

# Patterns of Clinical Trial Eligibility Criteria

Krystyna Milian<sup>1</sup>, Annette ten Teije<sup>1</sup>, Anca Bucur<sup>2</sup>, and Frank van Harmelen<sup>1</sup>

<sup>1</sup> Vrije Universiteit Amsterdam,

<sup>2</sup> Philips Research,

**Abstract.** Medical research would benefit from automatic methods that support eligibility evaluation for patient enrollment in clinical trials and design of eligibility criteria. In this study we addressed the problem of formalizing eligibility criteria. By analyzing a large set of breast cancer clinical trials we derived a set of patterns, that capture typical structure of conditions, pertaining to syntax and semantics. We qualitatively analyzed their expressivity and evaluated coverage using regular expressions, running experiments on a few thousands of clinical trials also related to other diseases. Based on an early evaluation we conclude that derived patterns cover the language of eligibility criteria to a large extent and may serve as a semi-formal representation. We expect that extending the presented method for pattern recognition with recognition of ontology concepts will facilitate generating computable queries and automated reasoning for various applications.

**Key words:** eligibility criteria, patterns of eligibility criteria, regular expressions for eligibility criteria, formalization of eligibility criteria

## 1 Introduction

New approaches to prevention, diagnostic, medication and treatment methods are tested during clinical trials. These can be carried out only when a sufficient number of eligible candidates is identified and enrolled. We are looking for a method that facilitates the formalization of eligibility criteria to support automated reasoning for various applications, i.e. determining patient eligibility for clinical trials or designing eligibility criteria. The observed similarity and repeatability of eligibility criteria of different breast cancer trials published at ClinicalTrials.gov [2] inspired us to investigate the possibility of capturing this specialized language by defined set of patterns, pertaining to syntax and semantics. In order to support automated mining of patterns and concrete eligibility criteria we proposed a multidimensional classification, providing metadata about the content. We will use the defined patterns in the formalization process as an intermediate step between free text and computable semantic queries. For instance, we can recognize in an inclusion criterion 'Has received chemotherapy within the past 14 days' the pattern 'prior () within ()' linked to the query 'select all patients with a prior () and timestamp of () less than ()'. We plan to generate computable queries by linking patterns with corresponding queries and

filling them with concrete retrieved data. This requires interoperability between the eligibility criteria and the patient data model. We expect that annotating criteria with ontology concepts and defining mappings to data items in EHR when needed facilitates the process.

This paper presents the defined set of patterns, a classification of these patterns, and an approach to evaluating their coverage and expressivity. Additionally we present findings about the most common types of identified patterns in clinical trials from different domains: breast cancer, lung cancer and diabetes. Finally, we describe our ideas on how to proceed from the presented semiformal representation to generating computable queries. The paper is organized as follows; section 2 introduces the patterns and their classification, section 3 describes the evaluation of the patterns in terms of coverage and expressivity, and presents observations about identified patterns. Section 4 provides information about related work, section 5 gives conclusions and describes plans for future work related to the formalization of eligibility criteria and supporting their design.

## 2 Classification of patterns

In this study we analyzed eligibility criteria of clinical trials published at ClinicalTrials.gov [2], a service of the U.S. National Institute of Health. Its search engine allows to specify various categories such as conditions being studied, interventions, outcome measures, recruitment status, study type and others. We focused on breast cancer trials since this group contains many examples and because we have access to domain knowledge related to this disease. The latter might be important in further steps of our research.

The analysis of eligibility criteria specified for various breast cancer trials, allowed us observe vast similarity and repeatability of criteria across the trials. It inspired us to define a set of patterns and analyze to which extent they capture the language, used to define eligibility criteria. We started an informal development process by extracting eligibility criteria from the description of all available breast cancer trials (3905). Further we focused the analysis on a randomly selected subset, containing approximately few hundreds of trials. To identify common ways of expression we manually grouping conditions by similar subject (demographic information, disease characteristic, prior- concurrent treatment) or similar syntax. We noticed that criteria differ in the level of complexity. Some are formulated as atomic phrases e.g. 'Not pregnant', others as complex sentences e.g. 'Brain metastases allowed provided they have been treated with surgery.' We aimed to define patterns covering both groups, incrementally extending a set of patterns.

