A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria

Samson Tu, MS[1], Mor Peleg, PhD[1,2], Simona Carini, MS[3], Michael Bobak, MS[3], Daniel Rubin, MS, MD,[1] Ida Sim, MD, PhD[3]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

[2]Department of Management Information Systems, University of Haifa, Haifa, Israel

[3]University of California, San Francisco, CA

Formalizing eligibility criteria in a computer-interpretable language to enable searching for studies or to identify candidate subjects for studies is an extremely labor intensive task. In the TrialBank project, we developed a methodology for incrementally capturing the semantics of eligibility criteria. The methodology defines an intermediate representation that is informed by both the complexity of natural language and the requirements of computable format. It allows us to apply computational methods to partially automate the formalization process and to derive values from the intermediate representation. We performed a feasibility study that demonstrates the practicality of this methodology.

INTRODUCTION

Human studies are the most important source of evidence for advancing our understanding of health, diseases, and treatment options. Finding better ways to design, perform, and make use of the studies is crucial to improve health care. Repositories of studies, such as ClinicalTrials.gov, allow queries based on health conditions studied, interventions, and other study items that describe the target population of a study with less detail than eligibility criteria. Currently, eligibility criteria are written in free text that cannot be reliably parsed or processed computationally. Making eligibility criteria computer-interpretable could enable several valuable uses. For example, at the design stage, study authors could query a library of standardized criteria to ensure that their study subjects are comparable to those of other studies. An author could determine whether the criteria that she uses are more generic than those used in other studies, making her target population more inclusive. Second, at execution stage, investigators could query an electronic health record to screen for eligible subjects. Finally, health-care providers could query a repository to find studies that match the characteristics of their patients.

Over the years, informatics researchers have developed representations of computable eligibility criteria. Some, like CDISC's ASPIRE project [1], seek to develop consensus on a core set of generic and disease-specific eligibility codes. Others, like Arden Syntax [2], GELLO [3], or other logic-based rule languages, create formal syntax for encoding computer-interpretable expressions that use external terminology systems. Nevertheless, encoding eligibility criteria using the existing methods is a difficult task. Eligibility codes have not been fully standardized, and it will take years to create eligibility codes with ASPIRE's disease-specific approach. The expression language approach is labor-intensive and requires encoders with special skills. Furthermore, these expression languages do not offer support for encoding the noun phrases, which contain much of the semantics of free-text criteria.

To address these problems, we created a methodology that helps us formalize natural language eligibility criteria in incremental steps. The methodology defines an intermediate representation, called **ERGO Annotation**, which is informed by both the complexity of natural language and the requirements of a computable format. We performed a feasibility study that demonstrates the practicality of a computer-assisted process for annotating eligibility criteria and the utility of ERGO Annotations for advancing the main use cases for eligibility criteria listed earlier. The annotation process starts from free-text eligibility criteria of clinical studies, incrementally classifies the criteria into well-defined statement types, and uses natural language processing (NLP) techniques to extract noun phrases and semantic connectors that capture the semantics of the criteria. Along the way, the noun phrases are mapped to terms in standard terminologies and to semantic types. Furthermore, we demonstrate how the resulting ERGO Annotations can be transformed into a computable and queryable language, such as Web Ontology Language (OWL), for the classification of eligibility criteria and the querying of studies by their target populations.

METHODS

We first describe the ERGO Annotation and then the design of the feasibility study.

ERGO Annotation

We have previously developed the Eligibility Rule Grammar Ontology (ERGO), a template-based expression language that can capture the full expressivity of eligibility criteria from any clinical domain [4]. Like

GELLO, ERGO is based on an object-oriented data model, and its expressions can be seen as a subset of GELLO, except that (1) ERGO allows not only Boolean combinations of statements, but also statements connected by semantic connectors such as *defined by* ("adult patients *defined by* age >= 18 years") and by examples ("Planned coronary revascularization *such as* stent placement *OR* heart bypass"), (2) ERGO explicitly incorporates constraints on temporal relationships (which are handled through object-oriented extensions in GELLO), and (3) ERGO explicitly models terminological expressions. ERGO defines three subclasses of *noun phrases*:

1. Primitive noun phrases which represent terms from vocabularies.
2. Logical combinations of noun phrases connected by *and*, *or*, or *not*. A noun phrase is interpreted as a set of terms that are the same or more specific than the named noun phrase (e.g., "acute MI" is a part of the set of terms denoted by "MI"'. The *and*, *or*, and *not* operators are interpreted as intersection, union, and complement of the corresponding sets.
3. Noun phrases with modifiers that place restrictions on the root noun phrase. Modifiers follow the entity-attribute-value model (e.g., asthma *induced by* exercise). In cases where the attribute of the modifier is unclear, we use a default *attribute* attribute that can be elided.

