
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 118 (2021) 103790

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

# Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials

Yingcheng Sun [a,1], Alex Butler [a,b,1], Latoya A. Stewart [c], Hao Liu [a], Chi Yuan [a], Christopher T. Southard [c], Jae Hyun Kim [a], Chunhua Weng [a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[b] Department of Medicine, Columbia University, New York, NY, USA
[c] Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA

ARTICLE INFO

*Keywords:*
Clinical trial
Eligibility criteria
COVID-19
Structured text corpus
Machine readable dataset

ABSTRACT

Clinical trials are essential for generating reliable medical evidence, but often suffer from expensive and delayed patient recruitment because the unstructured eligibility criteria description prevents automatic query generation for eligibility screening. In response to the COVID-19 pandemic, many trials have been created but their information is not computable. We included 700 COVID-19 trials available at the point of study and developed a semi-automatic approach to generate an annotated corpus for COVID-19 clinical trial eligibility criteria called COVIC. A hierarchical annotation schema based on the OMOP Common Data Model was developed to accommodate four levels of annotation granularity: i.e., study cohort, eligibility criteria, named entity and standard concept. In COVIC, 39 trials with more than one study cohorts were identified and labelled with an identifier for each cohort. 1,943 criteria for non-clinical characteristics such as "informed consent", "exclusivity of participation" were annotated. 9767 criteria were represented by 18,161 entities in 8 domains, 7,743 attributes of 7 attribute types and 16,443 relationships of 11 relationship types. 17,171 entities were mapped to standard medical concepts and 1,009 attributes were normalized into computable representations. COVIC can serve as a corpus indexed by semantic tags for COVID-19 trial search and analytics, and a benchmark for machine learning based criteria extraction.

## 1. Background

Since the first reported cases in December 2019, Coronavirus Disease 2019 (COVID-19) has spread rapidly from country to country and become a global pandemic [1]. As of February 2021, over 111 million confirmed positive cases have been reported worldwide, leading to 2.46 million deaths [2]. To deal with one of the worst pandemics in the world's history [3], numerous research studies assessing the efficacy and safety of COVID-19 treatments are being conducted at an unprecedented rate. As of September 31, 2020, over 3,500 clinical trials targeting COVID-19 were registered in ClinicalTrials.gov, the largest clinical trial registry in the world. However, the free-text eligibility criteria are not amenable for computational analyses [4] or direct cohort queries using from electronic health records (EHRs) [5].

Formal representations for eligibility criteria have been pursued by the biomedical informatics research community to facilitate cohort identification, trial search [6] and clinical analytics [7,8]. Several groups [4,9,10] have published datasets of annotated eligibility criteria, the largest of which containing criteria of 1,000 randomly selected clinical trials including 15 entity types and 12 relationships [11]. Manual annotation of clinical trials contributes to high-quality datasets but entails considerable costs of time and human labor. Automatic information extraction methods that leverage natural language processing (NLP) technologies may reduce the required curation time and effort but the performance is limited due the semantic complexity of eligibility criteria.

To address the challenge of lacking annotated COVID-19 trials, here we present a semi-automatic approach to annotating COVID-19 trial eligibility criteria by enhancing NLP-assisted manual concept recognition with machine-based concept normalization. The named entities in eligibility criteria are first annotated and classified by an automated Named Entity Recognition (NER) module, and then reviewed and

* Corresponding author at: Department of Biomedical Informatics, Columbia University, 622 W 168 ST, PH-20 room 407, New York, NY 10032, USA.
*E-mail address:* chunhua@columbia.edu (C. Weng).
[1] Equal-contribution first authors.

https://doi.org/10.1016/j.jbi.2021.103790
Received 23 October 2020; Received in revised form 21 February 2021; Accepted 10 April 2021
Available online 28 April 2021

updated by domain experts. The verified entities are converted to standard concepts by a concept mapping tool called Usagi [12]. Standard concepts ("concept" refers to "standard concept" in this paper) are medical terminologies that can be used as normative expressions of a clinical entity within standardized analytics. Translating source specific codes into standard concepts will promote the usage of eligibility criteria in a "common language" and leverage the power of rich clinical corpus and EHR database. Incorrectly mapped concepts are updated by experts in an iterative fashion. Associated relationships are also manually annotated. By integrating manual and NLP efforts, this approach could offer a cost-effective solution to generate trustworthy structured eligibility criteria. In this paper, we describe this semi-automatic eligibility criteria annotation framework and present our generated dataset COVIC with 700 clinical trials acquired from the ClinicalTrials.gov by querying all the trials indexed with "COVID-19" and including 11,710 annotated criteria formatted with the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [13]. Compared to other data models such as Integrating Biology and the Bedside (i2b2) data model [16], Sentinel Common Data Model (SCDM) [17] or Patient Centered Outcomes Research Network (PCORnet) [18], OMOP CDM has an international orientation and participation with terminology sets from multiple countries [19], which is widely used in the medical research community for health data standardization.

To the best of our knowledge, COVIC is the first structured eligibility criteria dataset for

COVID-19 clinical trials. COVIC provides standardized human- and computer-interpretable annotation and representation of COVID-19 trials, which can greatly facilitate patient recruitment and exhibits great analytical flexibility. The proposed scheme defines a hierarchical trial annotation model with four layers that accommodate different levels of annotation granularity, such as study cohort groups, named entities or standard concepts, providing a more comprehensive annotation specification. To address the imperative for sharing machine readable datasets to battle the current pandemic, we released version 1 of the dataset with 700 trials in September of 2020 at (https://github.com/WengLab-InformaticsResearch/COVID19-Structured_Trials). We also provide related source code and documentation along with the dataset, which allows for maximum understanding of the data and promotion of its use among non-experts. This dataset will be periodically updated to ensure accuracy of the data being shared.

## 2. Methods

Clinical trials are research studies performed on human subjects often to evaluate a new medical treatment, drug or device. The most complete repository of clinical trials is ClinicalTrials.gov, a web resource created and maintained by the U.S. National Library of Medicine (NLM) [14], which includes a collection of clinical trial registrations in key-value pairs. Most values within these records are very short and limited to a few discrete value types which can be considered structured data, such as target condition, start and end dates, phase, recruitment status, etc. In this paper, they are referred to as *metadata*. Eligibility criteria, although playing a central role in clinical research for specifying rules for screening clinical trial participants, exist as free text. We proposed a semi-automatic annotation method to generate computable representations for eligibility criteria and present them together with metadata as a machine-readable dataset. Fig. 1 is an overview of the methodology framework.