The method developed during this formalization process, inspired by observed concrete examples of eligibility criteria, can be summarized as follows. In order to cover sentence structure we started from basic forms e.g. "must be receiving ()" and added corresponding negated versions "can not be receiving", as well as past tense, both positive and negative e.g. "must have received

()” and future if applicable. Secondly we extended the resulting basic forms with common specifications, which restrict for example time frame, purpose of a treatment or co-occurrences. If applicable, these were combined. An example of a pattern containing two specifications: time frame and exclusions is ‘more than () since prior () except for ()’, capturing criteria like ‘More than 6 months since prior endocrine therapy, except tamoxifen’. Additionally, we defined patterns that capture atomic phrases, covering value restrictions for chosen parameters, expressed by arithmetic comparison or enumerated values, and their negations. Patterns that capture atomic phrases can be nested in the patterns reflecting sentence structure. As a result 130 different patterns were defined. We can relate this number to the average number of conditions specified in breast cancer trials which is 25, according to [8]. In section 3.3 we present the results of the experiment performed to identify most common patterns.

To support automated mining, the patterns were classified according to the dimensions and classes described below. Table 1 presents examples of patterns belonging to each dimension and class, concrete criteria from ClinicalTrials.gov [2] instantiating them, and the percentage of that class of patterns. For instance 58% of our patterns belong to the dimension “time independent status”, 29% belong to the class “present”. An example of a pattern of the class “present” is “diagnosis of”, an instance of the pattern “diagnosis of” is “diagnosis of malignancy”. Notice that a pattern can belong to several dimensions and classes, because they are not mutually exclusive. The dimensions and corresponding classes are:

- *Time independent status dimension.* Classes of this dimension are: present, absent, conditional, potential, not selective. This dimension reflects required status of data items in patient record.
- *Temporal status:* historical, current, planned. This dimension indicates whether a condition regards patient history, current findings or future plans.
- *Specification type:* time frame, including and excluding findings or therapies, value restrictions, purpose of a drug/treatment, co-occurrences, number of occurrences, confirmation, outcome constraint.
- *Medical content:* demographic data (age, gender), clinical data (pregnancy and nursing, menopausal status, adverse reactions), pathology and molecular data, interventions (prior and current therapies). Complete recognition of content will be done using ontologies, the patterns are only supposed to provide the context for annotation.
- *Data source of medical content.* This dimension is dependent on concrete EHR, it is meant to support automatic information extraction.
- *Variability and controllability:* stable, variable, controllable, subjective. This dimension was proposed in [12] and reflects the possibility of change of criteria evaluation over time. In most cases classification according to this dimension will require incorporating domain knowledge.
- *Subject:* candidate, family of a candidate

The possibility of classifying a pattern according to its each dimension depends on its details. Let us consider the pattern ‘no prior’. Its time independent

**Table 1.** Examples of categorized patterns and instances. The numbers denote the percentage of all 130 patterns classified accordingly.