Following SNOMED-CT, ERGO represents a context-dependent noun phrase such as "family history of colon cancer" as a post-coordinated noun phrase with a root phrase ("family history") and one or more modifiers (e.g., "associated-finding colon cancer").

Recognizing that encoding of eligibility criteria using ERGO is a labor-intensive process and that an ERGO execution engine or mappings of ERGO expressions to other executable languages do not exist, we developed ERGO Annotation as an intermediate representation that bridges the gap between natural language eligibility criteria and ERGO criteria. This intermediate representation focuses on noun phrases and three types of statements:

1. Simple statements making a single assertion
2. Comparison statements that have the form *Noun Phrase comparator* (e.g., >, < =) *Quantity*
3. Complex statements - multiple statements joined by Boolean connectives *AND*, *OR*, *NOT*, *IMPLIES* or semantic connectors (e.g., *evidenced by*).

We define a valid ERGO Annotation for a simple statement as a noun phrase (possibly post-coordinated with modifiers) that is either the most specific noun phrase that can be extracted from a criterion, a noun phrase semantically equivalent to the most specific one, or a generalization of the most specific noun phrase supported by the criterion text. For example, the criterion "asthma induced by exercise" would be annotated with UMLS code C00004096 (asthma) that has the attribute *induced by* and attribute value C0015259 (exercise). Alternatively, C0004099 (asthma, exercise-induced*)* or C00004096 (asthma) by itself are also valid ERGO Annotations. When the noun phrase is a logical combination of other terms, then this definition applies to each part of the combined noun phrase separately. We exclude phrases that are at the level of some UMLS Semantic Types (e.g., Disease or Syndrome) or that cannot be the subject of an assertion about a person (e.g., "pressure" in "high blood pressure"). These phrases are so general as to be non-informative or meaningless when applied to a person and therefore cannot be considered valid annotations.

The ERGO Annotation for a comparison statement is the triplet {*noun phrase*, *comparison operator*, *quantity*}, where *quantity* may be a string when the quantity cannot be expressed as a value and unit.

Finally, the ERGO Annotation for a complex statement includes the ERGO Annotations for its component simple and comparison statements joined by relevant Boolean connectives and semantic connectors.

Design of the Feasibility Study

We tested the feasibility and utility of this approach by developing partially automated methods for generating a database of studies that have their criteria annotated with ERGO Annotations and testing the method with a sample set of studies (Steps A - C). We then demonstrate the possibility of using the database for some of the use cases described in the Introduction (Step D).

A. Preparing eligibility criteria included the following steps:

1. Developing a Microsoft Excel-based instrument that allows us to record, classify, and rewrite eligibility criteria.
2. Selecting trials for the feasibility study by searching ClinicalTrials.gov.
3. Pre-processing the eligibility criteria by

- Eliminating criteria or parenthesized fragments that are either too vague or that do not have any discriminating power (e.g., "Men or women," "Low HDL cholesterol ('good cholesterol').
- Breaking down criteria that, though written on the same line (or bullet point), are really separate criteria, so that each criterion becomes stand-alone.

4. Classifying criteria as Simple, Comparison, or Complex.
5. Rewriting the criteria by decomposing complex statements into their constituent simple or comparison statements, and by:

- Making implied semantics explicit using the seman-

tic relationship *defined by* [e.g., "adult patients, 18-75 years of age" is rewritten as "adult patients *defined by* (age >= 18 years *AND* age <= 75 years)]

- Eliminating partial lists [e.g., "treatment with drugs raising HDL-C (e.g., niacin, fibrates)" is rewritten as "treatment with drugs raising HDL-C *OR* treatment with niacin *OR* treatment with fibrates"]
- Using implication as logical connector where needed [e.g., "Women must be post-menopausal or using an effective method avoiding of pregnancy" is rewritten as "women *IMPLIES* (post-menopausal *OR* using an effective method for avoiding pregnancy)"]
- Eliminating Boolean negation by changing inclusion criteria to exclusion criteria or vice versa.