First, 700 trials were exported from the Aggregate Analysis of ClinicaTrials.gov (AACT) database by querying all the trials indexed with "COVID-19" as its target condition. AACT is a publicly available relational database which contains information about studies registered in ClinicalTrials.gov and is provided for research purposes by the Clinical Trials Transformation Initiative (CTTI) [15]. Next, relevant metadata and eligibility criteria were extracted separately from each trial. A hierarchical trial annotation schema following OMOP CDM was developed to guide the iterative annotation process, with the results reviewed and validated by domain experts. Next, six commonly used criteria across all COVID-19 trials about current age, high-risk status (e.g., hospital worker), COVID-19 status (e.g., yes, cleared, no), days since diagnosis (if relevant), current hospitalization or intensive care unit (ICU) admission, and pregnancy status are specified as "key criteria" and recorded separately from the annotation data table. Finally, all data are packaged as a COVID-19 structured trial dataset COVIC. We list a few application scenarios with COVIC, such as semantic retrieval of COVID-19 trial, eligibility query generation for patient cohort identification, trial similarity and collaborative analytics, and machine learning model training for information extraction from eligibility criteria.

### 2.1. The annotation schema

To accommodate annotation at different granularity levels, we built a hierarchical trial annotation model with four layers: study cohort layer, eligibility criteria layer, named entity layer and standard concept layer. Fig. 2 shows the hierarchy and consisted elements of the four layers.

1) Study cohort layer

To make sure the structured eligibility criteria can be used effectively for patient recruitment, the number of sub-study or cohort groups utilized in each clinical trial was identified first during the annotation process. If a clinical study includes multiple sub-studies or cohort groups, it will be divided into a group of trials and each of them will be

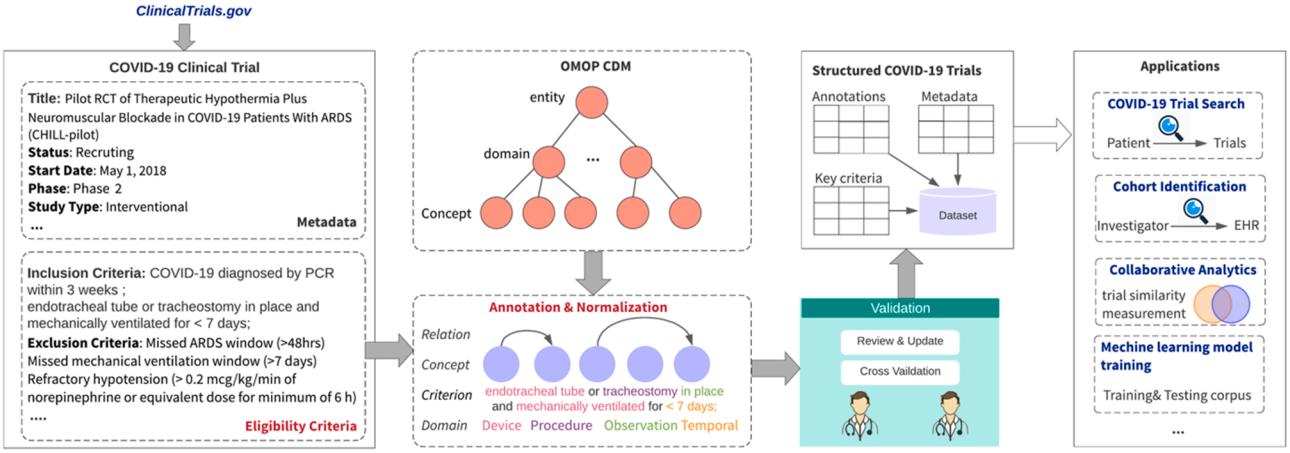


**Fig. 1.** Overview of the structured dataset generation framework.

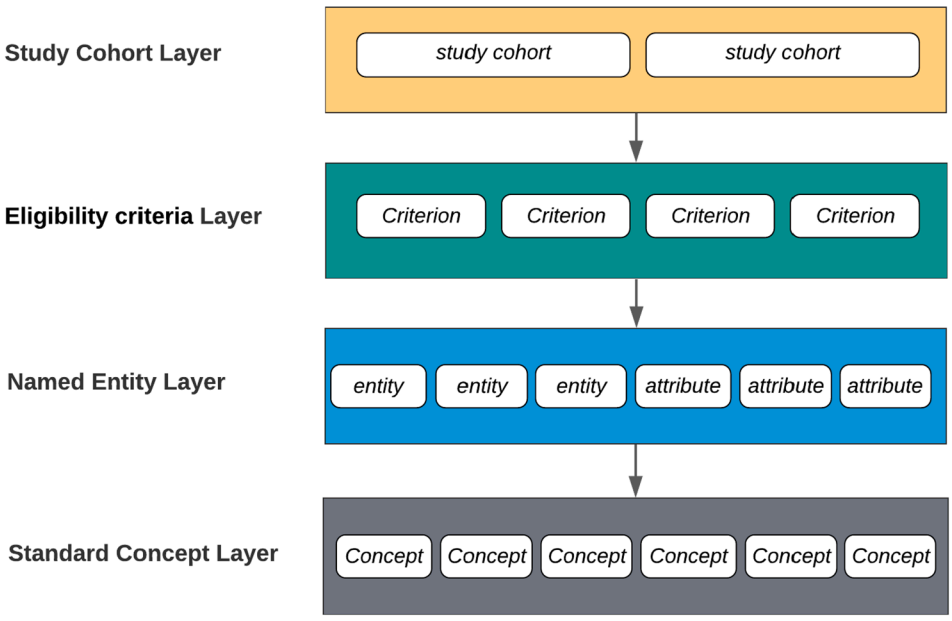


**Fig. 2.** Hierarchical trial annotation model with four layers: study cohort layer, eligibility criteria layer, named entity layer and standard concept layer.

treated as a new trial for annotation with a suffix added to their previous ID. Table 1 lists an example of a trial with three cohort groups and the names of three 'new' trials generated from the old one.

2) Eligibility criteria layer

In this layer, the "key criteria" and their types need annotation. First, criterion specifying any of the six "key criteria" discussed above will be identified and labelled. Next, "regular" criteria and "particular" ones are classified. Most of the criteria are "regular" ones annotated by terms, but there are four types of criteria labelled as a whole object because of their particular purposes, although they may contain entities or attributes. The four types are "non-query-able", "informed consent", "competing trial" and "post-eligibility". "non-query-able" means a given criterion cannot be used to query relevant patients from a relational database. Such criteria usually specify implicit or flexible eligibility instead of determinate requirements. "informed consent" criteria typically specify the signing of informed consents for participation in a clinical trial which is outside of the scope of database queries. "competing trial" typically excludes simultaneous participation in other clinical trials and reinforces exclusivity of participation. "post-eligibility" typically involves requirements that are met after the patient enrolls on the clinical trial. Table 2 shows examples of these four types of criteria.

