ERGO[1]: A Template-Based Expression Language for Encoding Eligibility Criteria

Samson Tu,[1] Mor Peleg,[2] Simona Carini,[3] Daniel Rubin,[1] Ida Sim[3]
[1]Center for Biomedical Informatics Research, Stanford University, Stanford, CA
[2]Department of Management Information Systems, University of Haifa, 31905, Israel
[2]Division of General Internal Medicine, UCSF, San Francisco, CA

I. Introduction

This document presents a simple summary of ERGO, a template-based expression language for encoding eligibility criteria, along with examples of its use. It describes the design objectives and components of the expression language and comments on the known limitations and intended semantics of the language with respect to negation and the open-world/closed-word question. ERGO was constructed using the Protégé Frame tool. The examples illustrate how one can use Protégé frame instances to encode eligibility criteria. The appendix shows a few of the patterns that HL7's TermInfo Committee recommends for encoding data using HL7 RIM and SNOMED CT. We recommend that these patterns be observed when encoding eligibility criteria.

II. Objectives

To capture the full expressivity of clinical eligibility criteria from any clinical domain, using an information model of components of criteria, consisting of noun phrases, statement and expression templates, which can be composed to specify individual's data that satisfy the criteria. The goal is to have ease of use without sacrificing expressiveness.

[1] ERGO stands for Eligibility Rule Grammar and Ontology, Despite its name, the use of ERGO is not limited to eligibility criteria and it does not yet contain a formal grammar definition, as it is still under development. It is formulated as an information model/frame-based ontology using the Protégé knowledge-modeling tool.

III. Components of Criteria

ERGO uses three types of components to model the structure of such criteria: (1) **Clinical_Statement_Template**, instances of which evaluate to TRUE, FALSE, or UNKNOWN when they are applied to data about individuals; (2) **Expression**, instances of which evaluate to non-Boolean values, and (3) Auxiliary constructs such as **ERGO Keyword**, **Modifier**, and **Constraint** that are not evaluated by themselves. Each eligibility criterion is an instance of a Clinical_Statement_Template.

A. Clinical_Statement_Template

There are two types of Clinical_Statement_Template: simple and complex. All Clinical_Statements are composed of Expressions (i.e., noun phrases, data values, variables, queries, or functions – see more in Section B) or of other Clinical_Statements.

1. Simple_Clinical_Statement

Simple_Clinical_Statements are statements about **Assessments** made about a person, **Interventions** done to him, or his **Behaviors** (e.g., person exercises at least 5 times a week). All eligibility criteria fall under one of these three types of Simple_Clinical_Statements, which thereby constitute a simple information model of what can be said about a potential study subject. Eventually, we probably want to substitute these statements with a standard information model (e.g., Clinical Statements Health Level Seven's Reference Information Model (HL7 RIM)) (Health Level Seven 2006). All Simple_Clinical_Statements have noun phrases (e.g., LDL_cholesterol) as the core of the statements, an effective time, and a mood. The effective time property specifies the time during which the statement is valid. For example, the effective time of a blood test result may be the time when the sample is taken. Similar to HL7 RIM mood codes, a mood modifies the simple clinical statement to specify whether the statement indicates an event that happened, an intent, or an order. Other mood codes may be added in the future. Current work on ERGO focuses on eligibility criteria that make use of Assessment statements; modeling Intervention and Behavior statements is future work.

2. Comparison_Statement

A Comparison_Statement compares two *Expressions* via comparators such as <,>, =

3. Complex_Clinical_Statement

Complex_Clinical_Statements are constructed from other clinical statements. They are of the following subtypes:

- Compound_Statement, a logical combination (and, or, not) of Clinical_Statements
- Semantically_Connected_Statement, where two Clinical_Statements are related via some semantic connector (e.g., causes, exacerbates, aka "act-act-relationship" in HL7 RIM terminology)
- Partially_Specified_Statement, where the encoder offers only an incomplete definition (e.g., “having major surgeries *such as* CABG or transplant”). This class is represented as a set of clinical statements linked via non-exhaustive and/or connectors. Thus, the statement "having major surgeries *such as* CABG or transplant” is true if either "having CABG" or "having transplant" is true. However, if we know that both "having CABG" and "having transplant" are false, the truth value of "having major surgeries *such as* CABG or transplant” is still unknown.