6. Rewriting terminological negation as *EXCLUDING*, to avoid confusion with Boolean negation [e.g., "Any life-threatening disease expected to result in death within 2 years (other than heart disease)" is rewritten as "life-threatening disease *EXCLUDING* heart disease *AND* life expectancy<=2 years"] .
7. Manually annotating the criteria with the maximally specific ERGO Annotations. For example, the maximally specific ERGO Annotation of "Dietary supplements containing phenolic compounds one month prior to study admission" is "dietary supplement" modified by "phenolic compound."

B. Semi-automated annotation process

The process of creating ERGO Annotations involves detecting linguistic noun phrases, linking them to standard terminologies, recognizing their semantic types, and formulating them as ERGO Annotations. As much as possible, the process should be assisted by automated tools. We experimented with three different NLP parsers [5-7] and used the National Center for Biomedical Informatics' Open-Biomedical Annotator (OBA) [8] and National Library of Medicine's LexAccess [9]. We used the following procedure:

1. Apply OBA to the preprocessed criteria to get the longest coded string and their UMLS CUIs and semantic types.
2. Apply LexAccess to the recognized strings to get their parts-of-speech
3. The phrases corresponding to CUIs are formed into single compound words (e.g., if "postauricular scar" is a phrase that correspond to some CUI, we will use "postauricular-scar"
4. Running preprocessed criteria through Open NLP Parser to find noun phrases
5. Use a heuristic algorithm to extract noun phrases and their modifiers for simple criteria and noun phrases, comparators, and quantities for comparison criteria, discarding everything else. The details of the algorithm are available at http://rctbank.ucsf.edu/home/ergo.html.

The output of the semi-automated annotation process are criteria with their acquired annotations. For example, OBA would recognize “heart failure” in "Severe heart failure" as an UMLS term, and the Open NLP Parser generate the parse tree [NP Severe/JJ heart-failure/NN]. Our heuristic algorithm, in this case, would use the noun (NN) as the root noun phrase and the adjective (JJ) as the modifier in the acquired ERGO Annotation for the criterion.

C. Evaluating the extent to which the annotations acquired through automated tools match the intended ERGO Annotations of the criteria.

For our initial feasibility study, we classified the ERGO noun phrase annotations acquired by the NLP tools as (1) a *match* when the acquired noun phrase is a valid ERGO annotation that is the most specific noun phrase of the manually created annotations or semantically equivalent to it, (2) a *non-match*, when the acquired noun phrase is not a valid ERGO Annotation for the statement, and (3) *partial match* if the acquired annotation is a valid ERGO Annotation but not semantically-equivalent to the most specific one. For example, for the criterion "high blood pressure," "blood pressure" is a valid ERGO Annotation that is not maximally specific, and therefore is a partial match. "Pressure" is not a sensible assertion about a person and therefore is not a valid ERGO Annotation. For a comparison criterion, a match requires that the acquired annotation includes not only the maximally specific noun phrase, but also the comparator and the quantity components.

D. Evaluating the possibility of supporting the use cases of an ERGO-annotation ontology: classifying criteria by noun phrases (to help researchers find appropriate criteria) and querying for studies (to help clinicians find studies, whose ERGO Annotations satisfy a query expression). We demonstrate the utility of ERGO Annotations by rewriting them as OWL expressions in Protégé.

Specifically, a noun phrase N with modifier M can be written as an OWL restriction (using the Manchester syntax) `(N and (modifier some M))`, where *modifier* is an OWL object property. A comparison annotation can be defined as restrictions on a noun phrase and a quantity. We treat the semantic connector *defined by* as an OWL equivalence relation. The other semantic connectors, such as *evidenced by* or *caused by*, become OWL object properties so that, for example, "coronary heart disease evidenced by angiography" becomes the OWL restriction "`coronary_heart_disease AND (evidenced_by some angiography)`."

RESULTS

We selected 4 trials from ClinicalTrials.gov for our feasibility study by using the search term "heart disease" on January 12, 2009 and by selecting the second,

sixth, eighth, and tenth open interventional studies. (Trial #2 and #4 have several criteria in common, so we excluded #4.) The 4 trials yielded 60 distinct criteria. Six were removed because they were too vague or non-discriminating. After rewriting, the criteria were decomposed into 110 simple and 12 comparison criteria, yielding 113 unique atomic criteria.