3) Named entity layer

For "regular" criteria, entities and their associated attributes and relationships will be extracted and annotated. We adapted a simplified version of the Chia annotation model created by Kury et al[11] by not using its "Scope" mechanism considering its complexity. Table 3 lists the categories and examples of eight domains of entities and seven categories of attributes in this layer.

Most entity types and attributes are defined by the literal meaning of their names in Chia model. One attribute which requires specific explanation is "Reference_point". It comes downstream (usually directly) from a parent "Temporal", and specifies a concept whose timestamp is pivoting that "Temporal". For example, in "within two

**Table 1**
Example of a trial with three cohort groups.

| # | New Trial ID | Cohort Group Topic |
|---|---|---|
| 1 | NCT04494893_exposed | Cohort 1. Exposed to coronavirus disease |
| 2 | NCT04494893_active | Cohort 2. Infected with coronavirus disease |
| 3 | NCT04494893_recovered | Cohort 3. Recovered from coronavirus disease |

**Table 2**
Examples of four particular types of criteria.

| Type | Example |
|---|---|
| Non-query-able | *Any condition unsuitable for the study as determined by investigators; Any other reason that the Investigator considers makes the patient unsuitable to participate* |
| Informed consent | *Patient or responsible family member or surrogate signs informed consent* |
| Competing trial | *Not participate in any other clinical trial for an investigational therapy through day 30* |
| Post-eligibility | *Able to attend all scheduled visits and to comply with all study procedures; Agrees to required laboratory data collected which will include the baseline organ function and regular ongoing assessments done as part of routine care.* |

**Table 3**
Entity and attribute with different categories and examples.

| | Category | Example |
|---|---|---|
| **Entity** | Condition | Patients with ***bleeding disorders*** |
| | Observation | ***History of*** stem cell transplant |
| | Drug | Has received treatment with ***systemic anticancer treatments*** |
| | Measurement | ***SpO2*** < 92% |
| | Procedure | requires ***supplemental oxygen*** 2 LPM |
| | Person | If ***female***, subject must meet one of the following conditions |
| | Visit | Patients ***admitted to hospital*** |
| | Device | Has a ***Ventricular Assist Device*** |
| **Attribute** | Value | SpO2 ***< 92%*** |
| | Temporal | Has had plasmapheresis ***within the previous 24 h*** |
| | Qualifier | ***Significant*** cardiovascular disease |
| | Reference_point | within two weeks of a ***blood transfusion*** |
| | Mood | Patients who ***require*** renal replacement therapy |
| | Negation | ***no*** alternative explanation for current clinical condition |
| | Multiplier | ***at least two of the following*** high-risk criteria |

weeks of a blood transfusion" this entire text string is one "Temporal", and it contains (overlaps) the "Reference_point" "blood transfusion." Eleven types of relationships were used to specify the association between pairs of entities (AND, OR) or entity and its corresponded attribute: SUBSUMES, HAS_NEGATION, HAS_MULTIPLIER, HAS_QUALIFIER, HAS_VALUE, HAS_TEMPORAL, HAS_INDEX (target argument is reference_point), HAS_MOOD and HAS_CONTEXT (target argument is observation and not included in above relationships). More details can be found in the Supplementary File of the Chia Annotation Model [11].

4) Standard concept layer

In this layer, extracted named entities are converted to the standard concepts in the OMOP CDM. Each standard concept has a unique ID and belongs to one domain, which defines the location where the concept would be expected to occur within data tables of the OMOP CDM. Table 4 lists various examples of extracted entities mapped to standard concepts. In the first two lines of examples, the name of entity is the same with its mapped concept, meaning the entity itself is already a standard concept. In the third example, the names of entity and concept are similar but not exactly the same. In example 4, the entity name is an abbreviation of the standard concept. The entity and mapped concept have totally different names in example 5 and 6, but they are semantically equivalent. In COVIC, all types of entities are mapped to standard concepts.

### 2.2. Semi-automatic annotation and normalization

To maximize annotation quality while minimizing manual annotation effort, manual and machine-assisted annotations are combined in a semi-automatic approach. First, trials with multiple cohorts are manually identified and labeled. Next, an information extraction tool is used to automatically locate and classify entities in eligibility criteria, and the recognized entities results are verified and corrected manually by medical domain experts to overcome the limitations in the tool. Next, Attributes and relationships are manually annotated by domain experts because existing criteria extraction tools could not achieve satisfactory performance due to the large number of attribute and relationship types defined in our annotation model. Finally, extracted entities are mapped to their corresponding concepts and normalized automatically by the concept mapping and normalization modules. Fig. 3 shows the semi-automatic annotation process.

An information extraction tool Critera2Query [5] is used in this study to automatically identify entities in eligibility criteria. Criteria2Query is a natural language interface that transforms eligibility criteria into executable query for cohort definition, and can be used for entity recognition. Compared to other NER tools like cTAKES [23] or MetaMap [24], Criteria2Query follows OMOP CDM with considerable precision (~90%) and recall (~71%) rates. Eligibility Criteria with identified entities are imported to an annotation tool Brat for verification, and to be annotated for attributes and relationships. Brat is a web-based, interactive annotation environment with a visualization frontend interface [20]. Entity or attribute annotation can be defined using a contiguous span, beginning at the start of a phrase and ending at the completion of the phrase to capture instances rather than individual word tokens. An entity can be linked by multiple entities or attributes to build various types of relationships. The annotation makes a directed acyclic graph, which can be easily transformed into Boolean logic to form a database query. Fig. 4 shows an example of annotated criterion with Brat tool.

The annotated eligibility criteria include extracted name entities and their domain tags, which can feed into the concept mapping module to obtain each entity's mapped concept name and ID. Usagi is used in this study to assist with concept mapping. It is a software tool created by the OHDSI team and used in the process of mapping codes from a source system into the standard terminologies stored in the OMOP vocabulary [12]. With the free-text string as input, medical concept with the highest 'mapping accuracy score' from the returned results will be selected as the mapped concept. If a mapped concept suggested by Usagi is not accepted by the expert after reviewing it, the expert will suggest a new mapping, and a "entity-concept" dictionary will be automatically updated. The dictionary is built and maintained to complement Usagi mapping in cases where there is no appropriately mapped concept. A segment of the dictionary is shown in Table 5.

