Automatic Translation of Clinical Trial Eligibility Criteria into Formal Queries

Chao XU [a,1], Walter FORKEL [b], Stefan BORGWARDT [b], Franz BAADER [b], and Beihai ZHOU [a]

[a] *Department of Philosophy, Peking University*
[b] *Institute for Theoretical Computer Science, TU Dresden*

Abstract. Selecting patients for clinical trials is very labor-intensive. Our goal is to develop an automated system that can support doctors in this task. This paper describes a major step towards such a system: the automatic translation of clinical trial eligibility criteria from natural language into formal, logic-based queries. First, we develop a semantic annotation process that can capture many types of clinical trial criteria. Then, we map the annotated criteria to the formal query language. We have built a prototype system based on state-of-the-art NLP tools such as Word2Vec, Stanford NLP tools, and the MetaMap Tagger, and have evaluated the quality of the produced queries on a number of criteria from clinicaltrials.gov. Finally, we discuss some criteria that were hard to translate, and give suggestions for how to formulate eligibility criteria to make them easier to translate automatically.

Keywords. automatic translation, natural language translation, eligibility criteria, clinical trials, patient cohort recruitment, query answering

1. Introduction

Automating the screening process for clinical trials is a major research topic [1,2,3]. As the demand for (semi-)automated patient recruitment based on electronic health records (EHRs) becomes more and more urgent, the representation and formalization of eligibility criteria (ECs) of clinical trials also have attracted considerable attention. To the best of our knowledge, however, there are no methods which can translate arbitrary ECs into logical expressions automatically (see Section 2 for related work).

Baader et al. [4] have proposed a framework for (semi-)automatically selecting patients for clinical trials, based on ontology-based query answering techniques from the area of Description Logic. Our goal is to build a prototype system that can be evaluated in practice. The users of such a system would be medical researchers rather than logicians, hence the tool must be able to formalize eligibility criteria (ECs) of clinical trials automatically. Since the available information is limited to EHRs, not all ECs can be evaluated by such a system, but it can support doctors in pre-selecting patients for a later, more thorough screening procedure.

[1] Corresponding Author: Chao Xu, Department of Philosophy, Peking University, Peking, China; E-mail: c.xu@pku.edu.cn

We present a prototype implementation that can automatically translate ECs into formal queries based on description logics. This can be seen as an instance of the larger field of translating natural language (NL) into a formal language with a precisely defined semantics. Rather than dealing with arbitrary NL expressions, we concentrate here on the restricted setting of ECs of clinical trials. These descriptions are specific to the medical domain, and there are many formal medical ontologies that can help us to recognize medical concepts. Additionally, by choosing a specific formal target language, we restrict the problem to recognizing the supported syntactical structures in NL.

Our formal query language, *metric temporal conjunctive queries with negation (MTNCQs)*, is based on several recent research results [5,6,7]. Our translation is based on annotating ECs formulated in NL by certain semantic roles and additional information. The semantic annotations we use focus on the kind of information that can be represented by our target query language, and hence can be seen as a filtering mechanism before the final translation to MTNCQs. Our prototype system uses existing NL techniques such as Word2Vec, Stanford NLP tools,[2] and MetaMap[3] [8,9,10]. We evaluate our implementation on a random selection of criteria from clinicaltrials.gov,[4] which contains more than 3.000.000 criteria from over 250.000 clinical studies. We identify which kinds of criteria are easy or hard to translate. From this, we develop some suggestions on how to formulate ECs so that processing them automatically becomes easier and more accurate.

Our prototype implementation with instructions on how to reproduce our results can be found at `https://github.com/wko/criteria-translation`. An extended version of this paper can be found at `https://tu-dresden.de/inf/lat/papers`.

2. Related Work

Our work combines two strands of research, namely representation and formalization of ECs and automatic translation from NL to formal languages.

Weng et al. [1] surveyed various representation methods of ECs and proposed a framework of five dimensions to compare them. According to different application scenarios, different representation methods for ECs are adopted. Bache et al. [2] proposed a general language for clinical trial investigation and construction (ECLECTIC) by analysing 123 criteria from 8 clinical trials. Based on our own investigation of ECs, we propose MTNCQs as formal representation language since it covers a wide range of criteria, profits from existing medical ontologies and is based on a large body of research on (temporal) ontology-based query answering [5,6,7].

Previous work has already considered translation of ECs. Tu et al. [3] proposed a practical translation method based on the ERGO annotation, which is an intermediate representation for ECs. However, ERGO annotation can only be done manually or semi-automatically. Milian et al. [11,12] focused on breast-cancer trials and summarized 165 patterns, and used these patterns and concept recognition tools to structure criteria. After that, they generated a formal representation by projecting the concepts in criteria to the predefined query template. There is also some work about extraction and representation of *partial* knowledge in ECs. Zhou et al. [13], Luo et al. [14] and Boland et al. [15]

[2] `https://nlp.stanford.edu/`

[3] `https://metamap.nlm.nih.gov/`

[4] `https://clinicaltrials.gov`

focused on the recognition and representation of temporal knowledge. Huang et al. [16] and Enger et al. [17] proposed several methods for detecting negated expressions.