We applied our semi-automated method for extracting ERGO Annotations to the criteria and, according to our metric, had 55 (48.67%) matches, 24 (21.24%) partial matches, and 34 (30.09%) non-matches. Compared to the state of art in identifying maximal noun phrases in radiology reports, where the recall rate can be 82% or higher [11], our recall rate is significantly lower.

To demonstrate applicability of OWL-based ERGO Annotations to our use cases, we defined an OWL ontology (Figure 1) that consists of a *Study* class that may have an object property called *has_eligibility_criteria*. The *has_eligibility_criteria* property is specialized into *has_exclusion_criteria* and *has_inclusion_criteria* subproperties. The *Criterion* class has a property called *has_ERGO_annotation*. ERGO Annotation may be a *Noun_phrase_type* or a *Comparison_Annotation*. *Noun_phrase_type* as subclasses: Demographic, Disease, and other UMLS semantic types. Comparison_Annotation has object properties *has_noun_ phrase* and *has_quantity*, and the *Quantity* class has a float value. When the quantity part of a comparison annotation is a string, it plays no part in the OWL formalization of ERGO Annotation. ERGO Annotations for complex statements are written as anonymous OWL expressions involving Noun Phrase Type and Comparison Annotation.

We illustrate the application of ERGO Annotations to our uses cases by annotating the three studies NCT00655538, NCT00655473, and NCT 00799903 taken from ClinicalTrials.gov. NCT0065553 8 and NCT00655473 have the inclusion criterion "adult patients, 18-75 years of age" and the exclusion criterion "poorly controlled diabetes." NCT00799903 has the inclusion criterion "age >= 18 years."

For "poorly controlled diabetes," the valid ERGO Annotations are a) C0011849 (`Diabetes_Mellitus`), b) C0743131 (`Uncontrolled_Diabetes`) and c) C0011849 (`Diabetes_Mellitus`) modified by C0205318 (`Uncontrolled`). We can write a necessary and sufficient definition of `Uncontrolled_Diabetes` as `Diabetes_Mellitus and (some attribute Uncontrolled)`. The annotation for "age > 18" is formalized as a Comparison Annotation "`has_noun_phrase some Age and has_quantity some (Quantity and has_value some float[> 18] and has_unit value year`), where `year` is an individual of the unit class.

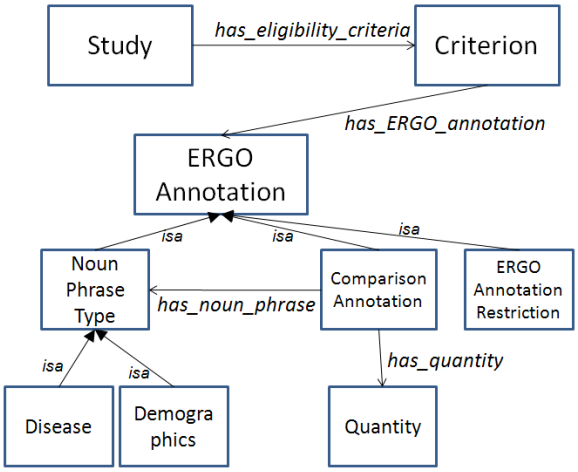

Figure 1. Predefined OWL ontology to illustrate how ERGO Annotations may be used to classify criteria and to search for studies

Examining our first use case, guideline authors searching for "Diabetes Mellitus" would retrieve all three ERGO annotations and their associated criteria and decide which one to use.

A query such as `(Study and exclusion_criteria some (Criterion and has_ergo_annotation some Diabetes_Mellitus))` will return studies with an exclusion criterion that includes any subclass of Diabetes_Mellitus (e.g, NCT00655538, NCT00655473). A query `Study and inclusion_criteria some (Criterion and has_ergo_annotation some (Comparison_Annotation and has_ quantity some (has_value some float [> 16.0])and has_noun_phrase some Age))` will return studies with inclusion criteria more restrictive than ages > 16 (e.g., all three of NCT00655 538, NCT00655473, and NCT0079990 studies).

DISCUSSION

We have shown how eligibility criteria can be annotated with ERGO Annotations that can be formulated as precise description-logic expressions involving terms from standard terminologies. With this approach,, we can build a library of eligibility criteria that are indexed by the classification hierarchy of ERGO Annotations. Our results show that such a library can be used for classifying studies, for querying studies that satisfy certain parameters, and as a source of standardized eligibility criteria for protocol authors. Furthermore, in preliminary work not reported here, we have seen initial promise on using ERGO Annotation to screen subjects for study eligibility.

