent*

4. Drug/Substance:

1) Medications:
   - *trazodone; atypical antipsychotics*
   - *typical antipsychotics*
2) Therapy or treatment with medication
   - *Use of any investigational therapy within 30 days or 5 half-lives, whichever is longer, and/or use of AD immunotherapy prior to screening.*
   - If the syntax is "treatment with X medications", only mark the medications:
     Treatment with memantine
3) Substance exposure:
   - *Current or recent participation in any procedures involving radioactive agents, including current, past, or anticipated exposure to radiation in the workplace.*

5. **Measurement:**

1) Lab test numeric results
   - *Neutropenia defined as absolute neutrophils count of < 1,500/microliter*
2) Drug/substance dosage
3) Scores for standard evaluation or criteria
   - *Mini-Mental State Examination (MMSE) 16-26 inclusive at the time of screening.*
4) Condition status numeric degree
   - *Congestive heart failure (New York Heart Association Class III or IV)*
5) Numeric value for other observations
   - *Males and females between the ages of 50-85*

6. Temporal constraints

1) Describe drug dosage stability: (*for* is not included in annotation schema)
   - *Stable doses of acetyl cholinesterase inhibitors for at least 3 months*
   - *Cholinesterase inhibitors taking a minimum of four weeks or more stable subject*

2) Describe an activity or condition happening in a certain period of time: (*within/in/during* is included because it defines a constraint. Usually the criterion will define the time period by using "prior to" some event, if it's left out, we assume it's prior to enrollment in the trial)
   - *Electroconvulsive Therapy (ECT) within 6 months*
   - *Vitamin B12 deficiency within one year prior to enrollment;*
   - *Any medication for Alzheimer's Disease in/during past 3 months*

3) Common qualitative temporal constraints:
   - *current; currently; concurrent; concomitant*
   - *Previously, previous; prior*
   - *recent; in the past*

4) When a temporal constraint is described in more than one way in the same noun phrase (such as a qualitative term + a numeric constraint; or time period length+ drug half-lives), it is considered a single constraint:
   - *Current or recent (past 6 months) alcohol or substance dependence*
   - *Have taken other investigational drugs within 30 days or 5 half-lives prior to randomization*

7. Qualifiers/Modifiers

1) Common qualifiers/modifiers:
   - **Describe condition status**: *major*; *active/chronic*; *uncontrolled/well-controlled; unstable/stable; untreated/treated; successfully treated* …
   - **Describe condition severity**: *serious; severe; mild; moderate; significant; clinically significant; probable; possible…*
   - **Describe qualitative status of lab test or examination results**: positive/negative; *high/low; normal/abnormal; abnormally high; higher; increased/decreased*
   - Define **status of drug dosage** as **qualifier/modifiers** (only in this case, the verbs is allowed to be annotated)*: stable; change; discontinuation; started…*

-They started or changed within 60 days prior to the screening visit the dosage of any drug
Change in dose or discontinuation of a drug

define granularity of modified concepts – follow the "longest concept" rule: define the 4 major categories of concepts (Condition, Observation, Drug/Substance, Procedure/Device) as the terms with longest length that can be found in UMLS (use OHDSI tool ATLAS http://www.ohdsi.org/web/atlas or UMLS online application https://uts.nlm.nih.gov/metathesaurus.html to search concepts).

E.g. *modifier/qualifier + concepts (condition/drug/procedure/observation)*
   - *Uncontrolled diabetes or hypertension*
   - *Severe AD*
     ("AD", UMLS CUI: C0002395*)*
   - *Active or uncontrolled epilepsy, Active hypothyroidism*
   - *Mild to moderate Alzheimer's disease*

("Alzheimer's disease", UMLS CUI: C0002395)

Need to notice:

- *major depression; major surgery within 6 months*

"major depression" (UMLS CUI C1269683) is marked together because it's a common description and an UMLS concept, even "*major*" is a common modifier; if not sure, lookup in UMLS or ATLAS)

- *major psychological illness; major disease*

For "major disease" the longest UMLS concept we can find in this term is "disease". *"major"* here is a modifier, defining status of the disease.

2) When adjacent qualifiers/modifiers are homogeneous, combine them into one annotation, otherwise, split into different ones.

