The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria

Nicholas J. Dobbins1✉, Tony Mullen2, Özlem Uzuner3 & Meliha Yetisgen1

Identifying cohorts of patients based on eligibility criteria such as medical conditions, procedures, and medication use is critical to recruitment for clinical trials. Such criteria are often most naturally described in free-text, using language familiar to clinicians and researchers. In order to identify potential participants at scale, these criteria must first be translated into queries on clinical databases, which can be labor-intensive and error-prone. Natural language processing (NLP) methods offer a potential means of such conversion into database queries automatically. However they must first be trained and evaluated using corpora which capture clinical trials criteria in sufficient detail. In this paper, we introduce the Leaf Clinical Trials (LCT) corpus, a human-annotated corpus of over 1,000 clinical trial eligibility criteria descriptions using highly granular structured labels capturing a range of biomedical phenomena. We provide details of our schema, annotation process, corpus quality, and statistics. Additionally, we present baseline information extraction results on this corpus as benchmarks for future work.

Background & Summary

Randomized controlled trials serve a critical role in the generation of medical evidence and furthering of biomedical research. In order to identify patients for clinical trials, investigators publish eligibility criteria, such as past history of certain conditions, treatments, or laboratory tests. These eligibility criteria are typically composed in free-text, consisting of inclusion and exclusion criteria. Patients meeting a trial's eligibility criteria are considered potential candidates for recruitment.

Recruitment of participants remains a major barrier to successful trial completion1, so generating a large pool of potential candidates is often necessary. Manual chart review of hundreds or thousands of patients to determine a candidate pool, however, can be prohibitively labor- and time-intensive. Cohort discovery tools such as Leaf2 and i2b23 may be used, providing researchers with a relatively simple drag-and-drop graphical interface in their web browser to create database queries to find potential patients in electronic health records (EHR). Learning how to use such tools nevertheless presents a challenge, as graphically-represented concepts may not align with researchers’ understanding of biomedical phenomena or trial eligibility criteria. In addition, certain complex queries may simply be impossible to execute due to structural limitations on the types of possible queries presented in these tools, such as complex temporal or nested sequences of events.

An alternative approach which holds promise is the use of natural language processing (NLP) to automatically analyze eligibility criteria and generate database queries to find patients in EHRs. NLP-based approaches have the advantage of obviating potential learning curves of tools such as Leaf, while leveraging existing eligibility criteria composed in a format researchers are already familiar with. Recent efforts to explore NLP-based approaches to eligibility criteria query generation have been published. These approaches can be generally categorized as (1) modular, multi-step methods which transform eligibility criteria into intermediate representations and finally into queries using rules4–6, (2) direct text-to-query generation methods using neural network-based semantic parsing5,6, and (3) information retrieval approaches to detect relevant sections of free-text notes meeting a given criteria7–10.

For each category of NLP approaches, a key element for accelerating research efforts is large, robust corpora which capture eligibility criteria semantics sufficiently for high-accuracy query generation. Such corpora can

1Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, WA, USA. 2Khoury College of Computer Science, Northeastern University, Seattle, WA, USA. 3Department of Information Sciences and Technology, George Mason University, Fairfax, VA, USA. ✉e-mail: ndobb@uw.edu
serve as reliable benchmarks for purposes of comparing NLP methods as well as training datasets. A number of
corpora designed for multi-step methods, which we focus on here, have been published. Past corpora cover only
a modest number of eligibility criteria, are narrowly focused on certain diseases only, are not publicly avail-
able, or have annotations insufficiently granular to fully capture the diverse, nuanced semantics of eligibility
criteria. Yu et al. released a corpus designed for direct text-to-query generation with semantic parsing, how-
ever given the relative simplicity of generated queries to date compared to the complexity of clinical databases,
it’s not clear this approach is yet viable for real-world clinical trials recruitment.

In this paper, we present the Leaf Clinical Trials (LCT) corpus. To the best of our knowledge, the LCT corpus
is the largest and most comprehensive human-annotated corpus of publicly available clinical trials eligibility
criteria. The corpus is designed to accurately capture a wide range of complex, nuanced biomedical phenomena
found in eligibility criteria using a rich, granular annotation schema. As the LCT annotation schema is uniquely
large, fine-grained and task-oriented, the corpus can serve as a valuable training dataset for NLP approaches
while significantly simplifying disambiguation steps and text-processing for query generation. The LCT annota-
tion schema builds upon the foundational work of EliIE, an Information Extraction (IE) system for eligibility
criteria, and Chia, a large corpus of clinical trials of various disease domains. Expanding the EliIE and Chia
annotation schemas, we developed the LCT annotation schema to greatly increase the variety of biomedical
phenomena captured while also annotating eligibility criteria semantics at a significantly more granular level.

Table 1 presents a comparison of the LCT corpus and these corpora.

| Measure                  | EliIE12 | Chia15 | LCT Corpus |
|--------------------------|---------|--------|------------|
| Disease domain           | Alzheimer’s Disease | All | All |
| No. of Eligibility Descriptions | 230     | 1,000  | 1,006 |
| No. of Annotations       | 15,596  | 68,174 | 105,816 |
| No. of Entity types      | 8       | 15     | 50         |
| No. of Relation types    | 3       | 12     | 51         |
| Mean Entities per doc.   | —       | 46     | 105        |
| Mean Relations per doc.  | —       | 19     | 49         |

Fig. 1 Example eligibility criteria annotated used the LCT corpus annotation schema (left) and corresponding
example SQL queries (right) using a hypothetical database table and columns. Annotations were done using the
Brat annotation tool. The ICD-10 codes shown are examples and not intended to be exhaustive.

Methods

Eligibility criteria and database queries. The NLP tasks involved in transforming eligibility criteria into
database queries include named entity recognition (NER) to tag meaningful spans of text as named entities, relation
extraction to classify relations between named entities, normalization to map named entities to common
coded representations (e.g., ICD-10), negation detection to detected negated statements (e.g., “not hyperten-
sive”) and so on. Gold standard corpora quality can thus directly affect performance and the validation of each
of these tasks. In this article, we only focus on design and development of the LCT corpus. Figure 1 illustrates
why corpora structure and integrity are important for the task of query generation, using examples of eligibility
criteria annotated using the LCT annotation schema and corresponding hypothetical Structured Query Language
(SQL) queries. In the first eligibility criterion, “preeclampsia” is explicitly named, and thus can be directly nor-
malized to an International Classification of Diseases-Tenth Revision (ICD-10) code. In the second criterion, “diseases”
in “diseases that affect respiratory function” is non-specific, and must be reasoned upon in order to determine appropriate codes, such as asthma,
chronic obstructive pulmonary disease (COPD), or emphysema. Programmatically reasoning to generate queries in such cases would be challenging and often impossible if the underlying semantics were not captured appropriately. With this in mind, we developed the LCT annotation schema in order to enable reasoning and ease query generation for real-world clinical trials use. As the second example in Fig. 1 shows, the LCT annotation captures the semantics of complex criteria, with changes to “respiratory function” annotated using a Stability[change] entity and Stability relation, and the cause, “diseases” annotated with a Caused-By relation. During query generation, a hypothetical algorithm can thus use LCT entities and relations to first normalize the span “respiratory function”, then reason that asthma, COPD, emphysema, and other conditions can affect respiratory function and thus the generated query should find patients with those diagnoses.