ERGO Annotation changes the problem of representing

eligibility criteria from that of formal encoding in some expression language to that of classifying and decomposing criteria and identifying noun phrases in simple criteria. The ERGO expression language informs the categorization and decomposition of criteria into simple and comparison criteria connected by Boolean and other semantic connectors. ERGO Annotation aims to capture the basic semantics of criteria by identifying linguistic noun phrases and formalizing them as terminological expressions, instead detailed modeling based on some information model of data.

The advantage of ERGO Annotation over formal expression languages like Arden Syntax or GELLO is the scalability of its annotation process. No knowledge of arcane syntax is required and the assistance of automated tools to detect noun phrases and recognize terms from standard vocabularies is available. However, ERGO Annotation trades expressiveness for the scalability of the annotation process. ERGO Annotations, for example, do not capture the temporal requirements that an eligibility criteria may impose on a medical condition or therapy.

ERGO Annotations, expressed as OWL expressions can complement ASPIRE's sets of standardized eligibility codes by giving them an ontological foundation. Instead of having enumeration of standard codes such as "Breast Cancer Estrogen-Receptor Status (Positive/ Negative/Unknown)", we can associate criteria codes with their corresponding ERGO Annotations that are organized in a classification hierarchy, making semantic relationships among the eligibility codes apparent.

Another contribution of the ERGO Annotation work lies in reducing the variability of eligibility criteria text. Variant criteria such as "treated appropriately for dyslipidemia" and "Current treatment with statin therapy unless the study doctor determines statins are not appropriate for the subject" in the context of heart failure trials may be trying to target similar subjects. The rewrite rules developed for this project can help study authors to write eligibility criteria more clearly and uniformly. A standard library of eligibility criteria can reduce unnecessary variability in the target populations of studies thus making study results more comparable.

Our feasibility study shows relatively low sensitivity. The result is not surprising, given the unsophisticated NLP techniques used in this early work. Much can be done to improve the recognition rate of the tools. For example, the statistical NLP parsers can be trained on eligibility text. Furthermore, we will evaluate the use of advanced biomedical NLP tools such as MedLEE [10] and ChartIndex [11]. In addition, we are in the process of automating the heuristic algorithm for extracting noun phrases.

CONCLUSION

We have tested the feasibility of using a semi-automated approach for transforming text-based eligibility criteria into a formal representation. We demonstrated the capture of eligibility semantics that supports queries of sufficient richness to enable two important use cases in clinical research: finding candidate studies for patients and finding studies that use particular criteria. ERGO Annotation and our semi-automated approach provide an expressive ontological foundation for computable representations of eligibility criteria.

ACKNOWLEDGEMENTS

Supported in part by grant R01-LM-06780.

REFERENCES

[1] Niland J. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. 2007 [cited 2008 August 20]; Available from: http://hsspcohort.wikispaces.com/space/showimage/APIRE+CDISC+Intrachange+July+10+2007+Final.ppt.

[2] Hripcsak G, Clayton PD, Pryor TA, Haug P, Wigertz OB, Van der lei J, editors. The Arden Syntax for Medical Logic Modules. Proc Annu Symp Comput Appl Med Care; 1990; Washington, DC.

[3] Sordo M, Boxwala A, Ogunyemi O, Greenes R, eds. Description and Status Update on GELLO: a Proposed Standardized Object-oriented Expression Language for Clinical Decision Support. Medinfo; 2004.

[4] Tu SW, Peleg M, Carini S, Rubin D, Sim I. ERGO: A Template-Based Expression Language for Encoding Eligibility Criteria 2008. In: http://128.218.179.58:8080/homepage/ERGO_Technical_Documentation.pdf

[5] OpenNLP: http://opennlp.sourceforge.net/

[6] Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml

[7] Apple Pie Parser: http://nlp.cs.nyu.edu/app/

[8] http://obs.bioontology.org/oba/OBA_v1.1_rest.html

[9] http://lexsrv2.nlm.nih.gov/SPECIALIST/Projects/lexAccess/current/index.html

[10] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc2004 Sep-Oct;11(5):392-402.

[11] Huang Y, Lowe HJ, Klein D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc 2005 12(3): 275-85.