<b>Dimension</b>	<b>%</b>	<b>Example of pattern</b>	<b>Example of instance</b>
<b>Dimension: Time independent status</b>			
present	29	Diagnosis of()	Diagnosis of malignancy.
absent	16	No concurrent ()	No concurrent endocrine therapy.
conditional	13	() allowed if ()	Multicentric breast tumors are allowed if all foci are ER-negative.
potential	3	known or suspected	Known or suspected pregnancy.
not selective	4	Prior () allowed	Prior biologic therapy allowed.
<b>Dimension: Temporal status</b>			
historical	26	No history of ()	No prior chemotherapy.
current	36	Allergy to ()	Allergy to bisphosphonates.
planned	4	Planned () required	Scheduled for prostatectomy
<b>Dimension: Specification type</b>			
time frame	15	at least () since prior ()	At least 3 weeks since prior steroids.
inclusions	3	No concurrent () including ()	No other concurrent anticancer therapies, including chemotherapy.
exclusions	5	No prior () except for ()	No prior malignancy, except for adequately treated basal cell.
value restrictions	25	T () stage; Age above ()	T2; Age >18
treatment purpose	5	Required prior () for ()	At least 1 prior chemotherapy regimen for advanced disease.
co-occurrences	2	No concurrent () with ()	No concurrent radiotherapy with chemotherapy.
confirmation	8	confirmed by ()	No metastasis to brain (confirmed by CT or MRI)
occurrences	2	completed () courses of ()	Received 4-7 courses of doxorubicin or taxane based regimen
specific value	15	can take oral medication	Able to swallow whole tablets.
<b>Dimension: Medical content</b>			
Age	5	() and over	18 and over.
Gender	2	female	Female.
Menopausal status	2	post-menopausal	Postmenopausal.
Pregnancy & nursing	3	not pregnant	Negative pregnancy test
Adverse reactions	3	no allergy to ()	No allergy to sulfonamides.
Pathology data	7	margins must be clear	Resected margins histologically free of tumor.
Molecular data	2	Known gene mutation	Documented BRCA1/2 mutation.
Therapy	20	required prior ()	Must have undergone lumpectomy
<b>Dimension: Variability &amp; controllability</b>			
stable	26	history of ()	History of breast cancer.
controllable	1	must use contraception	Patients must use effective nonhormonal contraception.
subjective	1	() in the opinion of investigator	Life expectancy of 12 weeks or more in opinion of investigator.
<b>Dimension: Subject</b>			
candidate's family	5	family history of ()	Family history of colon cancer.

and temporal status can be classified as 'absent' and 'historical' respectively. However its medical content depends on concrete instantiations, it is the cancer type for 'No prior breast carcinoma' and the treatment for 'No prior chemotherapy'. The purpose of described classification is to annotate the patterns with metadata, characterizing their content from various perspectives. Although most often the classification of patterns is case specific we can formulate general rules, based on correlations between dimensions, specifying which annotations can be expected for particular patterns. The general rules that we have identified are:

1. Patterns with 'historical' temporal status and 'molecular' medical status are classified as stable in variability and controllability dimension.
2. Patterns having specification type: exclusion, co-occurrences, confirmation contain implicit condition therefore can be classified as conditional in time independent dimension, if it supports evaluation.
3. Patterns having medical content: pathology, age, gender, pregnancy and nursing, menopausal status have specification type: value restriction.

Full classification according to each dimension is possible only after instantiating a pattern with concrete data.

The described classification of patterns will facilitate formalization and design of eligibility criteria. We will discuss this in more detail in section 5. In the next sections we evaluate this set of classified patterns.

### 3 Expressivity and coverage of patterns

This section presents the approach to evaluate the expressivity and coverage of defined set of patterns. The expressivity was analyzed by performing a case study for the single trial, Neo ALTTO (Neoadjuvant Lapatinib and/or Trastuzumab Treatment Optimisation) Study [3]. The choice of this trial was driven by the access to domain knowledge related to this study, which will be important in further stage of research. We evaluated the possibility of using the defined set of patterns to express eligibility criteria from this trial. Further we estimated the coverage by analyzing a large set of clinical trials related to breast cancer, lung cancer and diabetes, published at ClinicalTrials.gov [2], containing 3905, 2949, 5499 trials respectively. Since the patterns were derived using a subset of breast cancer trials we were interested to analyze the differences obtained by performing the same experiment for trials related to another cancer type, and trials from a completely different domain - diabetes. From each trial we extracted all eligibility criteria and counted the occurrences of the defined patterns. The experiment was performed in order to estimate the fraction of covered criteria, identify classes of most common patterns and analyze the differences in the selected medical domains.

#### 3.1 Expressivity of patterns

The expressivity of the defined set of patterns was evaluated by performing a case study for the Neo ALTTO trial [4]. The trial contains 39 eligibility criteria, 21

inclusion and 18 exclusion conditions. We analyzed all of them in order to identify corresponding patterns which could be used for semi-formal representation. After analysis we distinguished following cases:

1. Criteria with syntax corresponding to one of defined patterns (17/39).  
In this case a pattern which could be used for representation can be automatically suggested as demonstrated in table 2.
2. Criteria whose meaning can be reflected using defined patterns, but need reformulation (19/39).  
In this case a corresponding pattern needs to be manually chosen, as shown in table 3 .
3. Criteria whose meaning exceeds the scope of defined set of patterns, and can only partially be expressed (3/39).  
An example of such criterion is 'Over expression and/or amplification of HER2 in the invasive component of the primary tumor [Wolff et al 2006] and confirmed by a certified laboratory prior to randomization'. We can associate it with the pattern regarding over expression of HER2, but a pattern allowing to specify a corresponding tissue is missing.