Taken together, these statements should be able to state most of what one wishes to state about a Subject. They can also be used to describe a high-level clinical phenotype of a person.

B. Expression

Expressions are used to construct Clinical Statements. Expressions are of 5 types:

1. **Data values**, which can be quantities, terminology codes, or time entities (interval or time point)
2. **Noun_Phrases**, which are primitive or post-coordinated terms that should be linked to standard terminologies (e.g., SNOMED-CT) for interoperability. There are four subclasses of noun phrases:
   - Primitive noun phrases which represent terms from vocabularies.
   - Logical combinations of noun phrases – noun phrases combined via and/or, or a noun phrase that is negated. 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.
   - Modified noun phrases are noun phrase with modifiers that place restrictions on the root noun phrases. Modifiers follow the entity-attribute-value (EAV) model (e.g., 'asthma

exacerbated_by exercise'). In this way, the root noun phrase (e.g., 'asthma') is modified via a Relational_Modifier_Term (i.e., terms, such as 'exacerbated_by' that relate one noun phrase to another) to another term (e.g., "exercise")

The subject of all clinical statements, as well as modifier terms, refer to noun phrases.

3. **Variables** that have associated derivation expressions (e.g., the 'age' variable whose derivation expression is a function that take the difference between the current date and date of birth).
4. **Functions** that operate on variables and evaluate to a non-Boolean data type. For example, a function may take two variables representing the current date and a calendar date (e.g., date of birth), and compute the age in years.
5. **Queries** in two flavors: (1) Subject_Query that can select attribute values from the instances of a person's data that match the constraints specified in a Simple_Clinical_Statement. A Subject_Query can also specify an aggregate term (e.g., average) to aggregate the results that match the Simple_Clinical_Statement. A subject query is analogous to a SQL query. It constructs an expression of the form: *select {aggregate term} [attribute or *] from instances of Clinical_Statement_Tempate where {constraints on attributes of Clinical_Statement_Tempate*}. (2) Aggregate_Subject_Query that aggregate an expression or a clinical statement over a number of persons.

C. Auxiliary Components

1. ERGO Keyword

This ERGO class hierarchy contains keywords that are used to build complex statements, expressions, and noun phrases. Subclasses of the Keyword class include aggregation terms (e.g., maximum, most recent), logical statement connectors (and/or/not, non-exhaustive and/or), expression comparators (e.g., >, =), and temporal comparators (according to Allen's Interval Algebra, (e.g., before, meets) (Allen 1983)).

2. Modifier

A modifier (Modifier) consists of a modifier attribute, such as 'severity', 'exacerbated_by', and 'has_location', taken from a controlled vocabulary, and a modifier value that should be a noun phrase. The purpose of a modifier is to place restrictions on noun phrases to construct modified noun phrases such as 'asthma exacerbated_by exercise', 'bone marrow toxicity (with) severity moderate, and 'fracture_of_bone with finding_site [Bone structure of lower limb with laterality

left]' for fracture of the left leg.

3. Constraint

Constraints can be (1) simple value constraints that define minimum and maximum bounds and (2) constraints on time intervals (usually the effective times of query templates). The time-interval constraints may be constraints on duration (e.g., <3 days), constraints that compare two time intervals using Time_Interval_Comparators (discussed in the following section), and compound time constraints that are logical combinations of time constraints.

IV. Semantics of the Language

ERGO does not have a formally defined semantics. This section discusses the types of negation supported in ERGO and the open-versus-closed world issue, and gives some recommendations on preferred representational choices for encoding standardized eligibility criteria.