**Annotation schema.** We aimed to develop an expressive, task-oriented annotation schema which could capture a wide range of medical concepts and logical constructs present in eligibility criteria. To accomplish this, we first analyzed previously published corpora and expanded the list of included biomedical phenomena to fully capture the context and logic present in real clinical trials criteria. As one example, we introduced an entity called Contraindication to reflect where use of a given treatment is inadvisable due to possible harm to the patient.

The LCT annotation schema is designed with the following goals and assumptions:

1. The annotation schema should be **practical** and **task-oriented** with a focus on facilitating ease of query generation.
2. A greater number of **more specific, less ambiguous** annotated phenomena should be favored over a smaller number of possibly ambiguous ones.
3. Annotations should be **easily transformable** into composable, interconnected programmatic objects, trees, or node-edge graph representations.
4. The annotation schema should **model eligibility criteria intent and semantics** as closely as possible in order to ensure generated queries can do the same.

The LCT annotation schema is composed of **entities** and **relations**. Entities refer to biomedical, demographic, or other named entities relevant to eligibility criteria, and are annotated as a span of one or more tokens. We organized LCT entities into the following categories:

- **Clinical** - Allergy, Condition, Condition-Type, Code, Contraindication, Drug, Encounter, Indication, Immunization, Observation, Organism, Specimen, Procedure, Provider.
- **Demographic** - Age, Birth, Death, Ethnicity, Family-Member, Language, Life-Stage-And-Gender.
- **Logical** - Exception, Negation.
- **Qualifiers** - Acuteness, Assertion, Modifier, Polarity, Risk, Severity, Stability.
- **Temporal and Comparative** - Criteria-Count, Eq-Comparison (an abbreviation of “Equality Comparison”), Eq-Operator, Eq-Temporal-Period, Eq-Temporal-Recency, Eq-Temporal-Unit, Eq-Unit, Eq-Value.
- **Other** - Coreference, Insurance, Location, Other, Study.

The LCT corpus also includes 7 **Name** entities: Allergy-Name, Condition-Name, Drug-Name, Immunization-Name, Observation-Name, Organism-Name and Procedure-Name. Name entities serve a special purpose in the LCT corpus, as they indicate that a span of text refers to a specific condition, drug, etc., as opposed to any condition or drug. Name entities overlap with their respective general entities. For example, the span “preeclampsia” refers to a specific condition, and would thus be annotated as both a Condition and Condition-Name, while the span “diseases” is non-specific and would be annotated as only Condition. A full listing of the LCT annotation guidelines can be found at [https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki](https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki).

We defined a total of 50 entities in the LCT corpus. Examples of selected representative entities are presented in Table 2. In our representation, a subset of entities have **values** as well. For example, an Encounter may have a value of emergency, outpatient or inpatient. Values are optional in some entities (such as Encounters or Family-Member, where they may not always be clear or are intentionally broad) and always present in others. In the example annotations presented below, values are denoted using brackets (“[…]”) following entity labels.

Relations serve as semantically meaningful connections between entities, such as when one entity acts upon, is found by, caused by, or related in some way to another. We categorize relations into the following:

- **Alternatives and Examples** - Abbrev-Of, Equivalent-To, Example-Of.
- **Clinical** - Code, Contraindicates, Indication-For, Name, Provider, Specimen, Stage, Type.
- **Dependent** - Caused-By, Found-By, Treatment-For, Using.
- **Logical** - And, If-Then, Negates, Or.
- **Qualifier** - Acuteness, Asserted, Dose, Modifies, Polarity, Risk-For, Severity, Stability.
- **Temporal and Comparative** - After, Before, Criteria, Duration, During, Max-Value, Min-Value, Minimum-Count, Numeric-Filter, Operator, Per, Temporal-Period, Temporal-Recency, Temporal-Unit, Temporality, Unit, Value.
- **Other** - From, Except, Has, Is-Other, Location, Refers-To, Study-Of.

We defined a total of 51 relations in the LCT corpus. Examples of relation criteria are shown in Table 3.
In our annotations, some entity spans overlap with other entity spans in order to fully capture complex underlying semantics. Consider for example, the expression "Ages 18–55 years old". While an Age entity may be assigned to token "Ages", if an Eq-Comparison entity alone were assigned to the span "18–55 years old", the underlying semantics of the tokens "18", "−", "55", and "years" would be lost. In the following examples, we use the term fine-grained entity to refer to entities which are sub-spans of other general entities. Fine-grained entities are linked to general entities by relations. We use down arrow symbols (↓) to denote entity annotation and left and right arrow symbols (← and →) to denote relations. The (+) symbols denote overlapping entities on the same span.

The example expression “Ages 18–55 years old” would be annotated in three layers. In the first layer, the expression is annotated with Age and Eq-Comparison general entities with a relation between them:

```
Ages 18–55 years old
↓
Age Eq-Comparison
```

In the second layer, fine-grained entities with respective values are annotated:

```
"18" "−" "55" "years"
↓ ↓ ↓ ↓
Eq-Value Eq-Operator Eq-Value Eq-Temporal-Unit
```

Table 2. Examples of representative LCT annotation schema entities. A full listing of all entities can be found in the LCT annotation guidelines at https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki.

In our annotations, some entity spans overlap with other entity spans in order to fully capture complex underlying semantics. Consider for example, the expression “Ages 18–55 years old”. While an Age entity may be assigned to token “Ages”, if an Eq-Comparison entity alone were assigned to the span “18–55 years old”, the underlying semantics of the tokens “18”, “−”, “55”, and “years” would be lost. In the following examples, we use the term fine-grained entity to refer to entities which are sub-spans of other general entities. Fine-grained entities are linked to general entities by relations. We use down arrow symbols (↓) to denote entity annotation and left and right arrow symbols (← and →) to denote relations. The (+) symbols denote overlapping entities on the same span.