The "value" and "multiplier" attributes are normalized automatically as numerical values with upper and lower boundaries. For example, the value attribute '>20%' in the phrase '>20% of the body surface' will be coded into a minimum boundary as '20%' and maximum boundary as 'infinity' for 'body surface'. Supplementary Table 1 lists the mathematical operators in word (string type) expression and its corresponded numerical symbol. Regular expressions are used to identify and covert different type of operators. For the normalization of "temporal" attribute, all temporal expressions are unified to the same unit (days) by SUTime from Standard NLP group[21]. For example, "for at least 1 year before the screening visit" will be coded into 365 days before the 'the screening visit'. After normalizations, these attributes are converted from strings to numerical data types and are comparable in a quantitative manner. After labelling the five key criteria for COVID-19 trials, all annotated data are integrated together and reviewed and updated by domain experts. The structured dataset will be generated after that.

### 2.3. Annotation effort assessment

As discussed above, manual review was included in each step of the annotation schema to overcome the limitations of automated tools and ensure the corpus quality, we empirically assessed the annotation effort. Given the nature of the pandemic and the focus on rapid dissemination of annotated data at the outset of this project, specific tracking of the time needed to annotate each trial was not performed. However, in speaking with the annotators, early annotation work required around 45 min to one hour per trial which was reduced to around 15–20 min per trial at the end of the research effort as the annotators were more comfortable with the brat software and the iteratively generated mapping dictionary improved first-pass concept mapping results. The greatest amount of manual time for annotation review occurred at the eligibility criteria layer in adjusting entity recognition. Such an emphasis on manual review was due to our desire to ensure completely accurate data in our annotated corpus. To be specific, in the study cohort layer, the annotation is straightforward because most eligibility criteria contain explicit indicators such as "study group x" or "study cohort x"; in the edibility criteria layer and named entity layer, the annotation time depends on the length of criteria and number of included instances, and it usually takes 12–15 min; in the standard concept layer, new appeared concepts will be automatically added to the "entity-concept" dictionary and used for future annotations, so the annotation cost of this layer is supposed to continue to decrease as more and more concepts are

**Table 4**
Examples of entity and standard concept.

| # | Entity Name | Domain | Standard Concept | Concept ID |
|---|---|---|---|---|
| 1 | acute hepatic failure | Condition | acute hepatic failure | 4026032 |
| 2 | discharged from hospital | Observation | discharged from hospital | 4084843 |
| 3 | drug addiction | Condition | Dependent drug abuse | 4275756 |
| 4 | CPAP | Device | Continuous positive airway pressure (cpap) device | 2616666 |
| 5 | shortness of breath | Condition | Dyspnea | 312437 |
| 6 | solid tumor | Condition | Neoplasm | 4030314 |

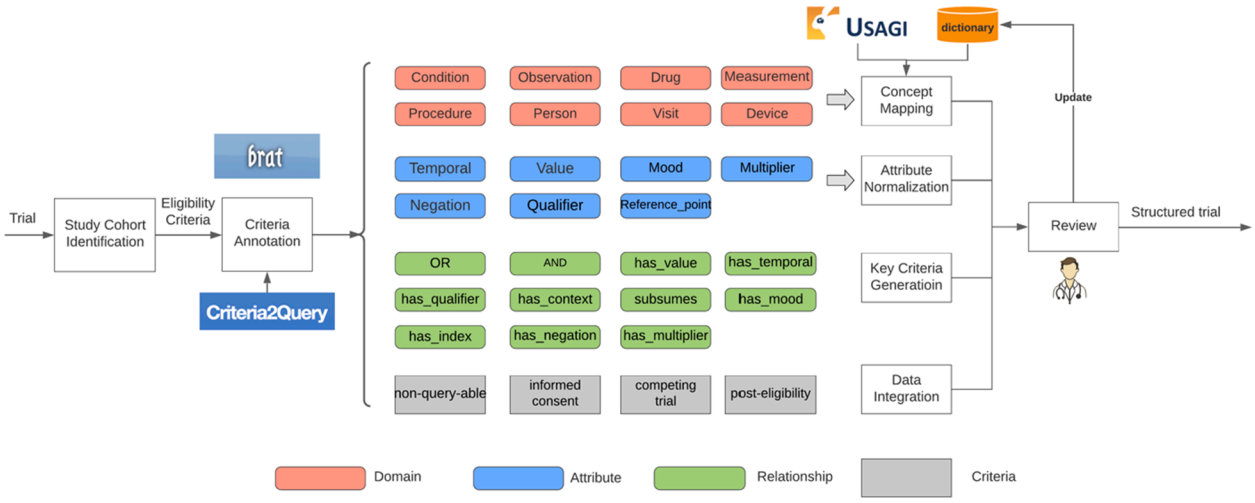


**Fig. 3.** Semi-automatic eligibility criteria annotation and normalization.

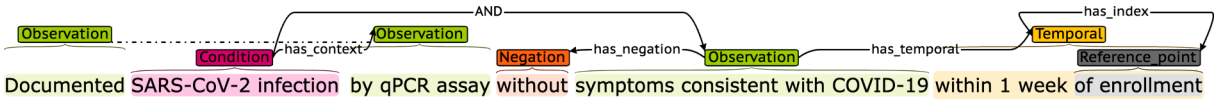


**Fig. 4.** Example of the eligibility criteria annotations on Brat tool.

**Table 5**
A short segment of the dictionary.

| Entity Source Text | Domain | Concept ID | Concept Name |
|---|---|---|---|
| GI perforation | Condition | 4202064 | Gastrointestinal perforation |
| encephalitis | Condition | 378143 | Encephalitis |
| oxygen saturation (pulse oximetry) | Measurement | 4310328 | Blood oxygen saturation |
| individual NEWS parameters | Measurement | 44808684 | National early warning score |
| hematopoietic stem cell transplant | Procedure | 4120445 | Hemopoietic stem cell transplant |
| transthoracic echocardiogram | Procedure | 4335825 | Transthoracic echocardiography |

The dictionary is a lookup table that saves all the manually corrected mappings and will be used for concept mapping before applying Usagi. It ensures the mappings updated by domain experts will be applied first in order to preserve the quality of annotations. The update of the dictionary follows an iteratively incremental learning process and all inappropriate mappings in previous annotations are avoided in mapping annotations of new trials.

included in the dictionary. It takes about 3–5 min for each trial in the current stage.