A. The Use of Negation

When writing criteria on clinical statements, it is often difficult to know how to express negative or absent information properly. In the case of an Assessment (e.g., an observation), sometimes you have choices about what to put into the code and value slots. For ERGO, we propose that, as much as possible, we use SNOMED-CT terms (Spackman 2000) and follow the recommendations of the HL7 TermInfo committee, which spelled out how HL7 clinical statements should be used with SNOMED-CT terms (HL7 TermInfo Project 2007). Some of the patterns recommended by the TermInfo committee are described in Appendix A. In ERGO, we use three patterns of negation:

1. Not known, not done, refused, etc.
   We propose to use, whenever possible, SNOMED-CT situation with explicit context that uses finding context, (e.g., known absent, known but not specified, unknown) to construct post-coordinated noun phrases (IHTSDO 2008). In Figure 1(b), we see an example of an Assessment, whose noun phrase is "clinical finding absent with associated-finding Frank Hematuria", which is constructed from the SNOMEDCT term "clinical finding absent" and the binary modifier "associated finding" = "Frank Hematuria".
2. The set complement of a collection of terms
   Use the *Not* logical noun phrase modifier as set complement. (e.g., Anti-diabetic agent not

metformin (see Figure 1(d)), pulmonary problem excluding asthma) to specify a noun phrase that uses negation to define a narrower set of terms

3. Boolean negation of statements
   Use Boolean negation of statement, for example, "NOT (Observation WBC < 4000 /mm3)"

B. Open/Closed-World Assumption

We propose that, when used as a specification language for generic eligibility criteria, ERGO makes the open-world assumption. If something is not explicitly false, then it is unknown. When criteria are implemented in a real system, the implementer has to decide whether a null query result means the object of query is truly absent in the patient (i.e., the implementer decides when and how to close).