The example expression “Ages 18–55 years old” would be annotated in three layers. In the first layer, the expression is annotated with Age and Eq-Comparison general entities with a relation between them:

```
Ages 18–55 years old
↓
Age Eq-Comparison
```

In the second layer, fine-grained entities with respective values are annotated:

```
"18" "−" "55" "years"
↓ ↓ ↓ ↓
Eq-Value Eq-Operator Eq-Value Eq-Temporal-Unit
```

| Category     | Entity       | Values     | Example Text                      |
|--------------|--------------|------------|-----------------------------------|
| Clinical     | Condition    | —          | Diagnosed with hypertension in past year |
|              | Contraindication | —         | any contraindications to vaginal delivery |
|              | Drug         | —          | β-blockers Drug                   |
|              | Encounter    | emergency, outpatient, inpatient | recently admitted to a hospital |
|              | Immunization | —          | received influenza vaccination     |
|              | Observation  | lab, vital, clinical-score, survey, social-habit | less than 500 |
|              | Procedure    | —          | Undergoing or scheduled for a colonoscopy |
| Demographic  | Age          | —          | 43 years old                     |
|              | Birth        | —          | born within the past 6 months    |
|              | Family-Member| mother, father, sibling, etc. | history of maternal breast cancer |
|              | Language     | —          | Speaks English or Spanish        |
| Logical      | Negation     | —          | with systemic disease            |
| Qualifier    | Assertion    | intention, hypothetical, possible | which cause conditions |
|              | Modifier     | —          | alcohol or substance abuse       |
|              | Polarity     | low, high, positive, negative | showing serum creatinine |
|              | Risk         | —          | at heightened risk for suicide   |
|              | Severity     | mild, moderate, severe | with complications from surgery |
|              | Stability    | stable, change | conditions known to affect mood |
| Temporal and Comparative | Criteria-Count | — | at least 3 of the following conditions: |
|              | Eq-Comparison | — | greater than 50ml |
|              | Eq-Temporal-Period | past, present, future | Active illness |
|              | Eq-Temporal-Recency | first-time, most-recent | Latest BMI > 35 |
| Other        | Location     | residence, clinic, hospital, unit, emergency-department | Seen at diabetes care clinic |

Table 2. Examples of representative LCT annotation schema entities. A full listing of all entities can be found in the LCT annotation guidelines at https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki.
In the third layer, relations connecting fine-grained entities to the general \(\text{Eq-Comparison}\) entity are added:

\[
\begin{align*}
\text{Value} & \quad \text{Eq-Comparison} \quad \text{Operator} \quad \text{Value} \\
\text{Temporal-Unit} & \quad \text{Eq-Value} \quad \text{“18”} \\
\text{Eq-Value} & \quad \text{“55”} \\
\end{align*}
\]

This multilayered annotation strategy allows significant flexibility in capturing entities and relations in a slot-filling fashion, simplifying the task of downstream query generation. We show examples of this in the Usage Notes section.

The LCT annotation schema contributes the following novel features: (1) deep granularity in entities and relations, which enables (2) rich semantic representation, closely capturing the intent of complex clinical trial eligibility criteria and facilitating accurate query generation.

**Deep entity and relation granularity.** We assume that more specific annotation labels are generally more straightforward to generate accurate queries with. For example, within the span, “preceding six months”, annotating the token “preceding” as \(\text{Temporal}\) (an entity type in Chia) may appear to be adequate, given that an English-speaking human would understand that this refers to the past. Without further information, however, a naïve algorithm would be unable to determine (1) whether such an entity refers to the past, present, or future, (2) that the token “six” refers to a numeric value, and (3) that “months” refers to a unit of temporal measurement.

| Category               | Relation          | Example Annotation                                      |
|------------------------|-------------------|--------------------------------------------------------|
| Alternatives and Examples | Abbrev-Of         | Post Concussion Syndrome (PCS)                        |
|                        | Equivalent-To     | Thrombocytopenia                                        |
|                        | Example-Of        | Condition                                             |
| Clinical               | Contraindicates   | conditions                                            |
|                        | Caused-By         | bleeding due to trauma                                 |
|                        | Found-By          | been seen on standard imaging                         |
|                        | Treatment-For     | coronary bypass surgery for atherosclerosis            |
|                        | Using             | total line arterioplasty with epidural anesthesia      |
| Logical                | If-Then           | BMI greater than 38 for women                          |
| Qualifier              | Risk-For          | death of death                                        |
|                        | Severity          | symptoms                                              |
|                        | Stability         | hemodynamically unstable                               |
| Temporal and Comparative | After             | infected following admission                          |
|                        | Before            | diagnosis of acute decompensation prior to visit       |
|                        | Duration          | type 1 diabetes for at least 1 year                    |
|                        | During            | mechanically ventilated while admitted                 |
|                        | Numeric-Filter    | body weight less than 110 pounds in past 6 months      |
|                        | Minimum-Count     | admitted at least twice                                |
|                        | Temporality       | seen within past 6 months                             |
| Other                  | Location          | admitted to the ICU                                    |

Table 3. Examples of representative relations. Direction of arrows indicates role, i.e., subject \(\rightarrow\) target entity.
In such cases, most query generation algorithms introduce additional rule-based or syntactic parsing modules, such as SuTime17 to further normalize the phrase to a value4,11. This ambiguity in label semantics creates unnecessary complexity in downstream systems, requiring that the same text be processed a second time.

In contrast, we designed the LCT annotation schema to favor discrete, explicit entities and relations where possible, with an aim toward reducing the need for additional normalization steps needed for query generation. In our annotation schema, this example would be annotated with the following fine-grained entities:

```
| Entity Type | Value |
|-------------|-------|
| Eq-Temporal-Period | [past] |
| Eq-Value | six |
| Eq-Temporal-Unit | [month] |
```

As shown in the example, each token is uniquely annotated, with the values [past] and [month] serving to more clearly disambiguate semantics of the temporal phrase to a normalized temporal value. Moreover, as fine-grained entities are connected by relations to general entities, which can in turn have relations to other general entities, the LCT annotation schema is able to capture eligibility criteria semantics at a deeper level than other corpora.

