Structuring Clinical Trial Eligibility Criteria with the Common Data Model

Gal Levy-Fix, MA[1], Anil Yaman, MS[1], and Chunhua Weng, PhD[1]
[1]Department of Biomedical Informatics, Columbia University in the City of New York, NY

Abstract

This paper presents a method for classifying and structuring free-text clinical trial eligibility criteria using the OMOP Common Data Model (CDM). Our method was applied to eligibility criteria text available from the largest clinical trial repository ClinicalTrials.gov. Structurally complex criteria were simplified and rewritten as simpler sentences connected by logical operators: AND or OR. Semantic annotation, using the Unified Medical Language System (UMLS) and other support semantics, was performed before criteria were clustered into groups, which were further classified into selected CDM domain categories according to the semantic type combination patterns in each cluster. Key criteria attributes were then extracted and annotated using the OMOP CDM.

Introduction

Structuring free-text eligibility criteria from clinical trial summaries has a multitude of applications. For example, structured clinical trial eligibility criteria (CTEC) can be effectively compared to electronic health records and facilitate personalized patient care by improving patient selection to clinical trials and allowing physicians to identify trial findings most relevant to the their patients. Structured CTEC can enable more rigorous comparison of trial populations and better meta-analyses of collective generalizability of related clinical trials and also pave the path to thorough studies of clinical trial participants and their representativeness of the general population.

We set out to standardize and structure CTEC available at ClinicalTrials.gov[1] through an annotation approach using the widely adopted Common Data Model (CDM) (http://omop.org/CDM) specification developed by the Observational Medical Outcomes Partnership (OMOP)[2]. One foreseeable advantage of using this model to standardize structured output of CTEC, is that more and more observational databases are modeled using the OMOP CDM data schema; therefore, structured CTEC modeled using OMOP CDM can be easily applied across a large number of observational databases for cohort selection, population health analytics, and other comparative effective research and translational medicine studies.

Of the 18 data tables in version 4.0 of the CDM[3], we chose to categorize the CTEC according to the tables most relevant to CTEC: *Person*, *Drug Exposure*, *Condition Occurrence*, *Observation*, *Procedure Occurrence*, and *Visit Occurrence*. We then extracted criteria attributes according the fields of the CDM tables for criteria categorized as *Person*, *Drug Exposure*, and *Condition Occurrence*. In comparison to a gold standard, the overall F-scores for criteria categorization and attribute extraction (for correctly classified criteria) were 75% and 81%, respectively. We created a web application (http://is.gd/EliXR) to process and structure CTEC text provided one or multiple trial NCT IDs separated by semicolons.

Our method builds on prior work done on text indexing and categorization of CTEC. Miotto et al.[4,5] utilize the Unified Medical Language System (UMLS)[6] lexicon to implement unsupervised mining of frequent semantic tags to index clinical eligibility. Hao et al.[7] automatically identify and cluster clinical trials with similar features. Luo et al.[8] combined unsupervised clustering and supervised methods to classify CTEC. Tu et al.[9] classified CTEC, extracted atomic elements, and structured them into a computable ontology. To the best of our knowledge, this study is among the first to conform CTEC data using a standard data model such as the OMOP CDM.

Methods

In structuring the CTEC, we performed two main tasks: 1) CTEC classification; and 2) extraction of data attributes from CTEC as structured output using the classification results from the previous task. In the first stage of our method, we processed the CTEC into individual sentences. Using the UMLS Metathesaurus semantic types and our own constructed support semantics (used to tag generic medical words such as "diagnosis"), each sentence in the criteria was annotated with semantics associated with the CDM domains (e.g. CONDITION and OBSERVATION). Following the general method outlined by Luo et al.[8], we used hierarchical clustering to group similar CTEC sentences. Each cluster was then classified into a CDM domain. According to the classification result, we searched for semantic patterns in the CTEC to identify key attributes to extract into structured data tables.