Weng et al. [18], Luo et al. [14], Bhattacharya et al. [19], and Chondrogiannis et al. [20] classified the clinical trials into limited semantic classes by using semantic pattern recognition or machine learning methods, which is helpful for figuring out the most prominent kinds of information expressed in clinical trials.

In the field of NL processing, automatic translation from NL into formal language, e.g., first-order logic formulas, is also known as *automatic semantic parsing*. Dong et al. [21] proposed an automatic semantic parsing method based on machine learning, different from traditional rule-based or template-based methods.

3. Preliminaries

Our approach is based on the paradigm of *ontology-mediated query answering* [22] in description logics, where an *ontology* (formalizing medical background knowledge) is used to answer a *query* (expressing clinical trial criteria) over a *dataset* (containing patient data from EHRs). We now describe the formal languages used for these ingredients.

3.1. Medical Information

We employ the existing large medical ontology SNOMED CT, which is expressed in the tractable description logic $\mathcal{EL}$. It consists of a large number of *concept definitions* of the form $A \equiv C$ or $A \sqsubseteq C$, where A is a *concept name*, e.g., the name of a disease or a surgical procedure, and its definition C is a *complex concept*, which can be built from concept names using the constructors $C_1 \sqcap C_2$ (conjunction of concepts) and $\exists r.C$ (existential restriction over a *role name* r). The semantics of complex concepts and concept definitions can be given by a translation into first-order logic (for details, see [23]). For example, SNOMED CT contains the definition

$$\mathsf{Asthma} \sqsubseteq \mathsf{DisorderOfRespiratorySystem} \sqcap \exists \mathsf{findingSite}.\mathsf{AirwayStructure},$$

saying that asthma is a disorder of the respiratory system that occurs in airway structures.

To use the ontological knowledge, the patient data need to be formulated in terms of the concept and role names occurring in SNOMED CT. We will in the following assume that all patient data are given in the form of an *ABox*,[5] which contains *concept assertions* $A(a)$, where A is a concept name and a is an *individual name*, denoting a specific patient or the disease of a patient, and *role assertions* $r(a,b)$, where r is a role name and a,b are individual names. For example, we can represent the simple fact that a patient (represented by some identifier p) has Asthma by the assertions $\mathsf{diagnosedWith}(p,d)$, $\mathsf{Asthma}(d)$. Note that diagnosedWith is not a role name from SNOMED CT, because this ontology was not intended to explicitly model patients; we introduce this new role name here to associate patients with their diagnoses. Similarly, we introduce takes to describe patients' medication, and undergoes to describe medical procedures like surgeries that were performed on the patients. Some formats for EHRs already contain diagnoses and procedures in a structured way, e.g., in the form of SNOMED CT concepts or other

[5] For now, we abstract from the also non-trivial task of translating patient data into this form (see, e.g., [24]).

formats that can be translated to a SNOMED CT representation. Apart from that, large parts of patient records are still made up of textual reports. To recognize SNOMED CT concepts in texts, one can use existing solutions such as the MetaMap tagger [10].

3.2. A Formal Language for Eligibility Criteria

The ECs of clinical trials are separated into inclusion criteria, which must be satisfied by an eligible patient, and exclusion criteria, which must not be satisfied by the patient. We focus here on translating single criteria such as 'History of lung disease other than asthma'[6] and do not distinguish between inclusion and exclusion criteria. After translating a criterion, one can negate the output in case it was an exclusion criterion.

Our goal is to translate ECs into logical queries that can then be evaluated over the ontology (SNOMED CT) and the data (EHRs). Our precise query language, proposed in [4], is based on *conjunctive queries*, but incorporates *negation* [7], *(metric) temporal operators* in the spirit of [6,5] as well as *concrete domains* [25].

A *conjunctive query with negation (NCQ)* is a first-order formula $\phi(\mathbf{x}) = \exists \mathbf{y}.\psi(\mathbf{x},\mathbf{y})$, where ϕ is a conjunction of (negated) concept atoms $(\neg)A(x)$ and (negated) role atoms $(\neg)r(x,y)$ over the variables $\mathbf{x} \cup \mathbf{y}$. The variables $\mathbf{x}$ are called the *answer variables*. For example, we can use $\phi(x) = \exists y.\mathsf{diagnosedWith}(x,y) \land \mathsf{DisorderOfLung}(y) \land \neg\mathsf{Asthma}(y)$ to find all patients (x) that have any lung disease (y) except asthma.

A *metric temporal conjunctive query with negation (MTNCQ)* is a formula in which NCQs can be combined via the constructors $\neg\phi$, $\phi_1 \land \phi_2$, $\phi_1 \lor \phi_2$, $\Diamond_I \phi$, $\Box_I \phi$, and $\phi_1 \mathsf{U}_I \phi_2$, where I is an interval over the integers. In this setting, we assume that each assertion in our ABox also contains a *time stamp* $i \in \mathbb{Z}$, which represents the time at which this fact was recorded. For example, $\mathsf{diagnosedWith}(p,d,i)$ says that the diagnosis took place at time i. In our system, we assume that each time stamp represents a single month.