### 2.4. Validation

The creation of COVIC was performed by medical professionals (AB and LS). An annotation manual was developed to guide the data processing and format conversion. A medical terminology search engine Athena (http://athena.ohdsi.org) that provides searching of concepts in the OMOP CDM is used to assist experts in annotations and reviews in case of doubt for some concepts. For the first 100 trials, both annotators regularly discussed adaptations to the annotation model based on their experience and reviewed each other's work to make sure their annotations were consistent. Next, each annotator received a separate set of criteria loaded in brat and hand-created the entities and relationships as expressed above. All annotated trials will be reviewed by a third annotator (YS) to check whether there are missing annotations.

When the annotations for all 700 trials are finished, we randomly selected 70 trials (10%) from them to evaluate the inter-annotator agreement. Each text-bound annotation in Brat is an annotation instance. Two identical instances are considered as instance-level agreement. If two annotators agree on the annotation of an entity, they agree on both the entity span and the category. For example, the italic phrase in the criteria "subjects with a history of *systemic autoimmune diseases*" was annotated as "Condition" type entity. An agreement is counted if and only if annotators consistently recognized the sequence "*systemic autoimmune diseases*" and classified it as a "Condition" type entity. Each document can be split up into multiple tokens. If annotations on the same token are identical, annotators achieve token-level agreement. For example, entity "tamoxifen citrate" annotated as "Drug" type is an instance. It contains two tokens: "tamoxifen" and "citrate", and they are both "Drug" type entities. Given a clinical trial annotated separately by two annotators, the "instance-level" agreement is calculated by $2|A \cap B|/(|A|+|B|)$[25], where $A$ and $B$ are sets of instances created by the two annotators. The "token-level" agreement is also be calculated by $2|A \cap B|/(|A|+|B|)$ while $A$ and $B$ are sets of tokens. For entities and attributes, both instance level and token level agreements are automatically measured by a Python tool "Bratiaa"[22]. For relationships, instance level agreements are manually calculated.

## 3. Results

In this section, we discuss the detailed description of the COVIC dataset. COVIC contains 44, 622 annotations for 11,710 inclusion and exclusion eligibility criteria from 700 COVID-19 trials conducted in the United States. 39 trials have multiple study cohorts and they were divided into separate trials each of which includes eligibility criteria for one study cohort. The descriptive statistics are computed from the trial metadata and listed in Table 6.

Most of the trials are interventional (71.86%) and randomly allocated (54.86%) studies. 27.71% of the studies focus on Phase 2. Places like New York, California and Massachusetts states have the most trial recruitment sites and trial seekers living around these states may find eligible trials more easily than others in US. We also notice that 40.00% of the trials have no information on the type of allocation and 43.14% of

**Table 6**
Statistical information of 700 trials: **6.A** Study Type, **6.B** Study Allocation, **6.C** Study Phase, **6.D** Trial Locations.

Table 6.A

| Study Type | Count | % |
|---|---|---|
| Interventional | 503 | 71.86% |
| Observational | 153 | 21.86% |
| Observational [Patient Registry] | 31 | 4.43% |
| Expanded Access | 13 | 1.86% |

Table 6.B

| Allocation | Count | % |
|---|---|---|
| Randomized | 384 | 54.86% |
| Non-Randomized | 36 | 5.14% |
| N/A | 280 | 40.00% |

Table 6.C

| Phase | Count | % |
|---|---|---|
| Early Phase 1 | 13 | 1.86% |
| Phase 1 | 43 | 6.14% |
| Phase 1/Phase 2 | 28 | 4.00% |
| Phase 2 | 194 | 27.71% |
| Phase 2/Phase 3 | 23 | 3.29% |
| Phase 3 | 73 | 10.43% |
| Phase 4 | 24 | 3.43% |
| N/A | 302 | 43.14% |

Table 6.D

| Locations (Top 10 States) | Count | % |
|---|---|---|
| NY | 49 | 23.67% |
| CA | 35 | 16.91% |
| MA | 24 | 11.59% |
| MD | 20 | 9.66% |
| IL | 15 | 7.25% |
| FL | 15 | 7.25% |
| TX | 15 | 7.25% |
| PA | 13 | 6.28% |
| MN | 12 | 5.80% |
| NC | 9 | 4.35% |

the trials do not specify any phase stage.

For all the trials in COVIC, 18,161 entities were extracted and annotated by various domain names. Table 7 lists the counts and percentage of annotated entities in eight domains. "Condition", "observation", "drug" and "measurement" have relatively higher annotations than the other four domains, which makes sense in terms of the purpose of clinical trials. Of the 7, 743 annotations for attributes, most of them fall into "value", "temporal" and "qualifier" categories, as show in Table 8. We have 16,443 annotations for 11 types of relationships among entities and their attributes, and the statistics is provided in Table 9. Categories of "OR", "has_value" and "has_temporal" were annotated more than other relationships, which is consistent with the statistical results of annotations in domains and attributes.

Besides the above annotations based on terms, we also have 2,210 annotations for criteria from four categories. They usually have particular purposes different than most of the criteria as discussed in section 2.1. The counts and percentage of them are listed in Table 10.

In the criteria normalization, we have 17,171 entities were mapped to 4,140 different medical concepts, and the top 10 of the most common concepts mapped in each domain are listed in Table 11. Each concept is listed by its ID and name retrieved from OMOP CDM. As the table shows, concepts like "Disease caused by severe acute respiratory syndrome coronavirus 2", "Detection of 2019 novel coronavirus using polymerase chain reaction technique", "Artificial respiration" and "Oxygen therapy" are mapped more frequently which corresponds with the targeted disease COVID-19 of our dataset.

The same 70 trials were provided to the two annotators (AB and LS) to annotate independently using the Brat annotation tool to evaluate the inter-rater agreement. In total, 331 out of 1,732 entities, 204 out of 726 attributes, and 288 out of 1,538 relationships were annotated with different tags, making the inter-rater agreement of the annotations 80.9% (1,401/1,732) for entities, 72% (522/726) for attributes, and 81.3% (1,250/1,538) for relationships. Table 12 and Table 13 list the instance-level and token-level agreement rates of entities and attributes in different type and their average. Inter-rater agreement for each trial is listed in Supplementary Table 2.