Our design is based on a strategy called "structured narratives" developed by Johnson et al.[10]. This model supports semi-structured annotation of free-text CTEC so that data attributes such as numerical amounts and measure units, are labeled with corresponding CDM domain data fields. Thus, a phrase like "*within 3 months of randomization*" is recognized as temporal constraint, but its components are not further analyzed. We can later apply a specialized temporal expression parser to perform deep parsing on such data elements[11,12]. The major benefit of this method is that it strikes the balance between cost-effectiveness and tradeoffs for annotation and deep natural language processing. Instead of fully parsing every linguistic unit, we focused on annotating the most important data elements that are useful for eligibility determination with the Common Data Model. Table 1 illustrates how three example criteria are structured into tables based on their classifications: *Person, Drug,* and *Condition*. Figure 1 depicts the steps of our structuring method. The details of each step are provided afterwards.

Table 1. Example output for structured eligibility criteria

Eligibility Criteria	Table Name	Table Entries			
"men , between the ages of 18 and 70 years of age , inclusive , at the time of randomization"	*Person*	**Type**	**Gender**	**Age Min.**	**Age Max.**
		inclusion	men	18	70
"oral prednisone < = 20 mg/day is permitted"	*Drug*	**Type**	**Drug Concept**	**Dosage**	**Temporal Constraint**
		inclusion	oral prednisone	< = 20 mg/day	N/A
"recent history of repeated infections"	*Condition*	**Type**	**Condition Concept**	**Temporal Constraint**	
		exclusion	infections	recent history	

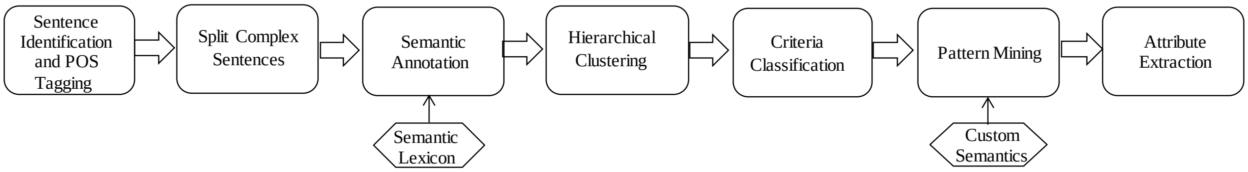

Figure 1. Our overall method framework

1. Criteria Sentence Identification and Complexity Reduction

CTEC summaries were downloaded from ClinicalTrials.gov and pre-processed to remove blank spaces and any unwanted symbols. Criteria were identified as "inclusion" or "exclusion" criteria when applicable and separated into sentences. Sentences that introduced a list of criteria (e.g. "administration of the following agents:") were removed, as they did not contain information that is useful for eligibility determination. Eligibility records were assigned part-of –speech (POS) tags using the package SharpNLP[13]. For instance, the sentence "*total CLASI activity must be >=10*" was assigned the following POS tags: "*total/JJ CLASI/NN activity/NN must/MD be/VB >/SYM =/SYM 10/CD*". Using the POS tags we broke down complex sentences with multiple concepts into simpler atomic criteria linked by logical operators: AND or OR. CTEC containing independent concepts separated by coordinating conjunctions like "or" and "and" were split into separate criteria. Criteria with lists of concepts were rewritten as new simplified criteria (see Figure 2 for examples).

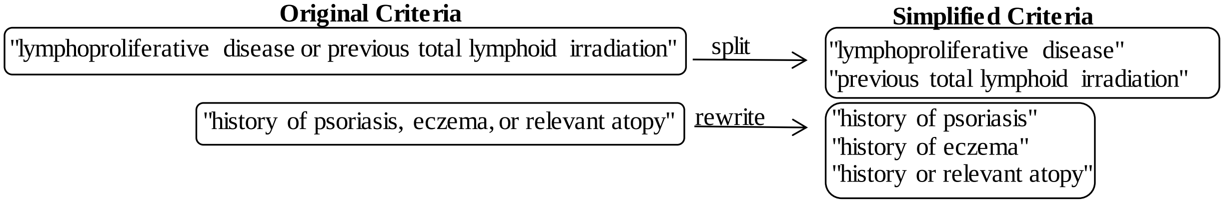
Original Criteria
Simplified Criteria
"lymphoproliferative disease or previous total lymphoid irradiation"
split
"lymphoproliferative disease"
"previous total lymphoid irradiation"
"history of psoriasis, eczema, or relevant atopy"
rewrite
"history of psoriasis"
"history of eczema"
"history or relevant atopy"