The temporal formulas $\Diamond_I \phi$, $\Box_I \phi$, and $\phi \mathsf{U}_I \psi$ express that ϕ holds at *some* point during the time interval I, at *all* points in I, and at all points *until* ψ holds (which happens within I), respectively. For example, $\Box_{[-6,0]} \exists y.\mathsf{Patient}(x) \land \mathsf{diagnosedWith}(x,y) \land \mathsf{Diabetes}(y)$ asks for patients x that have had diabetes for at least the past six months.

Finally, *concrete domains* allow MTNCQs to refer to measurements. For this, we include in the patient data assertions like $\mathsf{hemoglobinOf}(p, 15\ g/dl, i)$ to record a specific value of hemoglobin measured for patient p at time i. In the query, we extend NCQs by atoms such as $\mathsf{hemoglobinOf}(x) < 14\ g/dl$, e.g., to describe patients with abnormal measurements. We have developed an appropriate semantics and algorithms to efficiently answer MTNCQs,[7] and will extend this to deal with concrete domain atoms.

4. Methodology

The main idea is to use semantic annotations to bridge the gap between eligibility criteria and formal queries. The working of our system can be broadly divided into two stages: annotating the eligibility criterion, and then constructing a formal query from the semantic annotations. The outline of the system is shown in Figure 1.

[6] `https://clinicaltrials.gov/ct2/show/NCT02548598`

[7] A paper on this is submitted to a conference.

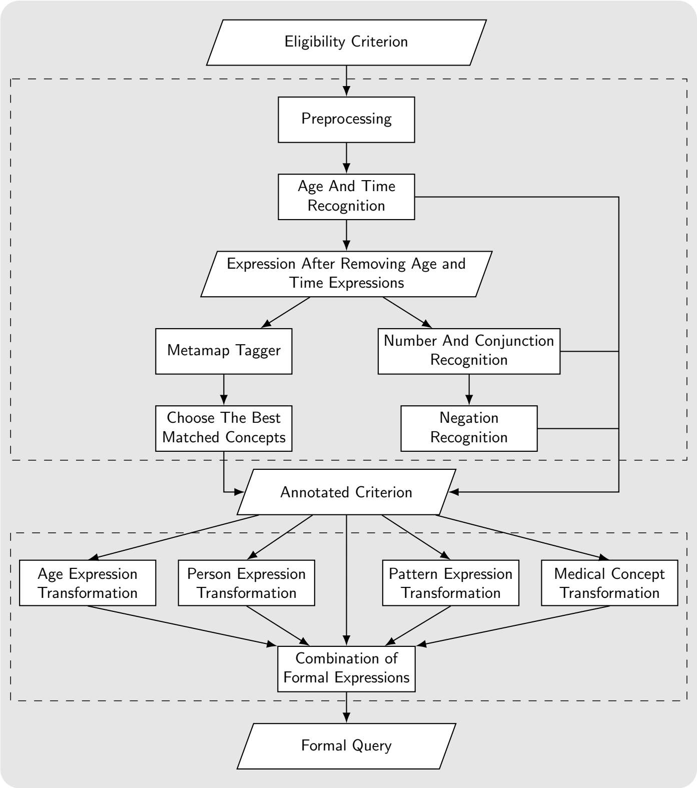

Eligibility Criterion
Preprocessing
Age And Time Recognition
Expression After Removing Age and Time Expressions
Metamap Tagger
Number And Conjunction Recognition
Choose The Best Matched Concepts
Negation Recognition
Annotated Criterion
Age Expression Transformation
Person Expression Transformation
Pattern Expression Transformation
Medical Concept Transformation
Combination of Formal Expressions
Formal Query

Figure 1. Outline of the translation system

4.1. Semantic Annotations

Our annotations identify pieces of information that can be translated to MTNCQ constructors, such as temporal operators, negation, and medical concepts. The design of the annotations also incorporates knowledge about frequently occurring types of ECs, and takes into account whether it can be reasonably expected that the queried information can be found in EHRs. We use the MetaMap tagger to recognize medical concepts, and we use keyword matching to recognize other concepts. As a preprocessing step, we homogenize the NL criteria, e.g., replace 'two' with '2' and replace 'greater than' with '>'.

4.1.1. The Selection of Semantic Roles

After looking at a number of ECs, we identified the following frequently requested types of information: *age, gender, diagnoses, medications, procedures, measurements*, and *temporal context* (e.g., 'history of . . . '). This analysis is consistent with the results of Weng et al. [18], Luo et al. [14], Bhattacharya et al. [19], and Chondrogiannis et al. [20], which all rank this kind of information high in their lists of prominent semantic classes.

Our formalization is based on SNOMED CT, which contains 19 top-level and more than 350 second-level categories. Out of these, we identified 8 categories that correspond to the above-listed information: *clinical finding, observable entity, product, substance, procedure, unit, family medical history, person.* For now, we discard other seman-

Table 1. List of semantic roles and representations in the semantic annotation

Semantic role	Examples	Representation
Age	age 18–70	$[\text{lower}, \text{upper}]$
Time	within 5 years	$[\text{start}, \text{end}]$
Comparison sign	greater than	$> \mid \geq \mid \leq \mid <$
Partial negation	other than	$\wedge\neg$
Main negation	no history of	$\neg$
Number	one, two, three, ...	Arabic numerals
Conjunction	and, or, defined by	$\wedge, \vee$
From SNOMED CT (e.g., clinical finding)	lung disease	Concept name

tic classes from SNOMED CT, such as *qualifier values* ('severe', 'known', 'isolated') or *devices*. This restriction helps to resolve some of the ambiguity of words or phrases. For example, in SNOMED CT 'female' can be mapped to '*Female structure (body structure)*' or '*Female (finding)*'; and 'scar' can be identified as '*Scar (disorder)*' or '*Scar (morphologic abnormality)*'. By excluding the types *body structure* and *morphologic abnormality*, we obtain a more uniform representation.