A few terms with multiple interpretations bring the disagreements for the two annotators. For example, "COVID-19 PCR" can be treated as an entity from "measurement" domain, but it also makes sense for "COVID-19" as a 'condition' entity and "PCR" as a measurement entity. Further, when assessing agreement between specific annotation types, the lowest level of agreement is observed in "reference_point" annotation types. The reason is that a "reference_point" attribute always overlaps a long "Temporal" type attribute, and easily to be ignored during annotation. The disagreements in annotations will not have effect to the availability of COVIC.

**Table 7**
Total count and percentage (%) of annotated entities in 8 domains.

| Domain | Count | % |
|---|---|---|
| Condition | 6,614 | 36.42% |
| Observation | 4,243 | 23.36% |
| Drug | 2,179 | 12.00% |
| Measurement | 2,112 | 11.63% |
| Procedure | 1,364 | 7.51% |
| Person | 1,008 | 5.55% |
| Visit | 470 | 2.59% |
| Device | 171 | 0.94% |
| **In total** | **18,161** | **100.00%** |

**Table 8**
Total count and percentage of annotated entities in 7 attributes.

| Attribute | Count | % |
|---|---|---|
| Value | 2,832 | 36.57% |
| Temporal | 1,803 | 23.29% |
| Qualifier | 1,613 | 20.83% |
| Reference_point | 537 | 6.94% |
| Mood | 520 | 6.72% |
| Negation | 380 | 4.91% |
| Multiplier | 58 | 0.75% |
| **In total** | **7,743** | **100.00%** |

**Table 9**
Total count and percentage of annotated entities in 11 relationships.

| Relationship | Count | % |
|---|---|---|
| OR | 3,155 | 19.19% |
| has_value | 2,910 | 17.70% |
| has_temporal | 2,651 | 16.12% |
| has_qualifier | 1,599 | 9.72% |
| AND | 1,549 | 9.42% |
| has_context | 1,529 | 9.30% |
| subsumes | 1,382 | 8.40% |
| has_mood | 572 | 3.48% |
| has_index | 529 | 3.22% |
| has_negation | 509 | 3.1% |
| has_multiplier | 58 | 0.35% |
| **In total** | **16,443** | **100.00%** |

**Table 10**
Total count and percentage of the four particular types of criteria.

| Type | Count | % |
|---|---|---|
| Non-query-able | 1,112 | 50.32% |
| Informed consent | 478 | 21.63% |
| Competing trial | 331 | 13.08% |
| Post eligibility | 289 | 13.08% |
| **In total** | **2210** | **100.00%** |

**Table 11**
The count (#) of most common (top 10) concepts mapped in each domain.

| Condition | | | Observation | | | Drug | | | Measurement | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** |
| 37311061 | Disease caused by severe acute respiratory syndrome coronavirus 2 | 660 | 4188893 | History of | 523 | 21605200 | Corticosteroids | 78 | 37310255 | Detection of 2019 novel coronavirus using polymerase chain reaction technique | 192 |
| 4299535 | Patient currently pregnant | 372 | 40218805 | CDC laboratory | 148 | 21602457 | Prolactine inhibitors | 66 | 4146380 | Alanine aminotransferase measurement | 118 |
| 437663 | Fever | 133 | 4185135 | Breastfeeding | 124 | 1777087 | Hydroxychloroquine | 53 | 4263457 | Aspartate aminotransferase measurement | 114 |
| 312437 | Dyspnea | 118 | 36685445 | on room air at rest | 91 | 21603891 | IMMUNOSUPPRESSANTS | 52 | 4310328 | Blood oxygen saturation | 96 |
| 254761 | Cough | 104 | 37310260 | Close exposure to 2019 novel coronavirus infection | 52 | 21601386 | ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS | 43 | 4233883 | Ratio of arterial oxygen tension to inspired oxygen fraction | 86 |
| 439727 | Human immunodeficiency virus infection | 83 | 4120014 | Polymerase chain reaction | 43 | 2718732 | Immunosuppressive drug not otherwise classified | 42 | 4267147 | Platelet count | 74 |
| 316866 | Hypertensive disorder | 68 | 45766517 | Confirmatory technique | 40 | 1507835 | Vasopressin (USP) | 41 | 4313591 | Respiratory rate | 72 |
| 255573 | Chronic obstructive lung disease | 66 | 4244251 | Confirmed by | 40 | 40171288 | tocilizumab | 36 | 44806420 | Estimation of glomerular filtration rate | 69 |
| 443392 | Malignant neoplastic disease | 61 | 4289014 | Normal breast feeding | 35 | 1792515 | Chloroquine | 30 | 3027315 | Oxygen [Partial pressure] in Blood | 66 |
| 4212484 | Multiple organ failure | 60 | 4142947 | Symptomatic | 33 | 1309944 | Amiodarone | 23 | 44789311 | Pregnancy test | 60 |
| **Procedure** | | | **Person** | | | **Visit** | | | **Device** | | |
| **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** | **ID** | **concept** | **Freq.** |
| 4230167 | Artificial respiration | 151 | 4265453 | age | 567 | 38004515 | Hospital | 298 | 4139525 | High flow oxygen nasal cannula | 28 |
| 4239130 | Oxygen therapy | 123 | 442986 | female | 200 | 38004311 | Inpatient Hospice | 29 | 4138614 | BiPAP oxygen nasal cannula | 16 |
| 4052536 | Extracorporeal membrane oxygenation | 71 | 442985 | Male | 100 | 38004519 | Home Health Agency | 21 | 2614925 | Cannula nasal | 15 |
| 4032243 | Dialysis procedure | 67 | 4323831 | old | 34 | 32037 | Intensive Care | 19 | 2616666 | Continuous positive airway pressure (cpap) device | 14 |
| 4202832 | Intubation | 60 | 4046779 | Adult | 23 | 9201 | Inpatient Visit | 13 | 4145528 | Nonrebreather oxygen mask | 12 |
| 4273629 | Chemotherapy | 59 | 4119673 | year | 7 | 38004522 | Department Store | 10 | 4030875 | Cardiac pacemaker | 6 |
| 44790095 | Invasive ventilation | 51 | 1332764 | children | 9 | 38004284 | Psychiatric Hospital | 8 | 4232657 | Vascular stent | 5 |
| 4208341 | Solid organ transplant | 45 | 4305451 | Infant | 3 | 8717 | Inpatient Hospital | 8 | 4234106 | Metal periosteal implant | 4 |
| 37018292 | Continuous renal replacement therapy | 37 | 42073776 | Newborn | 5 | 8676 | Nursing Facility | 8 | 4148006 | Epidural catheter | 3 |
| 40486624 | Noninvasive positive pressure ventilation | 34 | 2090691 | Mothers | 3 | 9202 | Outpatient Visit | 8 | 45760696 | Spinal catheter | 3 |