**Table 2.** Examples of criteria with correctly identified patterns

Criteria	Corresponding pattern
"Hemoglobin at least 9 g/dl"	() at least ()
"Histologically confirmed invasive breast cancer"	histologically confirmed ()
"Performance Status-ECOG 0-1"	value in range () - ()
"Diagnosis of inflammatory breast cancer"	diagnosis of ()

We performed this case study to evaluate the expressivity of the defined set of patterns. For our case study trial 36 criteria out of 39 could be represented using defined patterns, either directly or after reformulation. We conclude that at this stage of our research, the expressivity of patterns is sufficient to facilitate formalization. They can be used as a semi-formal representation, and serve as a starting point for generating computable queries. Nevertheless, being aware of limitations coming from performing a case study on a single trial, we expect that future work will provide insights which might motivate us to extend or modify the presented set of patterns as well as their classification.

### 3.2 Coverage of patterns

In order to evaluate the coverage of the defined set of patterns across medical domains we analyzed eligibility criteria from breast cancer, lung cancer and diabetes, published at [2]. We calculated a number of occurrences of each pattern in the set of eligibility criteria using regular expressions.

**Table 3.** Examples of criteria matching one of patterns after reformulation.

Criteria	Reformulated version	Corresponding pattern
No evidence of metastasis (M0) (isolated supraclavicular node involvement allowed).	No evidence of metastasis (M0) except for isolated supraclavicular node involvement.	no () except for ()
In the case of known Gilbert’s syndrome, a higher serum total bilirubin (< 2 x ULN) is allowed.	Higher serum total bilirubin (< 2 x ULN) is allowed if known Gilbert’s syndrome.	() allowed if ()
Exclusion criteria: Received any prior treatment for primary invasive breast cancer.	No prior treatment for primary invasive breast cancer.	no prior () for ()

When constructing regular expressions we had to consider both precision and recall. To increase recall, we aimed to capture various synonym forms of words such as allowed/permitted, and syntax e.g. "no (other)?concurrent.\* for" and "concurrent.\* for.\* is not (allowed|permitted)". To increase precision, we tried to capture only desired words by applying negative/positive lookbehind - which allow to specify string which cannot/must precede considered text. It was useful for instance in case of pattern capturing M stage of cancer, usually specified as M preceding a number or a range of numbers, to avoid matching also units of measurements (mm, or m<sup>2</sup>). Nevertheless some cases are impossible to distinguish without knowing the context.

In total we defined 342 regular expressions corresponding to 130 patterns.

Ideally pattern recognition should be performed condition by condition, however delimiting them is a challenging task. We approached the problem by delimiting sentences using existing NLP tools, starting with preprocessing to support the task. The preprocessing step regards trials which are edited according to the same template, specifying the domain followed by a colon and the condition e.g. "Age:.. Performance status:..Cardiovascular:.. Chemotherapy: ..., etc". Taking advantage of that clear separation which is rarely the case, we inserted full stops between conditions regarding different subjects. Secondly we delimited sentences using GATE [6], the open source framework for text processing. Our matching algorithm used the output of GATE, analyzing eligibility criteria sentence by sentence. Each sentence can correspond to more than one pattern. From the set of identified patterns in the sentence, we counted only those that cover longest phrases, and skipped patterns capturing segments subsumed by others. For example in the sentence 'No other concurrent hormonal therapy, including steroids', we identified two patterns 'no concurrent ()' and 'no concurrent () including ()', from which only the latter was calculated, because it reflects the content closer. Table 4 presents our obtained results with statistics about the identified patters.