Figure 2. Examples for splitting and rewriting complex sentences

2. Semantic Annotation

Our next objective was to annotate the CTEC with semantic types associated with the selected CDM domains (i.e., *Person*, *Drug Exposure*, *Condition Occurrence*, *Observation*, and *Visit*). We used the method developed for eTACTS[5] to extract and tag relevant n-grams from each criterion with a semantic type from the UMLS lexicon. Standardized UMLS semantic groups[14] were used to map the UMLS semantic types to CDM semantic types associated with the CDM domains listed above. We reassigned several semantics to CDM domains since the UMLS semantic groups were not compatible with the CDM domain definitions. For instance, since no UMLS group appropriately identified observation concepts we reassigned the semantic types *Laboratory or Test Result* and *Laboratory Procedure* to the *Observation* CDM category instead of *Procedure*, which was for therapy or treatment procedures.

In addition, to help further distinct between the criteria CDM categories, we compiled a list of generic medical words common in CTEC which we refer to as support words. To each support word we assigned a semantic type corresponding to a CDM domain. For instance, the appearance of the word "*diagnosis*", which we designated as a support word, helps determine that the sentence should be classified as a *Condition Occurrence*. The resulting annotation, with the UMLS and support semantic types, for the sentence "*diagnosis of discoid lupus erythematosus (DLE) with or without SLE*" is the following:

> *"[diagnosis]|CONDITION_SUPPORT| of [discoid lupus erythematosus]|CONDITION|*
> *([DLE]|CONDITION|) with or without [SLE]|CONDITION|."*

3. Hierarchical Clustering and CDM Domain Criteria Classification

The frequency of each CDM core semantic type (derived from the UMLS) and support semantic type was extracted to create a criteria–semantic matrix, with the features corresponding to the semantic types and rows corresponding to the criteria. The frequencies of core semantics (e.g. CONDITION) and their corresponding support semantics (e.g. CONDITION_SUPPORT) were combined to create one master CDM semantic feature since they imply the same criteria category. Moreover, the frequency of the support semantics indicating dosage units (e.g. "ml") and measurement units (e.g "pounds") were added to the drug semantic and observation semantic, respectively. Each value in the matrix represents the sentence- semantic-share of the corresponding semantic from all the semantics in the sentence. Using the criteria – semantic matrix, agglomerative hierarchical clustering was performed to group together criteria with a similar semantic type combination pattern. We used Pearson correlation as the distance measure between the sentence-semantic-shares[8]. We elected a distance threshold of 0.4, low enough so that clusters were composed of fairly similar sentences.

We then generated a cluster- semantic matrix by adding up the frequency of each semantic type in the clustered sentences and divided by the total number of semantic types. Sentences in each cluster were assigned a CDM category according to the semantic type with the majority vote within the cluster. Sentences identified to be of type *Condition Occurrence* or *Drug Exposure* but contained an OBSERVATION_SUPPORT semantic (associated with words such as "level" or "limit") were reclassified as *Observation* type. Criteria that were classified as one of the CDM domains (e.g. *Person*) but also contained semantics corresponding to other CDM domains (e.g. *Condition*) were reclassified as multiple domain criteria (e.g. *Person/ Condition*). Therefore, our classification supported poly-hierarchy. Clusters with semantics not corresponding to the CDM domains we selected were classified as *Other.*

4. Pattern Search and Attribute Extraction

At this stage of our work we limited our focus to the extraction of high level attributes according to the CDM table fields for the following criteria categories: *Person*, *Drug Exposure*, and *Condition Occurrence*. The method can be easily expanded to cover all other classes and other attributes in the CDM given more time, which will be part of our future work. The attributes extracted by category are presented in Table 3.

Table 3. Attributes extracted by OMOP CDM class

Criteria Class	**Attributes Extracted**
Person	gender; age
Drug Exposure	drug concept; dosage; temporal constraints
Condition Occurrence	condition concept; temporal constraints

The gender, age, drug, and condition concepts were recognized and extracted using CDM-based semantic annotation discussed earlier. Supporting attributes such as temporal constraints, dosage, and numeric values were extracted using a list of pre-specified semantic patterns, following an unpublished method. Figure 3 shows how key attributes are extracted from the sentence example: "*receipt of live vaccine with 3 months of randomization*".