However, SNOMED CT only contains medical concepts, and we additionally consider the semantic roles *age, time, number, comparison sign, negation*, and *conjunction*. Table 1 contains an overview of all semantic roles with examples. In addition to the semantic role, we record additional information in the annotations, e.g., the precise concept from SNOMED CT or a time interval.

Our choice of semantic roles determines the *vocabulary* that we will use to formulate MTNCQs. More precisely, the concept names are restricted to the subconcepts of the 8 categories in SNOMED CT identified above. We use the role names diagnosedWith, takes, and undergoes to connect patients to SNOMED CT concepts, but none of the role names from SNOMED CT itself. Additionally, we allow concrete domain predicates like hemoglobinOf that correspond to SNOMED CT *substances* (e.g., Hemoglobin) and *observable entities*, as well as ageOf. Finally, temporal information, negation, and conjunction are expressed by the logical connectives of our query language.

4.1.2. Concept Recognition and Semantic Role Annotation

To illustrate the annotation process, we consider the EC 'history of lung disease other than asthma';[8] Table 2 and the end result in Figure 2.

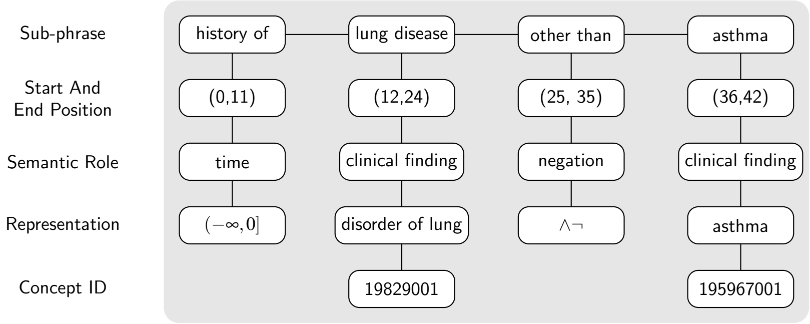
The first steps are to recognize and annotate age and temporal expressions using regular expressions. In our example, 'history of' is recognized by the regular expression '(a|any|prior|previous) (.*?)history of', and then annotated by the semantic role *time* and the temporal interval $(-\infty, 0]$. We then remove the identified age expressions and temporal expressions from the EC. They form complete semantic units, and thus removing them does not affect the meaning of the remaining part of the EC, while it allows us to avoid accidental translation of these expressions into SNOMED CT concepts.

On the remaining criterion, we then run the MetaMap tagger [10], a tool for recognizing concepts from the UMLS Metathesaurus, which subsumes SNOMED CT. Given

[8] https://clinicaltrials.gov/ct2/show/NCT02548598

Table 2. Example of the semantic annotation of an EC.

Stage	Output
Original EC	history of lung disease other than asthma
Age recognition	—
Time recognition	history of → *(time)*
Remove age/time	lung disease other than asthma
MetaMap	lung disease other than asthma → *Disorder of lung (disorder), Lung structure (body structure), Asthma (disorder)*
Restrict semantic roles	lung disease other than asthma → *Disorder of lung, Asthma*
Detect sub-phrases	lung disease, lung, disease, other, than, asthma
Compute most similar concept for each sub-phrase	(lung disease, *Disorder of lung*) : 0.91, (lung, *Disorder of lung*) : 0.81, (disease, *Disorder of lung*) : 0.89, (asthma, *Asthma*) : 1.0
Find best matches	lung disease → *Disorder of lung*, asthma → *Asthma*
Negation recognition	other than → *(negation)*
Other semantic roles	—



Sub-phrase	history of	lung disease	other than	asthma
Start And End Position	(0,11)	(12,24)	(25, 35)	(36,42)
Semantic Role	time	clinical finding	negation	clinical finding
Representation	$(-\infty, 0]$	disorder of lung	$\wedge\neg$	asthma
Concept ID		19829001		195967001

Figure 2. The semantic annotation for our example.

a phrase or sentence, it returns the most likely phrase-concept matches. In our example, MetaMap does not identify any sub-phrases, and outputs the following concepts for the whole phrase 'lung disease other than asthma': *'Disorder of lung (disorder)', 'Lung structure (body structure)', 'Asthma (disorder)'*. By restricting the types as described in Section 4.1.1, we immediately rule out *'Lung structure'*.

A larger challenge, however, is to obtain more exact phrase-concept matches. For this, we split all sub-phrases returned by MetaMap into more sub-phrases using the Stanford NLP tools [9]. Then we try to find the best phrase-concept matches, by calculating a similarity value (in $[0,1]$) of each sub-phrase to all candidate concepts using Word2Vec [8] and the Levenshtein distance; we also use the synonymous expressions provided by SNOMED CT to potentially obtain a higher similarity. To avoid spurious matches, we use a minimum threshold of 0.66 for the similarity. In our example, this excludes the words 'other' and 'than', because there is no candidate concept that is similar enough. The best matches for the phrases 'lung disease', 'lung', and 'disease' all refer to the same concept *Disorder of lung*, and we use the similarity values to choose the best of them, where we give preference to longer phrases.