**Table 4.** Coverage in different trial domains

	Breast cancer	Lung cancer	Diabetes
No. of trials	3905	2949	5499
No. of sentences processed	111334	119547	86526
Avg. no. of sentences in eligibility criteria per trial	28	29	15
Sentences with identified patterns	71 %	69 %	54%

Eligibility criteria from breast cancer trials have the largest number of sentences containing at least one identified pattern. This result is not surprising since the patterns were defined using eligibility criteria from breast cancer, and cover conditions which are typical for this tumor. Results for lung cancer are relatively similar, while in case of diabetes approximately 17% more sentences have no identified pattern. An interesting observation regards the average length of eligibility criteria: in diabetes trials they seem to be more compact, and on average half as long as breast or lung cancer eligibility criteria.

Among the sentences not covered by any of the patterns are criteria formulated as a list of excluding or including concepts.

There are two main reasons why obtained information about coverage is only an approximation:

- Unsuccessful identification of criteria which are in the scope of defined patterns. This regards criteria expressed using different synonym forms than those covered by regular expressions.
- Errors in identification, caused by insufficiently restrictive regular expressions. An example is matching the criterion 'No spontaneous menses for > 12 months' with a pattern 'no .\* for', which was supposed to match criteria restricting the purpose of a treatment e.g. 'No other concurrent therapy for cancer'. In order to avoid such errors we would need to either add patterns reflecting criteria with different meaning but similar lexical form, or use the help of an ontology annotator to recognize the semantic type of the criterion content. Both approaches will be considered in future work.

Nevertheless the results provide a useful estimation of the coverage of the defined pattern set and common practice of expressing eligibility criteria. More extensive evaluation of precision and recall is left for future work. It will be important in order to find out the fraction of criteria correctly assigned to a pattern, and among unrecognized patterns the fraction of criteria whose meaning can be reflected using the defined patterns but require reformulation and which exceed the scope of defined set.

### 3.3 Most common patterns

The results of the experiment described above were additionally used to analyze the most common patterns in eligibility criteria. Table 5 presents the statistics

about the number of patterns corresponding to each dimension. Additionally presented is the distribution of identified patterns over possible values, belonging to the same dimension.

**Table 5.** Percentage of identified patterns corresponding to each dimension, and their distribution among classes.

<b>Dimension</b>	<b>Breast cancer</b>	<b>Lung cancer</b>	<b>Diabetes</b>
Time independent status	46%	46%	40%
present	59%	61%	93%
absent	27%	29%	7%
conditional	3%	3%	0.2%
potential	0.3%	0.3%	1%
not selective	10%	7%	0.3%
Temporal status	42%	43 %	37%
historical	43%	44%	53%
current	57%	56%	46%
planned	2%	1%	2%
Specification type	46 %	46%	51%
time frame	7%	8%	2%
inclusions	0.9%	0.9%	0.1
exclusions	1.5%	1.6%	0.5
value restrictions	74%	73%	90%
purpose of treatment	6%	5%	1%
outcome	1.5%	2%	0
co-occurrences	0.2%	0.2%	0
confirmation	4%	5%	0.7%
occurrences	0.6%	0.5%	0.1
Medical content	22%	21%	24%
Age	6%	6%	11%
Gender	18%	15%	33%
Menopausal status	5%	1%	3%
Pregnancy & nursing	13%	17%	7%
Adverse reactions	6%	7%	11%
Pathology data	15%	16%	7%
Molecular data	0.6%	0.2%	0
Therapy	30%	32%	17%
Subject	0.1 %	0.05 %	0.2%
Candidate's family	0.1 %	0 %	0.2%

The obtained results are relatively similar for breast and lung cancer, while many differences can be observed for diabetes. This is to be expected taking into account the nature of mentioned diseases.

For all diseases, the most often identified patterns can be classified according to 'specification type', in case of breast and lung cancer 46% of identified patterns,

diabetes - 51%. In most cases these patterns fell into the value restriction class (74%, 73% and 90%), indicating e.g. criteria limiting lab results. The difference between the observed frequencies between diabetes and both cancer trials is significant in this case. The patterns that reflect time frame, purpose of a treatment or a confirmation constraints can be frequently found in cancer trials, whereas in diabetes trials they are hardly ever identified.