- *Active or uncontrolled epilepsy, Active hypothyroidism* (active and uncontrolled are heterogeneous)
- *Mild to moderate Alzheimer's disease* (mild and moderate are homogeneous)
- *an advanced, severe, progressive or unstable medical condition.*
- *low normal white blood cell counts*

3) Qualifiers/Modifiers can be both pre- and post-coordinations of the concepts they define.

- *Abnormally high or low serum levels of thyroid stimulating hormone (TSH) that is clinically significant in the opinion of the investigator*

General annotation rules:

1. Only complete noun phrases (NPs) and adjective phrases (APs) should be annotated (no verbs). Examples?

2. Not all criteria need to be annotated. If the eligibility that criterion defines is not in medical records, then this criterion will be skipped, even there are concepts in it that fit the concept semantic rules. For example:

- Criteria describing informed consent and compliance with the study requirements.

(e.g.) *-Written informed consent obtained and documented*

-Informant available with frequent (at least 1 hour/day or 1 day/week) contact with subject to verify functional status and CDR rating.

-Have the ability to cooperate and comply with all study procedures

- Criteria describing language and education.

- English as native language.

->= 8 years education

- Criteria describing care givers or study partners.

-Must have a reliable and capable caregiver who has regular interaction with them, will be present for all visits, can provide a collateral history, can assist in compliance with study procedures and who is willing to act as the Study Partner (provide written informed consent) and remain unaware of the results.

-Reliable and capable caregiver, who has regular interaction with them, will be present for all visits; can provide a collateral history,

- Criteria describing the willingness or agreement of the participants.

-Willing to undergo study procedures and remain unaware of the results

-Subject must agree to use effective contraceptive measures

- Criteria describing the contraception and birth control

-Require use of contraception during the course of study

-Both men who are able to father children and women of childbearing potential must be willing to use an adequate method of contraception (see section 7) to avoid conception throughout the study and for up to 30 days of study drug administration

3. Conjunctions and other syntax that denote lists should be included as one annotated concept if they occur within the modifiers or connected by a common set of modifiers. If the portions of the lists are otherwise independent, they should be different concepts. (ref i2b2 guideline)

- *hepatic/renal disorders* (is one concept in "Condition" category)
- *Any hepatic, cardiovascular, gastrointestinal or hematological illness.*
- *uncorrectable loss of hearing or eyesight*

4.When concepts are mentioned in more than one way in the same noun phrase (such as the definition of an acronym or where a generic and a brand name of a drug are used together), the concepts should be marked together. (ref i2b2 guideline)

E.g.

- *Alzheimer's disease (AD)*
- *atrio-ventricular [AV] block*

5. If a term with a low granularity (such as "medical condition", "medication") and followed by several more specific examples, both term and examples need to be marked.

- *Presence of diseases (systemic and/or ocular disease).*
- *Currently use of any of the following medications: Aricept, Cognex, Exelon, Reminyl or Hydergine.*

6. If a concepts or attribution is separated by other words, mark them separately:

- *Patient is male or female and at least 50 years of age or older.*

7. When there are two

- *A score of >= 8 on the Cornell Scale for Depression in Dementia*

8. Some pre-defined patterns to avoid ambiguity:

1) *able to/unable to/ability/inability/ intolerability:* When those terms refer to ability to perform some procedures or tolerability to some drugs, classify them to Observation class. (Notice: most of the descriptions in ability occur in the criteria that will be left out)

- *ability to tolerate a brain scan.*

2) Gender: Gender will not be marked if the criteria are not distinguished between different genders. Otherwise, genders will be classified to the Observation class.

- *Males or females aged 60 years or older.*
- *12 -lead ECG (including QTc greater than 450 millisecond for males and females.*
- *Clinically significant anaemia (i.e. haemoglobin < 11 g/dL for males or < 10 g/dL for females)*

3) Declare some specific ambiguity among different semantic categories:

	Classified to
smoking:	Observation
Contraindication/Allergy	Condition
complaints	Condition
pregnant/pregnancy	Condition
lacune :	Condition
microhemo/hemo :	Condition
lesion / focal lesion:	Condition
physical exam :	Observation
suicidal thoughts/ideation:	Condition
therapy (if it's a drug therapy)	Drug/Substance
Abnormality	Condition
dose/dosage	Procedure
HIV, HBV…	

- When referring to a test →Observation
  Positive hepatitis B virus, hepatitis C virus or HIV test at screening.
  Positive RPR or HIV
- When referring to a disease →Condition
  Subjects with HIV disease/infection
  Hepatitis B, C, HIV or Syphilis.