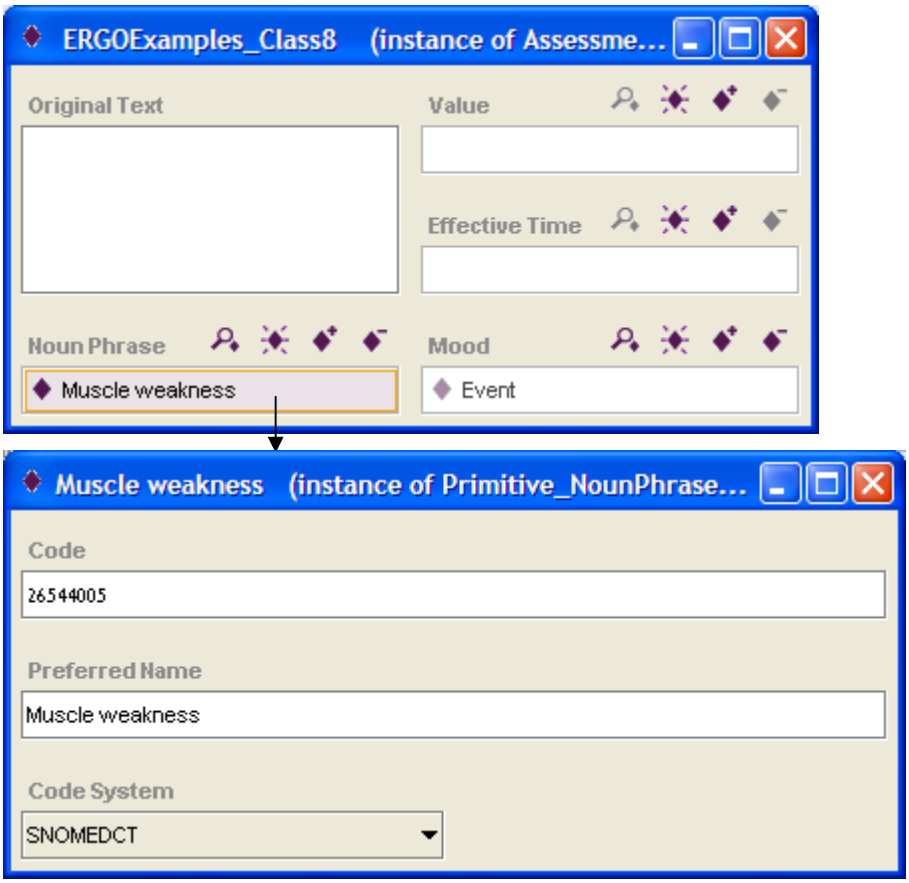
C. Preferred Representational Choices

Given an eligibility criterion such as "Not pregnant," ERGO permits multiple possible encodings: (1) an Assessment with noun phrase = "not pregnant" or (2) an Assessment with noun phrase = "clinical finding absent" whose associated finding is "patient currently pregnant," or (3) Boolean negation of an Assessment whose noun phrase is "patient currently pregnant." Option 1 (using a pre-coordinated term that incorporates the negation concept) should be avoided, as we want to make the semantics of criteria as explicit as possible. In general, option 2 (using SNOMED situation with explicit context) is preferred over option 3 because the semantics of option 3 is dependent on the open/closed-world assumption. Under closed-world assumption, the statement is false when it's not known whether the subject is pregnant, but has 'unknown' truth value under open-world assumption.

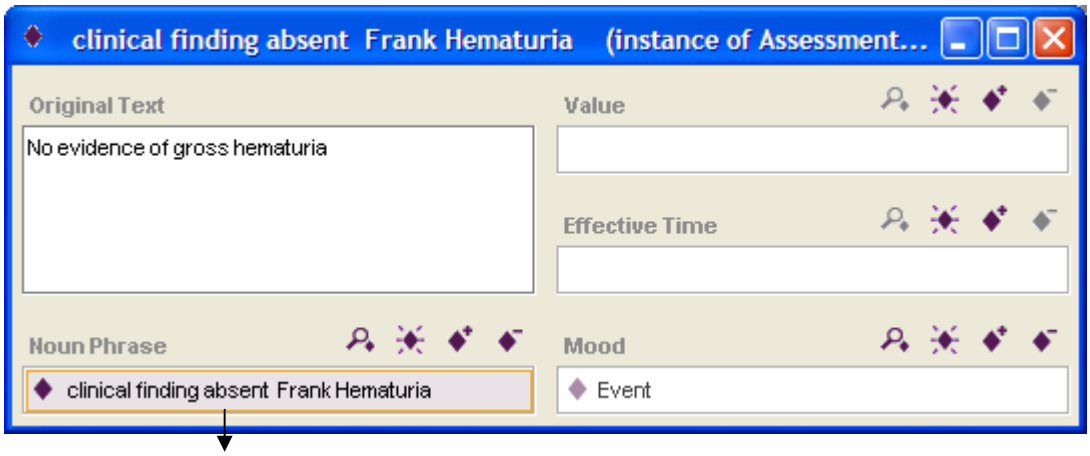
In cases not involving negations, we recommend using modifiers on root concepts to make the meaning of terms as explicit as possible, except when a term (such as 'breast cancer') is fully defined in SNOMED CT. In the latter case, using the existing terms does not cause the loss of semantic precision.