Another major difference is observed for time independent status: for all diseases most of the identified patterns require the presence of some findings. In case of breast cancer it is 59% of identified patterns, lung - 61% , for diabetes it is a significantly larger number 93%. Again eligibility criteria of diabetes trials seem to be simpler, conditional criteria cover only 0.2%.

Considering medical content in all cases approximately 20% of identified patterns can be classified according to this dimension. Some differences can be observed among the distributions. Eligibility criteria of cancer trials frequently mention pathology data and received or undergoing therapies in contrast to diabetes criteria. Another finding regards eligibility criteria considering family history, they are more frequently identified in diabetes trials.

Results of the described experiment allowed us to analyze similarities and differences between domains, and identify which patterns, developed for eligibility criteria of clinical trials related to breast cancer, could be reused for other diseases. Some of the patterns are common for all domains (e.g. patterns restricting lab values), others are typical for both cancer types (e.g. patterns related to pathology data), and the rest is breast cancer specific (e.g. patterns related to molecular data).

The presented results provide some observations about the most frequent patterns across domains. Our final goal is to reach a computable representation of criteria. Obtained information could be used to guide the order of formalization. The process could start from the most frequently observed conditions. However frequency is not necessarily correlated with importance and selectiveness, which will be also taken into account.

## 4 Related work

An informative analysis of eligibility criteria can be found in [9]. It was done using 1000 randomly chosen eligibility criteria from clinical trials published at CT.gov. Criteria were manually analyzed and categorized along several axes: complexity of conditions (simple or complex), high level clinical content (clinical attribute of the study participant, treatment, behavior of participant) and semantic and clinical patterns (related to demographic data, lab and test results, temporally related criteria, requiring clinical judgment, requiring data beyond criterion itself) . This classification presents a very informative overview of types of eligibility criteria. Our paper describes one step forward. We defined more fine-grained dimensions of comparison and described a method for automatic classification, which allows to analyze sets of significantly larger dimensions. Moreover, our

patterns together with the classification approximate the meaning of conditions, and therefore can facilitate generating computable queries.

There are several languages which could be applied for expressing eligibility criteria e.g. Arden syntax [13], Gello [10], ERGO [11] and others. The rich overview of existing options is presented in [14]. For our application we require applying ontologies and semantic reasoning, which determines the need of expressing eligibility criteria as semantic queries, rather than any of mentioned languages. Presented patterns create an intermediate representation between free text and formal semantic query.

There are also studies which approach the problem of matching patients to trials without formalizing eligibility criteria, just using semantic annotations. The TrialX system [1] selects trials which correspond to user entered queries, which is annotated with UMLS concepts. The system retrieves those trials from the database, which are indexed with the same concept. This approach is limited to the UMLS ontology and does not allow specifying complex queries. In our approach, patterns provide context information for ontology concepts, and allow to specify various restrictions, leading to better precision of matching patients.

## 5 Conclusions and future work

### 5.1 Conclusions

In this paper we have investigated the possibility of capturing and formalizing the jargon of clinical trial eligibility criteria. We approached the problem by defining a set of 130 patterns that differ in the complexity level. Some patterns reflect sentence structure, others capture phrases corresponding to specific medical parameters. We defined a detailed classification that capture following dimensions: time independent and temporal status, specification type, medical content, variability and controllability, and subject. For each dimension we specified corresponding classes.

We evaluated the expressivity and coverage of the defined set of patterns. Our experiment with the concrete clinical trial demonstrated that patterns could be used to express a high majority of criteria (36/39). In order to check the coverage of patterns across various medical domains, we analyzed eligibility criteria from several thousands of clinical trials related to breast cancer, lung cancer and diabetes (3905, 2949, 5499). We used 342 regular expressions to identify the patterns in extracted eligibility criteria and were able to find at least one pattern in 71%, 69% and 54% of lines of eligibility criteria, respectively. We obtained a method for automatic classification of eligibility criteria according to fine-grained dimensions.

Our findings indicate that the language used for expressing eligibility criteria is regular enough to be captured to a big extent by the set of defined patterns. We conclude that their expressivity and coverage is sufficient to continue the research in the directions described in the next section.