Table 3. Translation of basic query parts.

Semantic role	Formalization	Example
Age	$\mathsf{ageOf}(x) \geq \text{lower} \wedge \mathsf{ageOf}(x) \leq \text{upper}$	$\mathsf{ageOf}(x) \geq 18$
Time	$\Diamond_{[\text{start,end}]}$	$\Diamond_{[-12,0]}$
Person	$\mathsf{conceptName}(x)$	$\mathsf{Woman}(x)$
Clinical finding	$\exists y.\mathsf{diagnosedWith}(x,y) \wedge \mathsf{conceptName}(y)$	$\exists y.\mathsf{diagnosedWith}(x,y) \wedge \mathsf{HIV}(y)$
Product	$\exists y.\mathsf{takes}(x,y) \wedge \mathsf{conceptName}(y)$	$\exists y.\mathsf{takes}(x,y) \wedge \mathsf{Aspirin}(y)$
Procedure	$\exists y.\mathsf{undergoes}(x,y) \wedge \mathsf{conceptName}(y)$	$\exists y.\mathsf{undergoes}(x,y) \wedge \mathsf{Appendectomy}(y)$
Measurement pattern: substance/observable entity—comparison sign—number—unit		
Formula: $\mathsf{conceptNameOf}(x)$ $(> \mid \geq \mid \leq \mid <)$ number $\mathsf{conceptName}$		$\mathsf{hemoglobinOf}(x) < 14g/dl$
Group pattern: clinical finding—clinical finding—...		
Formula: $\mathsf{conceptName1}(y) \vee \mathsf{conceptName2}(y) \vee \dots$		$\mathsf{HIV}(y) \vee \mathsf{HepatitisC}(y)$
Negation pattern: clinical finding—partial negation—clinical finding		
Formula: $\mathsf{conceptName1}(y) \wedge \neg\mathsf{conceptName2}(y)$		$\mathsf{Diabetes}(y) \wedge \neg\mathsf{DiabetesType1}(y)$

It remains to recognize other semantic roles in the EC, i.e., *number, negation, comparison sign*, and *conjunction*. We mainly do this by keyword or pattern matching. The negation case is the most complex due to its various forms:

- explicit negation e.g., ‘not’, ‘except’, ‘other than’, ‘with the exception of’;
- morphological negation, e.g., ‘non-pregnant’, ‘non-healed’, ‘non-smoker’;
- implicit negation, e.g., ‘lack of’, ‘rule out’, ‘free from’.

In our prototype system, we focus on explicit negation, and consider two cases: either the whole sentence is negated (‘patient does not have ... ’) or only part of it (‘... other than ... ’). For conjunctions between parts of sentences, we use ‘$\vee$’ as default annotation, because there is no good way to map ‘and’ and ‘or’ in EC to conjunction or disjunction exactly, e.g., in the EC ‘... including cyclosporine, systemic itraconazole *or* ketoconazole, erythromycin *or* clarithromycin, nefazodone, verapamil *and* human immunodeficiency virus protease inhibitors’[9] both ‘and’ and ‘or’ have the same meaning.

The final semantic annotation for our example can be seen in Figure 2.

4.2. The Formal Queries

To obtain the final MTNCQ, we combine the different annotated phrases according to the composibility of semantic roles and structural information. There are four kinds of basic subformulas: age formulas, person formulas, medical formulas and pattern formulas, and their translation is described in Table 3. For measurements, we detect patterns in the semantic annotation that correspond to a comparison of a *substance* or *observable entity* with a specific numerical value (including unit). Additionally, we group adjacent SNOMED CT *findings* together, to translate them into a set of atoms joined by $\vee$ inside the same $\exists y.\mathsf{diagnosedWith}(x,y) \wedge \dots$ formula. We also translate negation between *clinical findings* into appropriate formulas, and do the same for *products* and

[9] `https://clinicaltrials.gov/ct2/show/NCT02452502`

Table 4. Experimental results. The right table shows the annotation of the translation quality for the 93 criteria that were marked as 'answerable' by all evaluators.

	Unanswerable	**Answerable**
evaluator 1	282	119
evaluator 2	254	147
evaluator 3	237	164

	Good	**Partial**	**Wrong**
evaluator 1	54	29	10
evaluator 2	56	27	10
evaluator 3	65	18	10

procedures. In our running example, 'lung disease other than asthma' is formalized as $(\exists y.\mathsf{diagnosedWith}(x,y) \land \mathsf{DisorderOfLung}(y) \land \neg\mathsf{Asthma}(y))$.