**Rich semantic representation.** Certain eligibility criteria cannot be directly translated into queries, but instead must first be reasoned upon. For example, a query to find patients meeting the criterion of “conditions contraindicating pioglitazone” requires reasoning to first answer the question, What conditions contraindicate use of pioglitazone? Such reasoning may be performed by a knowledge base or other methods, but cannot be done unless the contraindicative relation is detected:

```
| Entity Type | Value |
|-------------|-------|
| Condition | Conditions |
| Contraindication | contraindicating |
| Drug-Name | pioglitazone |
```

As the span “Conditions” is labeled Condition but does not have an overlapping Condition-Name entity, it is considered unnamed and thus would need to be reasoned upon to determine. “[P]ioglitazone”, on the other hand, includes a Drug-Name entity and is thus considered named. The absence of overlapping Name entities serves as an indicator to downstream applications that reasoning may be needed to determine relevant conditions or drugs. We examine additional cases of the LCT’s semantic representation and benefits in the next section, where we compare the LCT annotation schema to Chia.

**Comparison to Chia.** We designed the LCT annotation schema by building upon the important previous work of EliIE and Chia. As Chia itself builds upon EliIE and is more recent, in the next section we compare the LCT corpus and Chia by examining cases of ambiguity handling and annotation difference in entities, relations, and coupling of entity types to the OMOP18 (Observational Medical Outcomes Partnership) data model. For each case we examine the LCT annotation schema’s novel solutions and contributions. Figure 2 shows comparison examples of annotations of the same eligibility criteria using the two corpora in the Brat annotation tool19.

**Capturing entity semantics.** Example 2 of Fig. 2 demonstrates the need to closely capture semantics in clinical trials eligibility criteria for unnamed entities. The span “Medication[s] which affect glycemic control” refers to any drug which potentially affects glycemic control. As discussed, the LCT annotation schema uses Name entities to handle such cases, where the absence in this example of a Drug-Name entity indicates that “Medication” refers to any drug, and thus may need to be determined by downstream use of a knowledge base or other methods.

As can be seen, Chia does not differentiate between named and unspecified drugs, conditions, procedures and so on. While it is true that for query generation one may need to normalize these spans to coded representations (e.g., ICD-10, RxNorm or LOINC codes) and may in the process find that the span “Medication” is not a particular medication (and thus can be assumed to be any medication), such a workaround nonetheless complicates usage of the corpus in finding and handling such cases in a more direct, less error-prone way.

Consider also the phrase, “after age 30” in example 4 in Fig. 2. In Chia, disambiguation of time units, values, and chronological tense must be performed by additional processing, as the Chia Value entity provides no information as to the semantics of the component sub-spans. In contrast, in the LCT corpus the tokens “after”, “age”, “30” are annotated with the explicit entities and relations to enable more straightforward query generation.

**Capturing relation semantics.** Eligibility criteria frequently contain entities which relate to other entities in the form of examples, abbreviations, equivalencies, or explicit lists. Chia uses Subsumes relations to denote that one or more entities are a subset, depend on, or are affected by another entity in some way. In many cases however these entities leave significant semantic ambiguity which may complicate query generation. Consider the phrase:

“conditions predisposing to dysphagia (eg, gastroesophageal reflux disease [GERD] ...)”
In this case, "gastroesophageal reflux disease" is an example of a condition, while "GERD" is an abbreviation. However, both are Subsumes relations in the Chia annotation:

```
> conditions predisposing to dysphagia   (e.g., gastroesophageal reflux disease)   [GERD]
```

In contrast, in the LCT annotation schema this example would be annotated as:

```
> conditions  predisposing  to  dysphagia  (e.g., gastroesophageal reflux disease)  [GERD]
```

The LCT annotation uses Abbrev-Of and Example-Of relations to clearly differentiate relations between "gastroesophageal reflux disease", "GERD", and "conditions predisposing to dysphagia". Additionally, rather than grouping the latter into a single Condition entity, the LCT is also much more granular, with the annotation reflecting that dysphagia is a condition patients are hypothetically predisposed to (due to other conditions such as GERD), but not necessarily actively afflicted by.

Another example illustrating the importance of capturing relation semantics can be seen in the following Chia annotations:

```
> type 1 diabetes   for at least 1 year
```

Fig. 2 Examples of clinical trials eligibility criteria annotated with Chia and LCT annotation schemas. Each example shows a criterion from a Chia annotation (above) and an LCT annotation of the same text for purposes of comparison (below).
While syntactically similar, the semantics in chronology expressed in the two criteria are different. In (1), “type 1 diabetes for at least 1 year” suggests that the diagnosis of type 1 diabetes mellitus should have occurred at least 1 year prior to the present. In other words, a unit of temporal measurement (1 year), should have passed since initial diagnosis. In contrast, “Acute coronary syndrome in the past 6 months” (2) suggests that a range of dates between the present and a past event (past 6 months), should have passed since the diagnosis. In Chia, however, the same Has-Temporal relation is used for both, blurring distinctions between durations of time versus ranges of dates, potentially leading to errors during query generation.

In the LCT annotation schema, these would be annotated as (omitting fine-grained entities for brevity):

(3)

The LCT annotations distinguish these types of temporal semantics by using distinct Duration and Temporality relations, allowing downstream queries to more accurately reflect researcher intent. The LCT corpus also does not include “for” or “in the” as part of the entities.

Data model mapping The Chia annotation schema is mapped to the OMOP Common Data Model and is designed to ease integration with other OMOP-related tools and generation of SQL queries on OMOP databases. Chia OMOP-derived entities generally follow the naming convention of OMOP domains and SQL database tables, such as Person, Condition, Device, and so on.

The LCT annotation schema takes a different approach by intentionally avoiding direct mappings to data models. This approach was chosen to (1) allow the annotation entities and relations flexibility to be transformed to any data model (including but not limited to OMOP) and (2) provide flexibility in capturing criteria important to the task of query generation, even when such criteria are not represented in OMOP.

A disadvantage of directly coupling an annotation schema to a data model is evidenced by criteria such as:

“Males aged 18 years and above”

In Chia, spans related to gender and age share the same Person entity:

(4)

The use of the Chia Person entity across gender and age results in loss of information and complications for query generation. As with quantitative and temporal annotations, the generic Person entity again forces the burden of normalization and additional parsing to downstream applications. In contrast, this example would be annotated first using general entities and relations in the LCT annotation schema:

Followed by fine-grained entities and values:

Eq-Comparison

Eq-Value “18”

Eq-Temporal-Unit [year] “years”

Eq-Operator [GTEQ] “or above”
The LCT annotation captures the male and age spans as distinguishable entities, closely preserving the semantics of the original text.

**Annotation process.** We used eligibility criteria from https://clinicaltrials.gov as the basis for our corpus. We extracted 1,020 randomly selected clinical trials eligibility descriptions, 20 for training and inter-annotator comparison and 1,000 for post-training annotation. Documents were included only if they met the following criteria:

1. The combined inclusion and exclusion criteria text was at least 50 characters long.
2. The clinical trial was uploaded on or after January 1st, 2018. This date was chosen because we found that clinical trials performed further in the past appeared to exhibit less structural consistency in language, punctuation and indentation compared to more recent text.