## 5.2 Future work

We can apply the presented work for various applications.

Firstly, we will create a rich library of eligibility criteria, classified accordingly to all described dimensions and classes. As we explained before, identifying patterns in the text of condition allows classification only according to the dimensions and values associated with a pattern. Next step will cover extending it, by automatic analysis of criteria content. There are few ontology annotators which could be used for this purpose e.g. MetaMap [5]. Information about semantic types of identified ontology concepts for example from SNOMED CT will provide data needed to perform classification of medical content.

Additionally, we will incorporate domain knowledge to enable classification according to the variability and controllability dimension. Such knowledge is necessary to annotate e.g. tests with information whether its value can change in next exam, i.e. results of some blood tests are likely to change in contrary to the test indicating a gene mutation.

Created in this way library can facilitate the process of designing eligibility criteria. It will allow researchers to browse it and find criteria, defined for other clinical trials, related to their specific queries.

Secondly, we will approach the formalization of eligibility criteria to support e.g. matching patients for clinical trials. Based on annotations of criteria content with defined patterns and ontology concepts, we can start generating computable queries. It will be necessary to develop a method for combining identified patterns using logical operators or a grammar. We will also need to give consideration to ambiguous terms like 'high blood pressure', this issue was addressed in [7] in the context of formalizing medical decision rules. Another essential aspect is recognition whether a condition is inclusion or exclusion criterion, correct interpretation will influence the success of a matching algorithm.

Since semantic reasoning is expected to facilitate the process of criteria evaluation, we will formalize them in SPARQL, extending with SWRL rules if necessary. A computable representation of eligibility criteria will allow automatic determination of patients eligibility, and facilitate the recruitment process.

**Acknowledgments.** We would like to thank Rinke Hoekstra and Kathrin Dentler for the help with reviewing.

## References

1. Trialx. <http://trialx.com/>, June 2010.
2. Clinicaltrials. <http://clinicaltrials.gov/>, 2011.
3. Neo altto (neoadjuvant lapatinib and/or trastuzumab treatment optimisation) study. <http://clinicaltrials.gov/ct2/show/NCT00553358>, 2011.
4. Topoisomerase ii alpha gene amplification and protein overexpression predicting efficacy of epirubicin (top). <http://clinicaltrials.gov/ct2/show/NCT00162812>, 2011.

5. A. R. Aronson. Metamap: Mapping text to the umls metathesaurus. In *Proceedings AMIA Symposium*, 2001.
6. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.
7. S. Medlock, D. Opondo, S. Eslami, M. Askari, P. Wierenga, S. E. de Rooij, and A. Abu-Hanna. Lerm (logical elements rule method): A method for assessing and formalizing clinical rules for decision support. *International Journal of Medical Informatics*, 80(4):286 – 295, 2011.
8. L. Ohno-Machado, S. Wang, P. Mar, and A. Boxwala. Decision support for clinical trial eligibility determination in breast cancer. *Proceedings AMIA Symposium*, 1999.
9. J. Ross, S. W. Tu, S. Carini, and I. Sim. Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits on Translational Science Proceedings*, 2010.
10. M. Sordo, O. Ogunyemi, A. A. Boxwala, M.B.B.S., and R. A. Greenes. Software specifications for gello: An object-oriented query and expression language for clinical decision support. Technical report, Decision Systems Group, Brigham & Womens Hospital, Harvard Medical School, Boston, MA, 2003.
11. S. Tu, M. Peleg, S. Carini, D. Rubin, and I. Sim. Ergo: A templatebased expression language for encoding eligibility criteria. Technical report, 2009.
12. S. W. Tu, C. A. Kemper, N. M. Lane, R. W. Carlson, and M. A. Musen. A methodology for determining patients eligibility for clinical trials. *Methods of Information in Medicine*, 1993.
13. S. Wang, L. Ohno-Machado, P. Mar, A. Boxwala, and R. Greenes. Enhancing arden syntax for clinical trial eligibility criteria. *Proceedings AMIA Symposium*, 1999.
14. C. Weng, S. W. Tu, I. Sim, and R. Richesson. Formal representations of eligibility criteria: A literature review. *Journal of Biomedical Informatics*, 2009.