Finally, we combine these subformulas using the remaining connectives and negations and consider any time expressions. In our prototype system, we only express a single temporal operator of the form $\Diamond_{[-n,0]}$, which we found to be the most common in clinical trials. Such an operator is always applied to the whole formula, e.g., we obtain $\Diamond_{(-\infty,0]}(\exists y.\mathsf{diagnosedWith}(x,y) \land \mathsf{DisorderOfLung}(y) \land \neg\mathsf{Asthma}(y))$. If there is more than one temporal annotation, we choose the more specific one. For example, in 'history of myocardial infarction, unstable angina pectoris, percutaneous coronary intervention, congestive heart failure, hypertensive encephalopathy, stroke or TIA within the last 6 months'[10] there are 'history of' and 'within the last 6 months', and we choose the latter.

If there are no explicit connectives, we combine medical and measurement formulas by disjunction, and then combine them with age and person formulas by conjunction.

5. Experiments

To the best of our knowledge, there are no gold standard datasets for the translation of criteria into formal language. Therefore, we evaluated our approach on real-world studies taken from clinicaltrials.gov.[11] During the design phase we used 24 randomly selected studies, which contained approximately 300 criteria. Our prototype system was optimized to cover as many of these criteria as possible.

For testing, we randomly selected criteria across all studies on clinicaltrials.gov and manually evaluated them. Due to time constraints, we managed to process 401 criteria. We defined the following metrics: A criterion is *answerable*, if a) it is possible for a human to translate it into an MTNCQ using only the vocabulary chosen in Section 4.1.1; and b) it can in principle be answered by only looking at the EHR of a patient. Hence, criteria that refer to the future ('during study phase'), or ask for subjective information ('in the opinion of the investigator', 'willingness to'), are not considered answerable for the purposes of our system. For each answerable criterion, we then evaluated the quality of the translation. The resulting MTNCQ is labeled as *good* if it contains all (necessary) information; *partial* if it represents at least parts of the criterion; and *wrong* otherwise. These metrics are clearly subjective to some extent. To get a more reliable evaluation and to quantify the amount of subjectivity, we let three evaluators (three of the authors) vote independently on the test data. The results can be seen in Table 4.

The results indicate that the judgment on whether a criterion is answerable or not differs between the evaluators. We found that the difference is mainly caused by two

[10] `https://clinicaltrials.gov/ct2/show/NCT00220220`

[11] `https://clinicaltrials.gov/`

things: Firstly, it is sometimes difficult to judge whether a concept can be represented in SNOMED CT, because the concept name can differ significantly from the description in the text. Secondly, many criteria contain very specific phrases, for example ‘Active bowels inflammatory disease ([Crohn], chronic, diarrhea...)’.[12] The word ‘active’ cannot be translated into SNOMED CT, and we could translate it into a temporal constraint only under some assumptions on the semantics of ‘active’. Some might consider this to be not so important, while for others this renders the criterion unanswerable. Despite the differences, at least 60% of the criteria cannot be answered, even in the opinion of evaluator 3, who was the most optimistic. This is partially because of condition b) above. The second reason is that quite a number of criteria cannot be represented in our formal language, either because of a lack of vocabulary in SNOMED CT, or because of missing semantic roles (see Section 4.1.1). While the former cannot be improved on, the latter offers room for future optimizations.

To compare the quality of the translations, we consider only criteria that have been marked as *answerable* by all evaluators. This leaves 93 criteria that are analyzed on the right-hand side of Table 4. The difference in the translation quality is again due to the varying opinions of the evaluators regarding how detailed a translation needs to be in order to be considered good. Our system is able to translate more than 50% of the (confidently) answerable criteria, which is a promising first result. In the following, we give examples for a good, partial, and a bad translation of our system:

‘Has a history of diabetic ketoacidosis in the last 6 months.’[13]

$$\Diamond_{[-6,0]}\big(\exists y.\mathsf{diagnosedWith}(x,y)\land\mathsf{KetoacidosisInDiabetesMellitus}(y)\big)$$

‘History of, diagnosed or suspected genital or other malignancy (excluding treated squamous cell carcinoma of the skin), and untreated cervical dysplasia.’[14]

$$\Diamond_{(-\infty,0]}\Big(\exists y.\mathsf{diagnosedWith}(x,y)\land\big(\mathsf{MalignantNeoplasticDisease}(y)\lor\mathsf{DysplasiaOfCervix}(y)\big)\Big)$$

‘Primary tumors developed 5 years previous to the inclusion, except in situ cervix carcinoma or skin basocellular cancer properly treated’[15]

$$\Diamond_{(-\infty,0]}\Big(\exists y.\mathsf{diagnosedWith}(x,y)\land\big(\mathsf{CarcinomaInSituOfUterineCervix}(y)\lor\mathsf{SkinCancer}(y)\big)\Big)$$

The second translation is partially correct, because the temporal data and the main concepts have been recognized correctly, but ‘excluding . . . ’ was not translated. The last translation is wrong since neither the temporal information, the negation, nor the main concept ‘primary tumors’ have been recognized correctly. For more examples, we refer the reader to the appendix in the extended version.

6. Discussion and Ongoing Work

Formalizing ECs is a challenging task due to the gap between natural and formal language. In this paper, we have presented an automatic translation method from ECs into

[12] https://clinicaltrials.gov/ct2/show/NCT02363725
[13] https://clinicaltrials.gov/ct2/show/NCT02269735
[14] https://clinicaltrials.gov/ct2/show/NCT01397097
[15] https://clinicaltrials.gov/ct2/show/NCT01303029

formal queries, and developed a prototype system based on existing NLP tools. We have evaluated our prototype on 401 eligibility criteria. More than 50% of the answerable criteria have been translated correctly, which is an encouraging result that can be improved on by optimizing the translation process as we describe below. However, there remain certain criteria that are hard to translate (even for humans) due to their complex structure.