During annotation, 14 documents were found to be information poor (often with no spans to annotate) and discarded, resulting in 1,006 total annotated eligibility descriptions. Annotation was performed by two annotators, the first a biomedical informatician and the second a computer scientist. For initial annotation training, 20 documents were distributed to both annotators. Annotation was done in the following steps:

1. Annotation meetings were held bi-weekly for 3 months following initial annotation training in which the annotation guidelines were introduced. Initial meetings focused on discussion of annotation guideline implementation and revision. After each meeting, the annotation guidelines were revised to include new named entities and relationships and inter-annotator agreement was recalculated using F1-scores. Each annotator used the UMLS Terminology Services (UTS) Metathesaurus Browser (https://uts.nlm.nih.gov/uts/umls/home) to search for biomedical concepts whose meaning was unclear.

2. Inter-annotator agreement was calculated using F1 scoring for entities and relations with 20 double-annotated documents. The clinical trial was uploaded on or after January 1st, 2018. This date was chosen because we found that clinical trials performed further in the past appeared to exhibit less structural consistency in language, punctuation and indentation compared to more recent text.

2. After annotation guideline revisions and annotation training were completed, eligibility criteria were assigned to each annotator, with each clinical trial eligibility criteria annotated by a single annotator using the BRAT annotation tool19. Due to differences in time availability for annotation, roughly 90% (887 documents) of the annotation task was performed by the first annotator, and 99 documents by the second annotator.

3. At the point in which 50% of the corpus was annotated, we trained two neural networks (one for general entities and another for fine-grained entities) using the NeuroNER tool20 on our manually annotated eligibility criteria to predict annotations for the remaining 50%. NeuroNER is a state-of-the-art entity extraction system which has been successfully adapted to tasks such as de-identification and concept extraction. NeuroNER utilizes bidirectional Long Short-Term Memory and Conditional Random Fields (biLSTM + CRF) for token-level multiple-label prediction. We used the NeuroNER-predicted entities to auto-populate our remaining eligibility descriptions.

4. Manual annotation was completed on the remaining 50% of eligibility descriptions by editing and correcting the predicted entities from NeuroNER in (3).

The resulting corpus included 887 single-annotated and 119 double-annotated total notes.

**Data Records**

The LCT corpus annotated eligibility criteria and text documents can be found at https://doi.org/10.6084/m9.figshare.1720961021. Code for pre-annotation and analysis are available at https://github.com/uw-bionlp/clinical-trials-gov-data. The LCT corpus is annotated using the Brat "standoff" format. The Brat format includes two file types, " .txt" files and " .ann" files.

**Text (.txt) files.** The free-text eligibility criteria information in the 1,006 documents of the LCT corpus. Each file is named using the "NCT" identifier used by https://clinicaltrials.gov.

**Annotation (.ann) files.** The annotation files used by Brat for tracking annotated spans of text and relations. Each .ann file corresponds to a.txt file of the same name. Each row of a .ann file may begin with a "T" (for an entity) or "R" (for a relation), followed by an incremental number for uniquely identifying the entity or relation (e.g., "T15"). "T" rows are of the form "T<number>: <entity type>: <start character index>: <stop character index>", where start and stop indices correspond to text in the associated.txt file. "R" rows are of the form "R<number>: <relation type>: Arg1: <ID>:Arg2: <ID>", where ID values correspond to identifiers of entities. Additionally, for ease of annotation certain LCT relations are defined as arguments of Brat "events"; identified by "E". "E" rows are of the form "E<number>: <entity type>: <ID>: <relation type>: <ID>". More information on the Brat format can be found at https://brat.nlplab.org/standoff.html.

**Technical Validation**

**Inter-annotator agreement.** Inter-annotator agreement was calculated using F1 scoring for entities and relations with 20 double-annotated documents. Entity annotations were considered matching only if entity types and token start and end indices matched exactly. Relations annotations were similarly considered matching only if relation type and token start and end indices of both the subject and target matched exactly.

Initial inter-annotator agreement using the 20 training documents was 76.1% for entities and 60.3% for relations. Inter-annotator agreement improved slightly to 78.1% (+2%) for entities and 60.9% (+0.6%) for relations in the 99 additional double-annotated documents, indicating reasonably high annotator agreement considering the complexity of the annotation task.
We found two categories of annotation differences in double-annotated eligibility criteria. First, in spans such as:

"Reported being a daily smoker"

While both annotators tended to annotate "smoker" as both a Condition and Condition-Name, adjectives such as "daily" were often annotated as Eq-Comparison and Eq-Temporal-Unit[day] by one annotator and Stability[stable] by the other. After review, these were generally reconciled to the first pattern.

The second category of differences can be seen in spans such as "CDI Diarrhea", where "CDI" refers to Clostridium Difficile Infection. In these cases, the annotators may annotate this span as (omitting Condition-Name entities for brevity):

↓ ← ↓

−                              −

CDI Condition

Diarrhea Condition

(5)

Caused-By or

↓ Condition

"CDI Diarrhea"

(6)

The first annotation separates "CDI" and "Diarrhea" into two entities, with "Diarrhea" Caused-By "CDI", while the second annotation treats them as a single entity. Reconciliation in these cases was done by referring to the UMLS Metathesaurus to determine whether the combined span existed as a single concept within the UMLS. As "Clostridium Difficile Diarrhea" exists as a UMLS concept (C0235952), the annotations in this example were reconciled to use the second, multi-span entity.

Baseline prediction. To evaluate baseline predictive performance on the LCT corpus, we first created a randomly assigned 80/20 split of the corpus, with 804 documents used for the training set and 202 for the test set. For entity prediction, we trained NER models using biLSTM + CRF and BERT neural architectures. For BERT-based prediction, we used two pretrained models trained on published medical texts, SciBERT and PubMedBERT. For both biLSTM + CRF and BERT predictions, we trained one model to predict general entities and another for fine-grained entities.

For relation extraction, we evaluated SciBERT for sequence classification as well as a modified BERT architecture, R-BERT, following methods developed by Wu & He, also using the pretrained SciBERT model. Table 4 shows hyperparameters used for each task.

We achieved the highest micro-averaged F1 score of 81.3% on entities using SciBERT and 85.2% on relations using the R-BERT architecture with SciBERT. Results of representative entities and relations are shown in Tables 5 and 6. Among entities, we found two particular categories performed relatively well with F1 scores of 70% or often greater: (1) Entities which are syntactically varied but occurred relatively frequently in eligibility descriptions, such as Condition, Procedure, and Eq-Comparison, (2) Entities which sometimes occurred less frequently but with greater relative syntactic consistency and structure, such as Age, Contraindication, and Birth. Entities which occurred very infrequently, such as Death tended to have both low precision and recall.