**Table 12**
Instance-level and token-level agreement rates of entities in different types and the average (arithmetic mean for all trials).

| Type | Instance-level | Token-level |
|---|---|---|
| Measurement | 0.874 | 0.887 |
| Condition | 0.832 | 0.845 |
| Person | 0.79 | 0.830 |
| Drug | 0.781 | 0.825 |
| Visit | 0.765 | 0.792 |
| Procedure | 0.728 | 0.746 |
| Observation | 0.601 | 0.718 |
| Device | 0.565 | 0.714 |
| **Average** | 0.809 | 0.825 |

**Table 13**
Instance-level and token-level agreement rates of attributes in different types and the average (arithmetic mean for all trials).

| Type | Instance-level | Token-level |
|---|---|---|
| Temporal | 0.713 | 0.752 |
| Value | 0.927 | 0.954 |
| Negation | 0.653 | 0.693 |
| Qualifier | 0.66 | 0.741 |
| Multiplier | 0.678 | 0.764 |
| Reference_point | 0.344 | 0.372 |
| Mood | 0.624 | 0.643 |
| **Average** | 0.72 | 0.798 |

## 4. Usage discussion

As the first large structured dataset for COVID-19 clinical trials, COVIC can support the research community and facilitate the development of various applications on automatic cohort identification, fine-grained trial generalizability analysis or similarity comparison. It can also serve as a shared benchmark for machine learning based information extraction from free-text clinical trial eligibility criteria. We describe below four examples of data usages for different applications.

### 4.1. COVID-19 cohort identification

Researchers worldwide have worked diligently to understand the mechanisms of transmission and action for COVID-19 and to discover effective treatments and interventions. One important data source for COVID-19 research is patients' clinical data stored in electronic health records (EHRs). EHRs use structured data elements to document patient information with controlled vocabulary, but clinical trial eligibility criteria are usually written in free-text descriptions and the majority of which include semantically complex language hard for computational processing. It requires domain experts to understand the eligibility criteria and construct the query manually, which is time-consuming and error-prone.

With COVIC, the EHR data queries can be automatically generated with a few lines of extra codes. For example, in the criterion "Blood pressure<90 mm Hg systolic or 60 mm Hg diastolic recorded on at least two readings 30 min apart" (ID NCT04335123), "blood pressure systolic" and "blood pressure diastolic" are extracted and annotated as "measurement" entities and mapped to "systolic blood pressure (ID 4152194)" and "diastolic blood pressure (ID 4154790) " respectively. "<90 mm Hg", "<60 mm Hg" and "at least two" are normalized as numerical values. "Readings" and "30 min apart" are annotated as "observation" entities and normalized as "reading (ID 3243724)" and "Every thirty minutes as required (ID 45757503)". The original criterion can then be coded as "systolic blood pressure < 90 mm Hg" or "blood pressure diastolic < 60 mm Hg" with "≥ 2 reading" and "every thirty minutes as required". A few applications have been developed to convert such structured statements to queries for cohort definitions, such as Circle-be (https://github.com/OHDSI/circe-be) and Atlas (http://www.ohdsi.org/web/atlas/).

### 4.2. COVID-19 trial search

While we are fortunate to have a growing number of trials available for coronavirus therapies, it is often difficult to determine which specific studies are appropriate for individual patients. Keyword-based search engines allow users to search for trials using a combination of terms, though this method is often relatively nonspecific and can result in information overload when too many trials are returned. With the help of COVIC, semantic trial search engines can be developed for patients to find trials for which they are eligible and appropriate. Once the data are structured and coded, algorithms can be applied to match a patient to clinical trials based on their answers to eligibility questions. Rather than obtaining a long list of results, patients can receive a personalized list of a few trials for which they may be eligible. COVIC will be an enormous help to busy clinicians, allowing them to efficiently link patients with the trials best suited to them. A web application "COVID-19 Trial Finder" (https://covidtrialx.dbmi.columbia.edu) has already been developed by our team to provide patient-centered search of COVID-19 trials with the support of the COVIC dataset. It allows potential candidates to pre-screen their eligibility through a set of short medical questions with minimum effort. Its initial version was released in May 2020, and>2,300 page visits from 30 countries were recorded by Feb. 15th, 2021, including 1,348 visits from 40 states in the United States, and the average session duration was 2 min and 24 s, according to the report of Google Analytics.

### 4.3. COVID-19 trial collaboration analytics

Researchers across the world are racing to conduct COVID-19-related clinical trials. However, many of these trials have significant redundancy as well as limited or biased target patient populations. Clinical trial collaboration analytics tries to answer the question: given two clinical trials investigating the same scientific question, can we tell if they are studying comparable cohorts? The manual labor required from domain experts for such appraisal is prohibitive. Metadata such as study type, location, age, or sex information are usually used for coarse-grained trial comparison. COVIC provides more features within eligibility criteria to identify opportunities for collaborative recruitment such as the category or mapped concepts of annotated entities. They can be used for building recruitment optimization systems to minimize the unnecessary local competition among trial investigators in COVID-19 clinical trials. We have developed a prototype application "Collaboration Opportunities" (http://apex.dbmi.columbia.edu/collaboration/) based on COVIC dataset to support coordinated *meta*-analyses.

### 4.4. Machine learning based information extraction for clinical trials

Non-interoperable data cannot be interpreted by computers and this inhibits machine learning based NLP. Large machine-readable datasets are needed to assist the development and evaluation of the biomedical NLP systems. Most of previous datasets only include a few trials which are not sufficient enough to train robust machine learning models. Chia has a large size of annotations for 1,000 trials but only for the domains of entities, attributes, and their relationships. COVIC includes not only the three types of annotations that Chia provides, but also the mapped concepts for each identified entity. This allows COVIC to serve as a benchmark for both training and evaluative purposes for NLP related tasks like Named-entity Recognition or Named-entity Linking. As a manually annotated and validated large corpus, COVIC can be a knowledge base to catalyze clinical trial analytics and criteria extraction. A web application "Criteria2Query" (http://www.ohdsi.org/web/criteria2query/) was developed by our group for automatic clinical trial eligibility criteria extraction and query generation based on the structured datasets.