While it is unreasonable to expect medical doctors to formulate clinical trial criteria directly as MTNCQs, we nevertheless identify a few key points that can be observed during the formulation of ECs to make the automatic translation easier:

1. Split criteria whenever possible, e.g., divide ‘diagnosed with diabetes and hypertension’ into ‘diagnosed with diabetes’ and ‘diagnosed with hypertension.’
2. Formulate every EC as an independent description that does not depend on other criteria or the background knowledge of clinical trials, like in ‘Known hypersensitivity to any of the study drugs or excipients.’[16]
3. Avoid using nonadjacent words to express a concept, e.g., ‘… dermatologic, neurologic, or psychiatric disease’[17] should rather be formulated as ‘dermatologic disease, neurologic disease, or psychiatric disease.’

We can improve the quality of our translation by collecting more regular expressions and custom mappings, or employing specialized techniques from the literature for the recognition of semantic roles like *comparison sign* or *negation*. Other obvious steps are the inclusion of more concept categories from SNOMED CT such as *devices, qualifiers*, and *events*. For example, the criterion ‘severe aortic stenosis’[18] could be translated as $\exists y,z.\mathsf{hasDiagnosis}(x,y) \wedge \mathsf{AorticStenosis}(y) \wedge \mathsf{severity}(y,z) \wedge \mathsf{Severe}(z)$ if we annotate ‘severe’ with the SNOMED CT concept *severe (qualifier value)* and detect the pattern *qualifier value—finding*. It is also straightforward modify our system to output a ranked list of multiple candidate translations that the doctor may choose from.

Another interesting direction for future work is to develop a *controlled natural language* [26] based on our semantic annotations. Criteria formulated in this way can then easily be transformed into MTNCQs as we have described. With appropriate editing support, creating new ECs that conform with this controlled NL would be not much more difficult than writing them as free-form text. Of course, one should retain the possibility to add free-form criteria, which then have to be evaluated manually.

Acknowledgements

This work was supported by the DFG project BA 1122/19-1 (GOASQ), by DFG grant 389792660 as part of TRR 248, and by the China Scholarship Council.

References

[1] Chunhua Weng, Samson W Tu, Ida Sim, and Rachel Richesson. Formal representation of eligibility criteria: a literature review. *J. Biomed. Inform.*, 43(3):451–467, 2010.

[16] `https://clinicaltrials.gov/ct2/show/NCT01935492`

[17] `https://clinicaltrials.gov/ct2/show/NCT00960570`

[18] `https://clinicaltrials.gov/ct2/show/NCT01951950`

[2] Richard Bache, Adel Taweel, Simon Miles, and Brendan C Delaney. An eligibility criteria query language for heterogeneous data warehouses. *Method. Inform. Med*, 54(01):41–44, 2015.

[3] Samson W Tu, Mor Peleg, Simona Carini, Michael Bobak, Jessica Ross, Daniel Rubin, and Ida Sim. A practical method for transforming free-text eligibility criteria into computable criteria. *J. Biomed. Inform.*, 44(2):239–250, 2011.

[4] Franz Baader, Stefan Borgwardt, and Walter Forkel. Patient selection for clinical trials using temporalized ontology-mediated query answering. In *Proc. HQA Workshop*, pages 1069–1074. ACM, 2018.

[5] Franz Baader, Stefan Borgwardt, and Marcel Lippmann. Temporal query entailment in the description logic $\mathcal{SHQ}$. *J. Web Sem.*, 33:71–93, 2015.

[6] Franz Baader, Stefan Borgwardt, Patrick Koopmann, Ana Ozaki, and Veronika Thost. Metric temporal description logics with interval-rigid names. In *Proc. FroCoS Symposium*, pages 60–76. Springer, 2017.

[7] Stefan Borgwardt and Walter Forkel. Closed-world semantics for conjunctive queries with negation over $\mathcal{ELH}_\bot$ ontologies. In *Proc. JELIA Conference*, pages 371–386. Springer, 2019.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. ALC Meeting*, pages 55–60, 2014.

[10] Alan R Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proc. AMIA Symposium*, page 17, 2001.

[11] Krystyna Milian, Anca Bucur, and Annette Ten Teije. Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach. In *Proc. BIBM Conference*, pages 1–4. IEEE, 2012.

[12] Krystyna Milian and Annette ten Teije. Towards automatic patient eligibility assessment: From free-text criteria to queries. In *Proc. AIME Conference*, pages 78–83. Springer, 2013.

[13] Li Zhou, Genevieve B Melton, Simon Parsons, and George Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *J. Biomed. Inform.*, 39(4):424–439, 2006.

[14] Zhihui Luo, Meliha Yetisgen-Yildiz, and Chunhua Weng. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *J. Biomed. Inform.*, 44(6):927–935, 2011.

[15] Mary Regina Boland, Samson W Tu, Simona Carini, Ida Sim, and Chunhua Weng. EliXR-TIME: A temporal knowledge representation for clinical research eligibility criteria. *AMIA Transl. Sci. Proc.*, 2012:71, 2012.