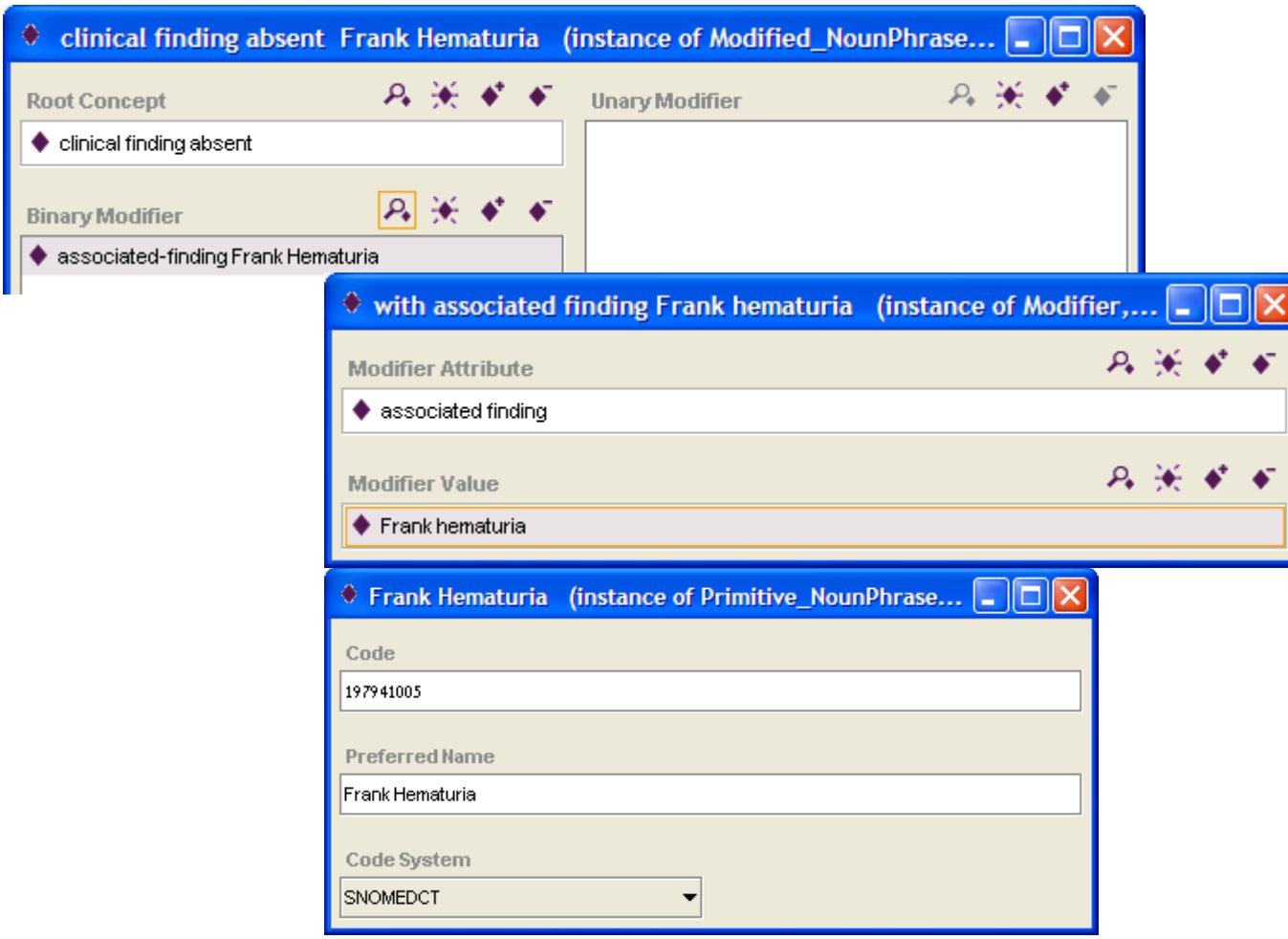
V. Examples of eligibility criteria encoded as clinical statements

a) A simple clinical statement (Assessment) with a simple noun phrase: Presence of weakness