6) Criteria about drug dosage

- Predefined rules:
- Dosage belongs to Procedure category
- Drug belongs to Drug/Substance category
- Dosage status belongs to qualifier/Modifier category (e.g. *stable, stability, start, change, termination*)
- Temporal description on dosage status belongs to Temporal constraints category

    - *Started or changed within 60 days prior to the screening visit the dosage of any drug*
    - *The dosage will likely remain stable throughout the trial*

Concept class	Corresponding CDM table	Definition from OMOP CDM v5	Examples (Concepts are marked as RED)
Condition	CONDITION_OCCURRENCE	The CONDITION_OCCURRENCE table captures records of a disease or a medical condition based on evaluation by a provider or reported by a patient.	Dementia or mini-mental state exam < 24; Appreciable accent schizophrenia, bipolar disorder or major depression;
Observation	OBSERVATION	The OBSERVATION table captures any clinical facts about a patient obtained in the context of examination, questioning or a procedure. The observation domain supports capture of data not represented by other domains, including unstructured measurements, medical history and family history.	Dementia or mini-mental state exam (MMSE) < 24; 50 - 90 years of age; Serum Creatinine > 1.5 mg/dL; Serum Glucose > 150mg/dL; Meets the National Institute on Aging-Alzheimer's Association criteria for amnestic or multi-domain MCI
Procedure/ Device	PROCEDURE_OCCURRENCE DEVICE_EXPOSURE	*The PROCEDURE_OCCURRENCE table contains records of activities or processes ordered by and/or carried out by a healthcare provider on the patient to have a diagnostic and/or therapeutic purpose. *The DEVICE_EXPOSURE table captures records about a person's inferred exposure to a foreign physical object or instrument that is used for diagnostic or therapeutic purposes through a mechanism beyond chemical action. Devices include implantable objects (e.g. pacemakers, stents, artificial joints), durable medical equipment and supplies (e.g. bandages, crutches, syringes), and other instruments used in medical procedures (e.g. sutures, defibrillators).	History of intra-vitreal injections; Contraindication to an MRI procedure, (i.e., pacemaker, metal plates)
Drug/Substance	DRUG_EXPOSURE	The DRUG_EXPOSURE table captures records about the inferred utilization of a biochemical substance with a physiological therapeutic effect when ingested or otherwise introduced into the body. Drugs include prescription and over-the- counter medicines, vaccines, and large-molecule biologic therapies. Drug exposure is inferred from clinical events associated with orders, prescriptions written, pharmacy dispensing, procedural administrations, and other patient- reported information.	Sedative hypnotics more frequently than 2 times per week within 4 weeks prior to screening; Medications with known significant cholinergic or anticholinergic side effects (such as pyridostigmine, tricyclic antidepressants, meclizine, oxybutynin)

Measurement	MEASUREMENT	A measurement is the capture of a structured value (numerical or categorical) obtained through systematic examination of a person or sample. The MEASUREMENT table captures measurement orders and measurement results. The measurement domain can contain laboratory results, vital signs, or quantitative findings from pathology reports.	Serum Creatinine > 1.5 mg/dL; Serum Glucose > 150mg/dL; Sedative hypnotics more frequently than 2 times per week within 4 weeks prior to screening;
Temporal Constraints	DRUG_ERA DOSE_ERA CONDITION_ERA OBSERVATION_PERIOD	*A Drug Era is defined as a span of time when the Person is assumed to be exposed to a particular active ingredient. *A Dose Era is defined as a span of time when the Person is assumed to be exposed to a constant dose of a specific active ingredient. *A Condition Era is defined as a span of time when the Person is assumed to have a given condition. *The OBSERVATION_PERIOD table contains records which uniquely define the spans of time for which a person is at-risk to have clinical events recorded within the source systems.	Sedative hypnotics more frequently than 2 times per week within 4 weeks prior to screening; Subject has used any tobacco products in the past 3 months
Qualifier	(not in the CDM definition)	(Including modifiers of the basic concepts and qualitative measurements)	Positive urine test; Clinically significant psychiatric condition; Unstable symptomatic CVD; Low neutrophil count (< 3 x 10^9/L)