To further improve the usage of the proposed schema and corpus, a couple of limitations might need more work. First, each criterion was annotated only one kind of label in this work, but sometimes a criterion with multiple interpretations might have various types of annotations. For example, "women who have been pregnant" can be considered as an "Observation" entity as a whole, but the token "women" and "pregnant" can also be labelled by "Person" and "Condition" types separately: "*women* [Person] who have been *pregnant* [Condition]". "COVID-19 PCR" can be treated as an entity from "measurement" domain, but it also makes sense for "COVID-19" as a 'condition' entity and "PCR" as a measurement entity: "*COVID-19* [Condition] *PCR* [Measurement]". Eligibility criteria with different types of annotations can be further explored. Second, although the annotation process is semi-automatic, it still took lots of human efforts to check and update the annotations due to the limitations of NER and concept normalization tools. Also, only new appeared criteria need to be focused by annotators and repetitive annotations should be automatically filtered or labelled. Future directions aim to improve the performance and efficiency of the automated tools. Third, COVIC contains a large number of annotations for 700 COVID-19 trials, but considering the nearly 4,800 COVID-19 trials registered in ClinicalTrials.gov as of Feb. 2021, it is worthy to expanding the corpus with more annotated trials.

## 5. Conclusions

In the context of the COVID-19 health crisis, a multitude of clinical trials have been designed to seek effective ways to manage this pandemic. In this research, we propose a semi-automatic clinical trial annotation method and contribute a multi-purpose large annotated dataset COVIC with structured and semantically annotated criteria for 700 COVID-19 clinical trials. We provide a detailed description of the method and the dataset, present examples of how it can assist in patient recruitment and provide insights into trial collaboration opportunities. COVIC is currently being used in a number of active systems and exhibits great analytical flexibility. It promises to support more applications.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the National Library of Medicine grant R01LM009886-11 (Bridging the Semantic Gap Between Research Eligibility Criteria and Clinical Data) and National Center for Advancing Clinical and Translational Science grants UL1TR001873 and 3U24TR001579-05.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103790.

## References

[1] Q. Zheng, F.K. Jones, S.V. Leavitt, L. Ung, A.B. Labrique, D.H. Peters, E.C. Lee, A. S. Azman, HIT-COVID, a global database tracking public health interventions to COVID-19, Sci. Data 7 (1) (2020) 1–8.
[2] COVID-19 Map - Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html Accessed September 31, 2020.
[3] N.A. Sansa, Effects of the COVID-19 Pandemic on the World Population: Lessons to Adopt from Past Years Global Pandemics (2020). Available at SSRN 3565645.
[4] T. Kang, S. Zhang, Y. Tang, G.W. Hruby, A. Rusanov, N. Elhadad, C. Weng, EliIE: An open-source information extraction system for clinical trial eligibility criteria, J. Am. Med. Inform. Assoc. 24 (6) (2017) 1062–1071.
[5] C. Yuan, P.B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang, C. Weng, Criteria2Query: a natural language interface to clinical databases for cohort definition, J. Am. Med. Inform. Assoc. 26 (4) (2019) 294–305.
[6] Y. Sun, A.M. Butler, F. Lin, H. Liu, L.A. Stewart, J.H. Kim, B.R.S. Idnay, Q. Ge, X. Wei, C. Liu, C. Yuan, C. Weng, The COVID-19 Trial Finder, J. Am. Med. Inform. Assoc. 28 (3) (2021) 616–621.
[7] Desvars-Larrive, A., Dervic, E., Haug, N., Niederkrotenthaler, T., Chen, J., Di Natale, A., Lasser, J., Gliga, D.S., Roux, A., Chakraborty, A. and Ten, A., 2020. A structured open dataset of government interventions in response to COVID-19. medRxiv.
[8] Y. Sun, K. Loparo, Information extraction from free text in clinical trials with knowledge-based distant supervision. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), IEEE 2019, July, Vol. 1, pp. 954–955.
[9] J. Ross, S. Tu, S. Carini, I. Sim, Analysis of eligibility criteria complexity in clinical trials, Summit Transl. Bioinform. 2010 (2010) 46.
[10] C. Weng, X. Wu, Z. Luo, M.R. Boland, D. Theodoratos, S.B. Johnson, EliXR: an approach to eligibility criteria extraction and representation, J. Am. Med. Inform. Assoc. 18 (Supplement_1) (2011) i116–i124.
[11] Observational Health Data Sciences and Informatics. Usagi, https://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi (2018).
[12] F. Kury, A. Butler, C. Yuan, L.H. Fu, Y. Sun, H. Liu, I. Sim, S. Carini, C. Weng, Chia, a large annotated corpus of clinical trial eligibility criteria, Sci. Data 7 (1) (2020) 1–11.
[13] C. Reich, P.B. Ryan, R. Belenkaya, K. Natarajan, C. Blacketer, OHDSI Common Data Model v6.0 Specifications, https://github.com/OHDSI/CommonDataModel/wiki (2019).
[14] J.S. Ross, T. Tse, D.A. Zarin, et al., Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis, d7292-d7292, BMJ 344 (2012), https://doi.org/10.1136/bmj.d7292.
[15] Clinical Trials Transformation Initiative. Aggregate Analysis of ClinicalTrials.gov, https://aact.ctti-clinicaltrials.org/ (2016).
[16] i2b2 Common Data Model. https://i2b2.org/software/files/PDF/current/CRC_Design.pdf. Accessed 25 Aug 2020.
[17] Sentinel Common Data Model. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model. Accessed 25 Aug 2020.
[18] S. Toh, L.J. Rasmussen-Torvik, E.E. Harmata, R. Pardee, R. Saizan, E. Malanga, J.L. Sturtevant, C.E. Horgan, J. Anau, C.D. Janning, R.D. Wellman, The National Patient-Centered Clinical Research Network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. JMIR research protocols, 6(12) (2017) p.e222.
[19] E.A. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, F.J. DeFalco, A. Londhe, V. Zhu, P.B. Ryan, Feasibility and utility of applications of the common data model to multiple, disparate observational health databases, J. Am. Med. Inform. Assoc. 22 (3) (2015) 553–564.
[20] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J.I. Tsujii, April. BRAT: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
[21] A.X. Chang, C.D. Manning, Sutime: A library for recognizing and normalizing time expressions, in Lrec, 2012, May ,Vol. 2012, pp. 3735–3740.
[22] T. Kolditz, C. Lohr, J. Hellrich, L. Modersohn, B. Betz, M. Kiehntopf, U. Hahn, August. Annotating German Clinical Documents for De-Identification. In MedInfo, 2019, pp. 203–207.
[23] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[24] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, 2001, p. 17. American Medical Informatics Association.
[25] G. Hripcsak, A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, J. Am. Med. Inform. Assoc. 12 (3) (2005) 296–298.