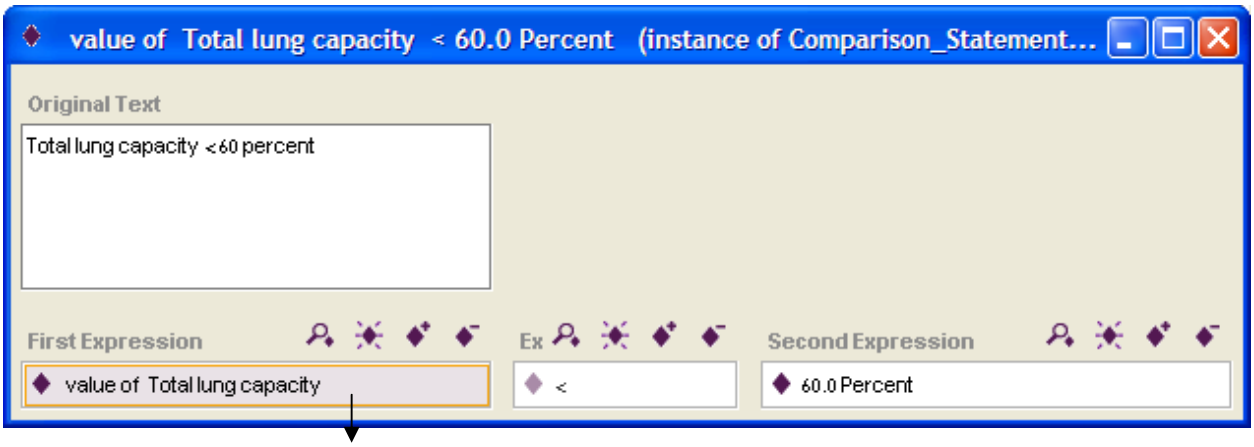


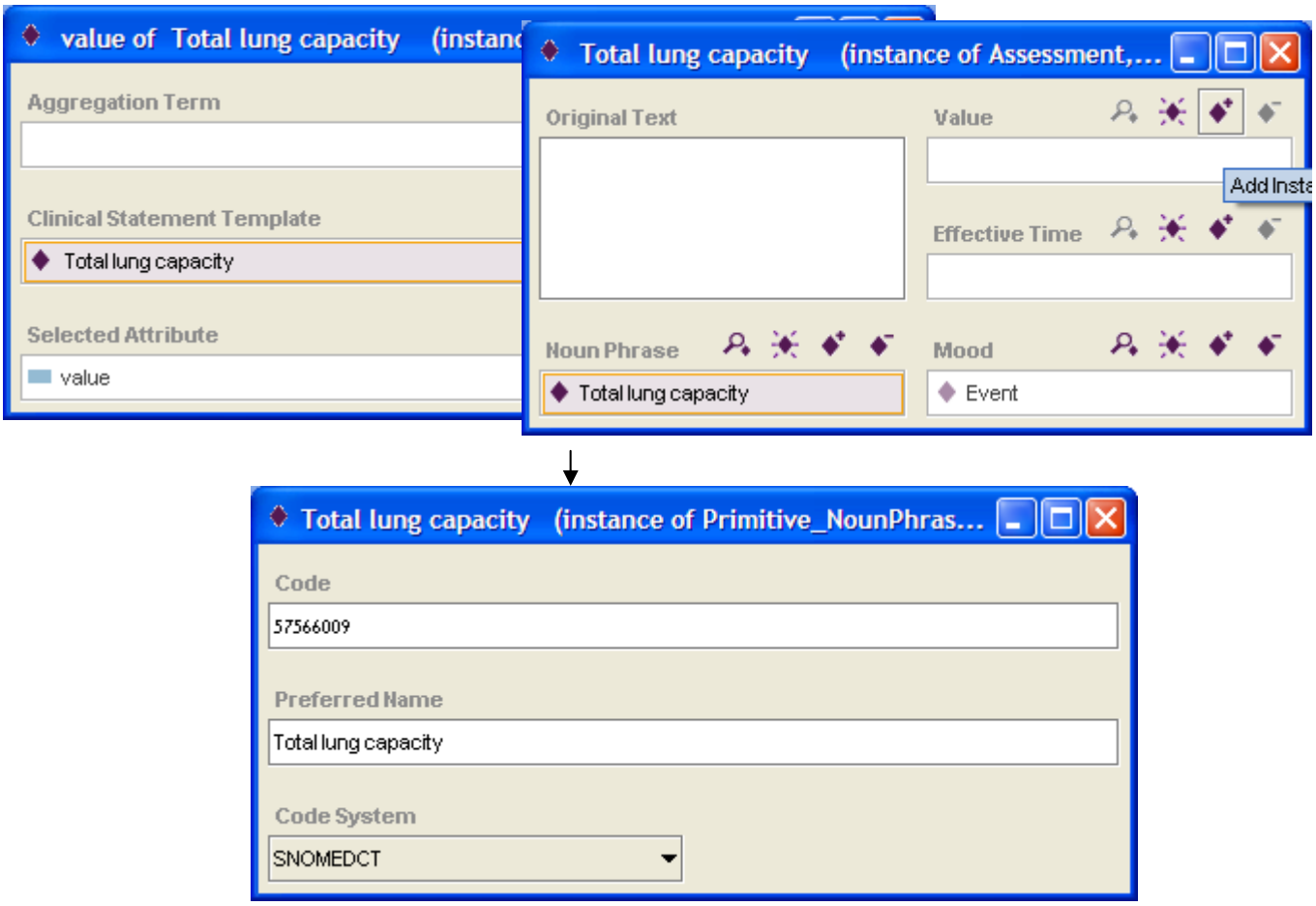
b) A simple clinical statement (Assessment) that uses the SNOMEDCT finding with associated context pattern. The criterion "No evidence of gross hematuria" is modeled as an Assessment with the noun phrase composed from SNOMEDCT term "clinical finding absent" (373572006) and the binary modifier "associated finding" = "Frank Hematuria" (197941005).