[16] Yang Huang and Henry J Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *J. Am. Med. Inform. Assn.*, 14(3):304–311, 2007.

[17] Martine Enger, Erik Velldal, and Lilja Øvrelid. An open-source tool for negation detection: A maximum-margin approach. In *Proc. SemBEaR Workshop*, pages 64–69, 2017.

[18] Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. EliXR: An approach to eligibility criteria extraction and representation. *J. Am. Med. Inform. Assn.*, 18(Supplement_1):i116–i124, 2011.

[19] Sanmitra Bhattacharya and Michael N Cantor. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. *J. Biomed. Inform.*, 46(5):805–813, 2013.

[20] Efthymios Chondrogiannis, Vassiliki Andronikou, Anastasios Tagaris, Efstathios Karanastasis, Theodora Varvarigou, and Masatsugu Tsuji. A novel semantic representation for eligibility criteria in clinical trials. *J. Biomed. Inform.*, 69:10–23, 2017.

[21] Li Dong and Mirella Lapata. Language to logical form with neural attention. In *Proc. Annual Meeting of the ACL*, 2016.

[22] Meghyn Bienvenu. Ontology-mediated query answering: Harnessing knowledge to get more from data. In *Proc. IJCAI Conference*, pages 4058–4061. AAAI Press, 2016.

[23] Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.

[24] Jon Patrick, Yefeng Wang, and Peter Budd. An automated system for conversion of clinical notes into SNOMED Clinical Terminology. In *Proc. ACSW Symposium*, pages 219–226, 2007.

[25] Franz Baader and Philipp Hanschke. A scheme for integrating concrete domains into concept languages. In John Mylopoulos and Raymond Reiter, editors, *Proc. IJCAI Conference*, pages 452–457, 1991.

[26] Tobias Kuhn. A survey and classification of controlled natural languages. *Comput. Linguist.*, 40(1):121–170, 2014.

A. Evaluation Examples

In Table 5, we list additional examples of criteria that were considered *unanswerable* in our evaluation due to various reasons. Table 6 lists some *answerable* criteria that were not translated correctly; all of them could be translated if our system was extended by appropriate regular expressions.

Table 5. Examples of unanswerable criteria.

Type	Example
Future information	*Planned* primary unilateral THA or TKA `https://clinicaltrials.gov/ct2/show/NCT02405104`
Time point close to the study	Recent tooth extraction or major dental procedure *within 3 weeks of study entry* `https://clinicaltrials.gov/ct2/show/NCT00102908`
Subjective attitude	female patients of childbearing potential *unwilling to* use a medically acceptable form of contraception `https://clinicaltrials.gov/ct2/show/NCT00891683`
Incomplete/unclear description	Known hypersensitivity to any of the *study drugs or excipients* `https://clinicaltrials.gov/ct2/show/NCT02923739`
	Having a history of *diseases stimulated by heat* `https://clinicaltrials.gov/ct2/show/NCT01362192`
	Cardiac arrhythmia *requiring* medical therapy `https://clinicaltrials.gov/ct2/show/NCT00343525`
Concepts not in SNOMED CT	Any other *hormone treatment contraindications* `https://clinicaltrials.gov/ct2/show/NCT01057511`
Concepts outside our vocabulary	*severe* diabetes `https://clinicaltrials.gov/ct2/show/NCT00521053`
Concepts not recognized by MetaMap	*Ejection fraction* is required if the patient is $>$ 50 years of age, or history of cardiac disease or anthracycline exposure `https://clinicaltrials.gov/ct2/show/NCT00040846`
Overly detailed description	Prior Therapy - Patients are eligible if they have been treated with clofarabine, mitoxantrone, or a combination of both in the past. However, the maximal lifetime cumulative previous anthracycline dose should not exceed doxorubicin dose equivalent of 450 mg/m2 (see Table 1). Patients who received more than one anthracycline prior to study entry should have each individual agent cumulative dose converted to doxorubicin equivalent and added together (eg, a patient who received cumulative dose of Daunorubicin at 200 mg/m2 and Mitoxantrone 48 mg/m2 has a total doxorubicin dose equivalent of 358.6 mg/m2 (200 mg/m2 x 0.833 + 48 mg/m2 x 4). `https://clinicaltrials.gov/ct2/show/NCT01842672`

Table 6. Examples of answerable criteria that could not be translated.

Reason	Example
Age expressions are not recognized	patients aged *under 18 years* `https://clinicaltrials.gov/ct2/show/NCT02710877`
Time expressions are not recognized	*12 months of* spontaneous amenorrhea `https://clinicaltrials.gov/ct2/show/NCT02865538`
Negations are not recognized	Participant has a transplanted organ, *excluding* corneal transplant, performed $>$ 3 months prior to the first dose of trial medication `https://clinicaltrials.gov/ct2/show/NCT01651936`
Comparison signs are not recognized	Calcium serum values *below* 7.0 mg/dl or *above* 10.0 mg/dl `https://clinicaltrials.gov/ct2/show/NCT01815021`
Conjunctions are recognized wrongly	Disease and/or medical conditions *accompanied by* hypercalcaemia and/or hypercalciuria `https://clinicaltrials.gov/ct2/show/NCT01480869`