For relations, we found the most frequently occurring relations, such as Eq-Comparison, Temporality, Modifies, and Example-Of to perform well, with F1 scores greater than 85%. Among less frequently occurring relations, we found a number of cases where relations which tend to occur in similar positions within sentences and grammatical structures were frequently mistaken during prediction. For example, Eq-Comparison (e.g., “greater than 40”) and Minimum-Count (e.g., “at least twice”) were sometimes incorrectly predicted. We found similar incorrect predictions for relations such as Treatment-For (e.g., “Surgery for malignant pathology”) and Using (e.g., “knee joint replacement with general anaesthesia”).

In future work we intend to examine approaches improving prediction of less frequently occurring entities and relations. A full listing of baseline prediction results can be found with the annotation guidelines at https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki/Named-Entity-Recognition-and-Relation-Extraction-performance.

| Task                      | Architecture                  | Hyperparameter/Embeddings          | Training Value |
|---------------------------|-------------------------------|-----------------------------------|----------------|
| Named Entity Recognition  | biLSTM + CRF                  | Character Dimensions              | 25             |
|                           |                               | Token Embedding Dimensions        | 100            |
|                           |                               | Learning Rate                     | 0.005          |
|                           |                               | Dropout                           | 0.5            |
|                           |                               | Pretrained Embeddings              | GloVe28        |
| Relation Extraction       | BERT & R-BERT                 | Pretrained Model                   | SciBert        |
|                           |                               | Learning Rate                     | 0.00003        |

Table 4. Hyperparameters and pre-trained embeddings used for named entity recognition and relation extraction baseline results. For the NER task, the same architecture and hyperparameters were used for both general and fine-grained entity models. For the relation extraction task, the same hyperparameters were used with both the BERT and R-BERT architectures.
To determine the quality of single-annotated documents compared to those which were double-annotated, we trained NER models (one for general and another for fine-grained entities, as in earlier experiments) using SciBERT with the 887 single-annotated documents and evaluated on the 119 double-annotated documents. The results were a precision of 79.7%, recall of 82.5%, and an F1 score of 81.4%, which are very close to the highest performance of our randomly split train/test set results shown in Table 5. These results indicate relative uniformity and consistency in the corpus across both single- and double-annotated documents.

As the latter near-half (493 documents) of the LCT corpus was automatically annotated, then manually corrected, we also evaluated the quality of the manually annotated portion versus the semi-automatically annotated portion to ensure consistency. We first trained NER models with SciBERT using the manually annotated portion and tested on the semi-automated portion, then reversed the experiment and trained on the semi-automated portion and tested on the manually annotated portion. Results are shown in Table 7.

Results of the experiments when training on both the manually and semi-automatically annotated halves of the corpus show comparable results, with the greatest difference being in precision, with the manual annotation-trained model performing slightly worse (−4.7%) in prediction versus the semi-automated annotation-trained model. Overall F1 scores were similar at 78.6% and 80.0%, suggesting reasonable consistency across the corpus.

### Annotation quality evaluation

To determine the quality of single-annotated documents compared to those which were double-annotated, we trained NER models (one for general and another for fine-grained entities, as in earlier experiments) using SciBERT with the 887 single-annotated documents and evaluated on the 119 double-annotated documents. The results were a precision of 79.7%, recall of 82.5%, and an F1 score of 81.4%, which are very close to the highest performance of our randomly split train/test set results shown in Table 5. These results indicate relative uniformity and consistency in the corpus across both single- and double-annotated documents.

As the latter near-half (493 documents) of the LCT corpus was automatically annotated, then manually corrected, we also evaluated the quality of the manually annotated portion versus the semi-automatically annotated portion to ensure consistency. We first trained NER models with SciBERT using the manually annotated portion and tested on the semi-automated portion, then reversed the experiment and trained on the semi-automated portion and tested on the manually annotated portion. Results are shown in Table 7.

Results of the experiments when training on both the manually and semi-automatically annotated halves of the corpus show comparable results, with the greatest difference being in precision, with the manual annotation-trained model performing slightly worse (−4.7%) in prediction versus the semi-automated annotation-trained model. Overall F1 scores were similar at 78.6% and 80.0%, suggesting reasonable consistency across the corpus.

### Usage Notes

The LCT corpus is designed to facilitate query generation and question answering for real-world clinical trials and clinical research, specifically for a future version of the Leaf cohort discovery tool. Figure 4 visualizes an example of a transformation of LCT annotated data into a Directed Acyclical Graph (DAG) structure, which can then be potentially compiled into SQL, FHIR, SPARQL, or other query methods.

To demonstrate the value and utility of the corpus, using the trained baseline Named Entity Recognition and Relation Extraction models, we developed a simple prototype web application to test named entity and relation prediction on unseen text. Figure 3 shows a screenshot of the models correctly predicting entities and relations on an input sentence not present in the LCT corpus. As can be seen, the models are able to predict entities and relations with very high accuracy on new text, demonstrating the power of the corpus.