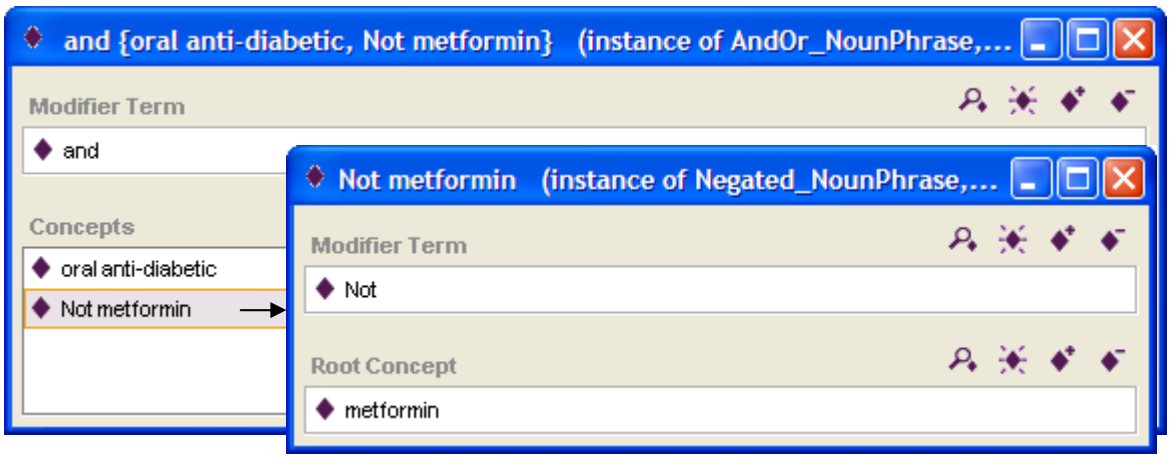


c) A comparison statement that compares the queried value of total lung capacity to a threshold





d) A logical combination noun phrase using negation (set complement): oral anti-diabetic excluding metformin



VI. Possible Uses of ERGO

ERGO could be used in various kinds of applications, such as an application that checks encoded eligibility criteria against an electronic medical record and finds patients who match the eligibility criteria. It can also be used to formulate expressions that define outcome variables of studies or decision criteria used in formalizing clinical practice guidelines for decision support.