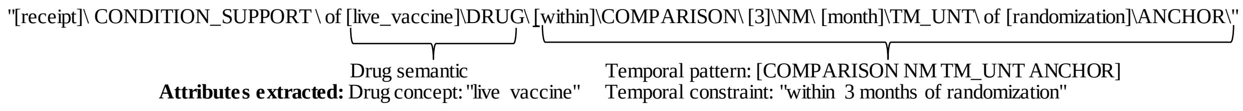
"[receipt]\ CONDITION_SUPPORT \ of [live_vaccine]\DRUG\ [within]\COMPARISON\ [3]\NM\ [month]\TM_UNT\ of [randomization]\ANCHOR\"
Drug semantic
Temporal pattern: [COMPARISON NM TM_UNT ANCHOR]
Attributes extracted:
Drug concept: "live vaccine"
Temporal constraint: "within 3 months of randomization"

Figure 3. Attribute extraction

The attributes were extracted and entered into a structured CDM-compliant table. The table was then checked for duplicative information once the attributes were extracted from the entire CTEC summary.

Results

1. Online System Demo

Our CTEC structuring tool is available online at http://is.gd/EliXR. A single NCT ID can be entered to structure a single trial (e.g. NCT01164917) or multiple NCT IDs can be entered with a semicolon between them to structure multiple trials at the same time (e.g. NCT01164917; NCT01597050).

2. Preliminary System Evaluation

To evaluate the accuracy of the CTEC classification and attribute extraction, five trials were randomly selected for manual classification and attribute extraction. The 95 CTEC sentences extracted from these five trials were assigned to two evaluators. The two human evaluators independently classified each sentence in the CTEC and identified key attributes for classes *Person*, *Drug Exposure*, and *Condition Occurrence*. A gold standard was created based on the consensus of the two evaluators. First, the gold standard was compared to the resulting automated classification. Then we compared the gold standard to the attribute extraction for the criteria classified correctly, as one of the three classes stated above. Three main accuracy measures were calculated: Precision, Recall, and F-score (their definitions are below). Each accuracy measures was calculated for each class category and transformed into a weighted average using the relative frequency of each category as its weight.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \qquad \text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \qquad \text{F-Score} = \frac{2{*}\text{Precision}{*}\text{Recall}}{\text{Precision} \ + \text{Recall}}$$

As shown in Table 4, the precision measure for the criteria classification was found to be 76%, the recall 76%, and F-score 75%. For the 38 criteria correctly classified as *Person, Condition Occurrence,* or *Drug Exposure* the overall attribute extraction precision rate was 88%, the recall was 75%, and F-score 81%.

Table 4. Accuracy evaluation results for 95 eligibility criteria eligibility criteria from five random clinical trials

	Precision	**Recall**	**F-score**
Criteria Categorization (weighted average)	76%	76%	75%
Attribute Extraction (weighted average)	88%	75%	81%

Discussion

Error Analysis: Incorrect Semantic Annotation

Errors in classification of the test criteria were mostly caused by erroneous or no semantic annotation, sensitivity in multi-category classification rules, meaning changes due to complexity reduction, and context hidden in the free-text structure. Several semantic annotation errors were caused by dubious UMLS semantic type annotations, also documented by Fan and Friedman[15]. For instance, a criterion was classified as a *Condition Occurrence* instead of *Other* since the n-gram "*non-English speaking*" was tagged with the UMLS semantic type "Finding", mapped to the disorder semantic group. Some key phrases were missing semantic annotations due to the incompleteness of the UMLS lexicon or the exact match requirement in eTACTS method. The phrase "*Ashkenazi Jewish*" was not tagged with a semantic type since the exact phrase was not found in the lexicon. Allowing for a partial phrase match would have matched to the phrase "*Ashkenazi Jewish religion*" contained in the UMLS and assigned the semantic type "Population Group" which we mapped to the CDM sematic PERSON.