| Category | Entity      | Count | biLSTM + CRF | PubMedBERT | SciBERT |
|----------|-------------|-------|-------------|------------|---------|
| Clinical | Condition   | 7,087 | 76.6/78.1/78.3 | 76.1/79.4/77.7 | 78.4/83.3/80.8 |
|          | Contraindication | 142 | 93.7/78.9/85.7 | 77.4/80.0/78.6 | 100.0/96.5/98.3 |
|          | Drug        | 1,404 | 76.8/81.3/79.0 | 74.1/80.9/77.4 | 73.4/80.9/77.0 |
|          | Encounter   | 302  | 64.1/58.1/60.9 | 51.7/61.7/56.3 | 58.3/74.5/65.4 |
|          | Observation | 2,558 | 74.3/66.1/69.9 | 67.9/73.5/70.6 | 72.1/77.6/74.7 |
|          | Procedure   | 3,016 | 68.3/75.5/71.9 | 67.0/75.9/71.2 | 71.3/79.4/75.1 |
| Demographic | Age       | 708   | 91.3/95.4/93.3 | 82.4/88.5/85.3 | 99.1/98.3/98.7 |
|          | Birth       | 27    | 100.0/80.0/88.8 | 100.0/62.5/76.9 | 100.0/62.5/76.9 |
|          | Death       | 35    | 33.3/33.3/33.3 | 0.0/0.0/0.0 | 100.0/20.0/33.3 |
|          | Family-Member | 147 | 40.0/19.0/25.8 | 33.3/55.5/41.6 | 44.9/61.1/51.7 |
|          | Language    | 194   | 92.5/96.1/94.3 | 73.8/100.0/84.9 | 96.6/93.5/95.0 |
| Logical  | Negation    | 952   | 74.3/82.7/87.2 | 60.9/73.1/66.4 | 73.5/82.9/77.9 |
| Qualifier| Assertion   | 1,157 | 66.6/62.8/64.7 | 56.1/58.9/57.5 | 62.1/65.8/63.9 |
|          | Modifier    | 3,464 | 65.0/58.3/61.5 | 59.2/64.0/61.5 | 58.5/65.4/61.8 |
|          | Polarity    | 360   | 82.5/88.0/85.1 | 74.6/67.4/70.8 | 81.4/79.5/80.4 |
|          | Risk        | 117   | 93.1/96.4/94.7 | 91.3/91.3/91.3 | 95.4/91.3/93.3 |
|          | Severity    | 569   | 86.8/90.8/88.7 | 76.7/79.5/75.8 | 86.5/94.1/90.2 |
|          | Stability   | 397   | 84.2/67.6/75.0 | 79.4/75.0/77.1 | 75.3/84.7/79.7 |
| Temporal and Comparative | Criteria-Count | 33  | 50.0/66.6/57.1 | 28.5/40.0/33.3 | 12.5/20.0/15.5 |
|          | Eq-Comparison | 5,298 | 83.1/83.8/83.4 | 81.4/85.0/85.2 | 85.3/89.3/87.3 |
|          | Eq-Temporal-Period | 2,057 | 88.7/89.2/88.9 | 70.0/73.9/71.9 | 82.6/86.3/84.4 |
|          | Eq-Temporal-Recency | 131  | 68.7/84.6/75.8 | 43.4/55.5/48.7 | 50.0/66.6/57.1 |
|          | Eq-Temporal-Unit | 1,808 | 95.1/97.6/96.4 | 97.4/98.1/97.8 | 98.2/99.4/98.8 |
|          | Eq-Value    | 3,835 | 91.8/95.3/93.5 | 95.5/96.2/95.9 | 96.4/97.1/96.7 |
| Other    | Location    | 371   | 68.5/58.7/63.2 | 65.4/71.6/68.3 | 73.4/78.3/75.8 |
|          | Total       | 56,146 | 80.2/79.6/79.9 | 75.3/78.7/77.0 | 79.0/83.7/81.3 |

Table 5. Baseline entity prediction scores (% Precision/Recall/F1). Corpus-level micro-averaged scores are shown in the bottom row. For brevity a representative sample of entities is shown. Count refers to the total count of unique spans annotated in the entire corpus. Entities included in the total count and scores but omitted for brevity are Acuteness, Allergy, Condition-Type, Code, Coreference, Ethnicity, Eq-Operator, Eq-Unit, Indication, Immunization, Insurance, Life-Stage-And-Gender, Organism, Other, Specimen, Study and Provider.
Limitations. The LCT corpus is designed as a granular and robust resource of annotated eligibility criteria to enable models for entity and relation prediction as means of query generation. The corpus does have a number of limitations however which should be recognized. First, the corpus is largely singly annotated, with 119 of 1,006 documents (11%) double annotated and reconciled, while double annotation is generally considered to be the gold standard in the NLP research community. As discussed in the Technical Validation section, the reasonably high F1 score from experiments to evaluate NER when training on the singly annotated portion of the corpus suggests relative consistency of annotation across both single and double annotated documents. Additionally, entities in roughly half of the LCT corpus (493 documents) were automatically predicted, then manually corrected. This can potentially lead to data bias if predicted entities are not thoroughly reviewed and corrected by human annotators. Similar results from our experiments to detect differences in performance by training on the manually annotated portion versus the semi-automatically annotated portion (F1 scores of 78.6% and 80.0%) suggest this may not be not a significant issue. Last, though the ultimate goal of the LCT corpus is to facilitate more accurate query generation, the corpus itself is not composed of queries by which it can be compared to similar corpora and thus cannot necessarily be proven to be more effective. Similarly, as we do not formally define a quantifiable means for measuring semantic representation within annotations, it is difficult to demonstrate that the LCT corpus enables more accurate query generation.

Table 6. Baseline relation prediction scores (%). Corpus-level micro-averaged scores are shown in the bottom row. For brevity a representative sample of relations is shown. Count refers to the total count annotated in the entire corpus, including relations not shown. The count total excludes general to fine-grained entity relations, which as overlapping spans are not used for relation prediction. Relations included in the total count and scores but omitted for brevity are Acuteness, Code, Criteria, Exception, From, Indication-For, Is-Other, Max-Value, Min-Value, Polarity, Provider, Refers-To, Specimen, Stage, Study-Of and Type.

| Category               | Relation                  | Count | SciBERT          | R-BERT + SciBERT |
|------------------------|---------------------------|-------|------------------|------------------|
| Alternatives and Examples | Abbrev-Of               | 462   | 95.2/90.9/93.0   | 92.3/93.1/94.2   |
|                        | Equivalent-To            | 516   | 61.5/69.5/65.3   | 59.6/67.3/63.2   |
|                        | Example-Of              | 1,497 | 94.8/92.9/93.8   | 90.5/91.7/91.1   |
| Clinical               | Contraindicates         | 153   | 90.9/90.9/90.9   | 90.9/90.9/90.9   |
|                        | Caused-By               | 726   | 63.0/86.4/72.9   | 78.6/86.4/82.3   |
|                        | Found-By                | 293   | 90.4/59.3/71.7   | 9.3/71.8/75.4    |
|                        | Treatment-For           | 457   | 69.2/69.2/69.2   | 61.7/74.3/67.4   |
|                        | Using                   | 405   | 73.8/83.7/78.4   | 66.6/64.8/65.7   |
| Logical                | And                     | 821   | 54.1/60.0/56.9   | 53.8/53.8/53.8   |
|                        | If/Then                 | 261   | 57.6/65.2/61.4   | 55.5/65.2/60.0   |
|                        | Negates                 | 984   | 74.3/91.0/81.8   | 74.5/88.7/81.0   |
|                        | Or                      | 4,156 | 85.1/93.2/89.0   | 88.4/92.2/90.2   |
| Qualifier              | Asserted                | 1,184 | 83.7/89.0/86.3   | 85.9/89.0/87.5   |
|                        | Modifies                | 3,400 | 90.9/94.2/92.5   | 92.2/95.4/93.8   |
|                        | Risk-For                | 90    | 92.3/85.7/88.8   | 92.8/92.8/92.8   |
|                        | Severity                | 529   | 80.2/96.6/87.6   | 86.3/96.6/91.2   |
|                        | Stability               | 395   | 76.0/92.6/83.5   | 76.4/95.1/84.7   |
| Temporal and Comparative | After                   | 166   | 75.0/70.5/72.7   | 72.2/76.4/74.2   |
|                        | Before                  | 320   | 70.2/86.6/77.6   | 78.1/83.3/80.6   |
|                        | Duration                | 243   | 59.3/79.1/67.8   | 64.5/83.7/72.7   |
|                        | During                  | 350   | 66.6/68.7/67.6   | 63.6/65.6/64.6   |
|                        | Numeric-Filter          | 1,957 | 84.6/93.3/88.7   | 85.7/92.3/88.8   |
|                        | Minimum-Count           | 173   | 64.2/69.2/66.7   | 71.4/76.9/74.0   |
|                        | Temporality             | 2,645 | 80.7/90.7/85.4   | 81.8/92.2/86.7   |
| Other                  | Location                | 207   | 64.2/94.7/76.6   | 69.2/94.7/80.0   |
|                        | Total                   | 24,379| 80.2/88.2/84.0   | 82.5/88.0/85.2   |