VII. Known Limitations

A. To Be Addressed

1. The set of Simple_Clinical_Statements requires refinement and alignment with a standard information model.
2. We do not have a satisfactory approach for building standard sets of modifiers (e.g., "severity severe") and semantic connectors (e.g., "caused by"). UMLS has a list of semantic connectors, but they have heterogeneous meanings. OBO Relation Ontology is a formal model of semantic relationships. However, the relationships enumerated there are very basic and not directly usable for modeling clinical relationships. At the moment we are using SNOMED CT's "concept model attribute" (410662002) and "unapproved attribute" (408739003) as starting points.

B. Currently Out of Scope

a) Iteration: (e.g., "less than a total of 4 months therapy with didanosine", which requires iteration over each didanosine medication records and sums up their durations.)

b) Queries whose temporal constraints involve coalescing the temporal extents of multiple intervals (e.g., the maximum duration of lisinopril medication regardless of dose levels should be less than 4 months (which may involve concatenating a number of lisinopril medication records)).

IX. Appendix Examples of Encoding Patterns Recommended by the HL7 Terminfo Committee

Taken from *Using SNOMED CT in HL7 Version 3; Implementation Guide, Release 1.4*.

PATTERN ONE: Observation.code [(<<363787002 | Observable entity) OR (<<386053000 | Evaluation procedure)] ; Observation.value = not null (e.g. numeric, nominal, ordinal, coded result).

Example 5. Observation code/value: observable entity with result

```
<observation classCode="OBS" moodCode="EVN">
  <code code="50373000"
    codeSystem="2.16.840.1.113883.6.96"
    displayName="Height"/>
  <text>Height: 177 cm</text>
  <value xsi:type="PQ" value="1.77" unit="m"/>
</observation>
```

PATTERN TWO: Observation.code = "ASSERTION" (codeSystem="2.16.840.1.113883.5.4"); Observation.value [(<<413350009 | Finding with explicit context) OR (<<404684003 | Clinical finding)] .

Example 6. Observation code/value: assertion of a clinical finding

```
<observation classCode="OBS" moodCode="EVN">
  <code code="ASSERTION"
    codeSystem="2.16.840.1.113883.5.4"/>
  <text>Headache</text>
  <value xsi:type="CD" code="25064002"
    codeSystem="2.16.840.1.113883.6.96"
    displayName="Headache"/>
</observation>
```

An alternative form of Pattern 2, when more explicit context is needed, is the pattern of clinical finding with explicit context:

Example 7. Observation code/value: assertion of a clinical finding with explicit context

```
<observation classCode="OBS" moodCode="EVN">
  <code code="ASSERTION"
    codeSystem="2.16.840.1.113883.5.4"/>
  <text>Presence of headache</text>
  <value xsi:type="CD" code="373573001"
    codeSystem="2.16.840.1.113883.6.96"
    displayName="Clinical finding present">
    <qualifier>
      <name code="246090004" displayName="Associated finding"/>
      <value code="25064002" displayName="Headache"/>
    </qualifier>
  </value>
</observation>
```

X. References

Allen, J. (1983). "Maintaining knowledge about temporal intervals." Communications of the ACM.

Health Level Seven. (2008). "HL 7 Reference Information Model." accessed 2008, from http://www.hl7.org/library/data-model/RIM/modelpage_mem.htm.

HL7 TermInfo Project (2007). Using SNOMED CT in HL7 Version 3; Implementation Guide, Release 1.4, Health Level 7.

IHTSDO (2008). SNOMED CT Style Guide: Situations with Explicit Context.

Spackman, K. (2000). "SNOMED RT and SNOMEDCT. Promise of an international clinical terminology." MD Comput **17**(6): 29.