Errors in rewriting criteria also led to classification errors. For instance the sentence "*history of alcohol or drug abuse in past 5 years*" was rewritten as "*history of alcohol in past 5 years*" and "*history of drug abuse in past 5 years*". The semantic annotation of "*alcohol*" instead of "*alcohol abuse*" led to the classification of the sentence as *Drug Exposure* rather than *Condition Occurrence.* Errors in attribute extraction of drug and condition concepts were mostly caused by faulty semantic annotation described above. One such example is the phrase "*new diagnostic*" which was erroneously extracted as a condition from the sentence "*any active disease process requiring new diagnostic and therapeutic plans"* since it was assigned the UMLS semantic type "Finding" and mapped to the CDM domain *Condition Occurrence*. Such a mistake could be avoided if the generic medical word "*diagnostic*" was included in our support list or if the UMLS semantic type was corrected to be "*Diagnostic Procedure*", mapped to the CDM domain *Observation*. Issues with temporal extraction were caused by semantic patterns that we overlooked in our predefined list of patterns. For instance, the pattern "*current or history*" was unidentified in the sentence "*cerebral vascular or coronary artery disease (current or history)*".

Limitation and Future Work

Our future work lies in two areas: NLP and data modeling. To improve the performance of the semantic annotation we plan to improve sentence simplification, temporal expression extraction, numerical expression extraction, and coreference resolution. We will continue expanding the attribute extraction and standardization to other CDM domains such as *Observation* and *Procedure*. We will also continue to expand the CDM to better cover the data elements and classes in CTEC text. Additional tables may need to be added to structure criteria describing *Patient Behavior* or *Attitude* concepts such as "*tanning*" or "*consent*". Furthermore, to fully conform the criteria standardization to the CDM additional temporal features would be required in order to capture the time frame information, common in CTEC categorized as *Drug Exposure* or *Condition Occurrence*. For instance, the sentence "*receipt of a live vaccine within 3 months of study randomization"* indicates that patients who received a live vaccine 3 months or closer to the study randomization time are eligible to participate. As we mentioned above, we will use our previously developed temporal parser to enrich this semantic annotation method presented here. Moreover, under the current CDM specification such a sentence would be categorized as *Drug Exposure;* however, the start and end exposure date fields in the CDM table are not sufficient to represent such a temporal condition.

Conclusion

We developed a method to automate CTEC classification and annotate data attributes as tables following the specifications of the CDM standard. Our comparison to a gold standard shows that the F-score for our method was 75% for CTEC classification 81% for the attribute extraction. Further expansion of our method to the remaining criteria categories will improve the comprehensiveness of this method and enable effective study of the populations represented in clinical trials as well as the comparison among clinical trials and with electronic health records.

Acknowledgments

This research is sponsored by National Library of Medicine Grant **R01LM009886** (PI: Weng) and National Center for Advancing Clinical and Translational Science (NCATS) **U01 TR000040** (PI: Ginsberg).

References

1. *ClinicalTrials.gov*. September 2014; Available from: http://clinicaltrials.gov/.
2. Observational Medical Outcomes Partnership. Common Data Model; Available: http://omop.org/CDM.
3. Observational Medical Outcomes Partnership. CDM specifications version 4.0. April 2014.
4. Miotto R and Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform 2013;46(6):1145-51.
5. Miotto R, Jiang S, Weng C. eTACTS: a method for dynamically filtering clinical trial search results. J Biomed Inform 2013;46(6):1060-7.
6. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The unified medical language system. J Am Med Inform Assoc 1998;5(1):1-11.
7. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform 2014; pii:S1532-0464(14):00011-2.
8. Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. J Biomed Inform 2011;44(6):927-35.
9. Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform 2011;44(2):239-50.
10. Johnson SB, Bakken S, Dine D, Hyun S, Mendonca E, Morrison F, Bright T, Van Vleck T, Wrenn J, Stetson P. An electronic health record based on structured narrative. J Am Inform Assoc 2008;15(1):54-64.
11. Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. AMIA Annu Symp Proc 2011; 2011:843-52.
12. Hao T, Rusanov A, Weng C. Extracting and normalizing temporal expressions in clinical data requests from researchers. International Health Informatics Conference, Beijing, China 3-4 August 2013;41-51.
13. SharpNLP. September 2014; Available from: http://sharpnlp.codeplex.com/.
14. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001;84(Pt 1):216-20.
15. Fan JW, Friedman C. Semantic reclassification of the UMLS concepts. Bioinformatics 2008;24(17):1971-3.