Table 7. Results of NER experiments using the manually annotated and semi-automated portions of the corpus. The manually annotated portion includes 513 documents while the semi-automatically annotated portion is 493 documents.
Future work. As discussed, evaluation of generated query accuracy and semantic representation in annotations is difficult and can potentially be done by different methods, such as ROUGE scoring to compare generated query syntax to expected syntax, or by including UMLS Concept identifiers within the LCT annotation schema and comparing the number of UMLS concepts to those found in other corpora.

Fig. 3 Screenshot of a prototype web application for real-time entity and relation prediction on custom user input text.

Fig. 4 Example of an LCT annotated document (top) transformed into a Directed Acyclical Graph (bottom). LCT entities and relations are readily transformable into tree, graph, or object-oriented representations used for query generation.
Taking a different approach, in future work, we intend to evaluate the LCT corpus and query generation methods by evaluating generated queries in the context of real clinical trials which have taken place at the University of Washington (UW). As the UW EHR system maintains clinical trial enrollments and patient identifiers alongside clinical data, it is possible to query our EHR databases to compare patients who actually enrolled in clinical trials versus those found by our queries had they been run at the time of a given trial. We believe this means of evaluation is uniquely valuable as it uses real world clinical trials and EHR data while scoring queries by the accuracy of their ultimate results rather than less consequential factors such as syntax.

## Code availability

All code used to generate, pre-annotate, and analyze the LCT corpus is freely available at https://github.com/uw-bionlp/clinical-trials-gov-data. The LCT annotation guidelines can be found at https://github.com/uw-bionlp/clinical-trials-gov-annotation/wiki.

Received: 1 February 2022; Accepted: 28 June 2022;
Published online: 11 August 2022

## References

1. Richesson, R. L. et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association* 20, e226–e231 (2013).
2. Dobbins, N. J. et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. *Journal of the American Medical Informatics Association* 27, 109–118 (2019).
3. Murphy, S. N. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association* 17, 124–130 (2010).
4. Yuan, C. et al. Criteria2Query: A natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association* 26, 294–305, https://doi.org/10.1093/jamia/ocy178 (2019).
5. Wang, F., Shi, T. & Reddy, C. K. A translate-edit model for natural language question to sql query generation on multi-relational healthcare data. arXiv preprint arXiv:1908.01839 (2019).
6. Yu, X. et al. Dataset and Enhanced Model for Eligibility Criteria-to-SQL Semantic Parsing. 5828–5837 (2020).
7. Koopman, B. & Zuccon, G. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 669–672 (2016).
8. Liu, S. et al. Implementation of a cohort system for clinical data repositories using the observational medical outcomes partnership common data model: Proof-of-concept system validation. *JMR medical informatics* 8, e17576 (2020).
9. Park, J. et al. A framework (socratex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database development and usability study, *JMR medical informatics* 9, e23983 (2021).
10. Truong, T. H. et al. ITTC@ TREC 2021 Clinical Trials Track, arXiv preprint arXiv:2202.07858 (2022).
11. Weng, C. et al. EliXR: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association* 18, i116–i124, https://doi.org/10.1136/amiajnl-2011-000321 (2011).
12. Kang, Y. et al. ElIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association* 24, 1062–1071, https://doi.org/10.1093/jamia/ocx019 (2017).
13. Tu, S. W. et al. A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics* 44, 239–250, https://doi.org/10.1016/j.jbi.2010.09.007 (2011).
14. Milian, K. et al. Enhancing reuse of structured eligibility criteria and supporting their relaxation. *Journal of biomedical informatics* 56, 205–219 (2015).
15. Kury, F. et al. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific data* 7, 1–11 (2020).
16. Boland, M. R., Tu, S. W., Carini, S., Sim, I. & Weng, C. EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2012, 71–80 (2012).
17. Chang, A. X. & Manning, C. D. Sutime: A library for recognizing and normalizing time expressions. *Lrec* 3735, 3740 (2012).
18. Hripcsak, G. et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and Informatics* 216, 574 (2015).
19. Senetor, P. et al. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapters of the Association for Computational Linguistics*, 102–107 (2012).
20. Dernoncourt, F., Lee, J. Y. & Szolovits, P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks, arXiv preprint arXiv:1705.05487 (2017).
21. Dobbins, N., Mullen, T., Uzuner, O. & Yetisen, M. The Leaf Clinical Trials (LCT) corpus: an annotated corpus of clinical trials eligibility criteria. figshare https://doi.org/10.6084/m9.figshare.17209610 (2022).
22. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
23. Bellamy, L. I. O. K. & Cohan, A. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019).
24. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* (HEALTH) 3, 1–23 (2021).
25. Wu, S. & He, Y. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2361–2364 (2019).
26. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81 (2004).
27. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, D267–D270 (2004).
28. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543 (2014).

## Acknowledgements

This study was supported in part by the National Library of Medicine under Award Number R15LM013209 and by the National Center for Advancing Translational Sciences of National Institutes of Health under Award Number UL1TR002319. Experiments were run on computational resources generously provided by the UW Department of Radiology.
Author contributions
N.D. created the annotation guidelines, was primary annotator, and drafted the original manuscript. T.M. served as secondary annotator and reviewed and revised the manuscript. O.U. reviewed the task and goals as well as reviewed and revised the manuscript. M.Y. conceived of the annotation task, supervised, and reviewed and revised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to N.J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022