Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach

Krystyna Milian[1,2], Anca Bucur[2], Annette ten Teije[1]
1 Vrije University, Amsterdam, the Netherlands
2 Philips Research, Eindhoven, the Netherlands

***Abstract*—The semi-automatic evaluation of eligibility criteria can facilitate the recruitment for clinical trials, timely completion of studies and generation of clinical evidence about new approaches to treatment, prevention and diagnosis. Because eligibility criteria are represented as free text, automatically extracting their meaning and evaluating them for a particular patient is challenging. This paper presents our approach to the problem of automatic interpretation of criteria meaning. It is based on detecting in text semantic entities (diseases, treatment, measurements etc.) using ontology annotators and semantic taggers, and detecting predefined patterns providing the contextual information in which these entities occur. Evaluation of the approach is the main subject of the paper. It covers several aspects: precision and recall of the pattern detection algorithm and the assessment of the implications of using the identified patterns to find potential candidates. It was performed manually using a subset of patterns and randomly selected 33 trials from ClinicalTrials.gov. The average precision and recall of pattern detection algorithm calculated for selected patterns is 0.9 and 0.91, meaning that in most cases using the patterns can lead to correct interpretation of criteria and can support patient recruitment.**

***Keywords*-formalization of eligibility criteria, supporting clinical trial recruitment, clinical trials**

I. Introduction

Clinical trials, if successfully completed, provide evidence about new approaches to treatment, prevention and diagnostic methods. A trial can only be completed when sufficient participants (to achieve statistically-sound results) are enrolled. Each clinical trial specifies eligibility criteria, defining the population that can be recruited. The criteria concern, among others, demographic information, the history of the patient treatment and current health conditions. The process of identifying eligible candidates is time- and effort-consuming, requiring health care providers to be aware of all relevant currently running trials and to identify candidates that suit the trials.

We aim to support the task of recruitment by implementing a method for the formalization of eligibility criteria, enabling their semi-automatic evaluation. The method relies on: 1) detection of syntactic patterns based on regular expressions, providing the contextual information, 2) detection of semantic entities, 3) evaluation of criteria composed of the two types of entities based on patient data. For example, consider the eligibility criterion ”No prior malignancy except for nonmelanoma skin cancer”. First, we detect the pattern ”No prior () except for ()”, and second, the concepts ”malignancy” and ”nonmelanoma skin cancer”. To evaluate criteria, the patterns can be linked to pre-defined incomplete queries, which after filling with the semantic entities identified can be executed to verify patient eligibility.

In this paper, we first describe the method, then evaluate the approach to eligibility criteria formalization, with the focus on pattern detection. The feasibility is influenced by the number of patterns, the algorithm of patterns detection, and the restrictiveness and the variety of synonym forms covered by the regular expressions.

The paper is organized as follows. In the next section we describe the related work. Further, in section III introduce the method for formalization of eligibility criteria. In section IV we describe the evaluation experiment and the sets of selected patterns and clinical trial eligibility criteria. Next, in section V we analyze the obtained results, and present our conclusions in section VI.

II. Related work

The analysis of clinical trial eligibility criteria and their formalization has been already addressed in the literature from other perspectives. Tu et al [1] provides a detailed and informative analysis of eligibility criteria, based on randomly chosen 1000 cases from ClinicalTrials.gov. Criteria were categorized along several axes: complexity of conditions, high level clinical content and semantic and clinical patterns. Our paper describes one step forward, 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. Various representation languages that can be applied for expressing eligibility criteria (e.g. Arden syntax [2], Ergo [3], Gello[4]) are presented in the overview study [5]. However no complete solution has been presented to the problem of automatic transforming free text of criteria to any of this representations. In [6] authors describe the semi-automatic approach, focused on ERGO language, which provides insights into challenges of the task. However, since

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

it requires some manual effort it cannot be directly applied for another set of trials.

III. Formalization Method

The formalization of eligibility criteria supports their automatic evaluation, providing information about whether a patient satisfies all necessary conditions to be enrolled in a clinical trial. After analyzing eligibility criteria from clinical trials published at ClinicalTrials.gov and observing vast similarity among them, we decided to capture the most typical expressions by patterns (see Figure 1). We defined a set of 165 patterns which cover conditions related to demographic information (e.g. "Age over ()"), disease characteristics (e.g. "T () stage" , "allergy to ()") and prior and concurrent therapies (e.g. "At least () since prior () therapy, "No concurrent () except for ()"). The patterns were defined through an iterative process of assessing and improving the expressivity of the entire set. They are described in detail in our previous work [7].

Formalizing eligibility criteria consists of several steps, depicted in Figure 2. First, we recognize the general meaning of a criterion by detecting the syntactic patterns providing the contextual information about the semantic entities mentioned in the criterion. Next, we identify these semantic entities, which can be instantiated by diseases, treatments, lab measurement, value or temporal constraints. To detect them we apply MetaMap [8], the ontology annotator, and GATE [9], the open source framework for text processing, providing semantic taggers for measurements and numbers. Since these are state of the art tools, their annotation performance is not evaluated here.

The algorithm for pattern detection is based on regular expressions. In total we defined 468 regular expressions corresponding to the 165 patterns. The algorithm processes eligibility criteria delimited using GATE sentence splitter. Each sentence can correspond to more than one pattern. From the set of patterns identified in the sentence, the algorithm chooses only those that cover the longest phrases, and ignores patterns capturing segments subsumed by others. For example in the sentence 'No other concurrent hormonal therapy, including steroids', it identifies two patterns 'no concurrent ()' and 'no concurrent () including ()', from which it selects only the latter because it more closely reflects the content and meaning of the criterion. In addition, it recursively searches for nested patterns. In the sentence: 'No history of other malignant neoplasms except for curatively treated nonmelanoma skin cancer or surgically cured carcinoma of the cervix in situ' the algorithm first identifies the pattern 'no prior () except for ()' and, second, the one nested in the second parameter 'recovery from ()'.

The results of applying the algorithm are the subject of our evaluation as described in the next sections.

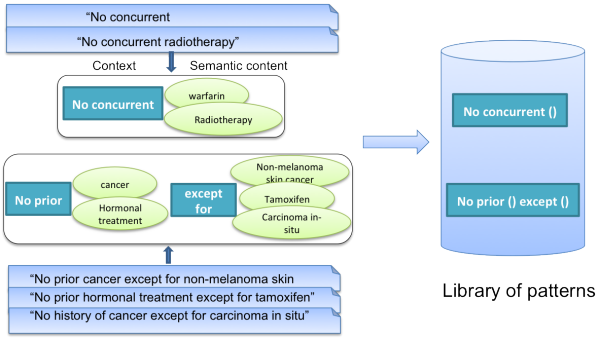

Figure 1. Patterns of eligibility criteria.

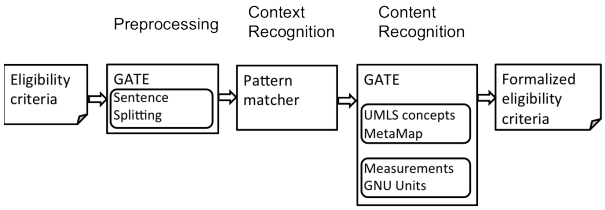

Figure 2. Pipeline of formalization steps.

IV. The Evaluation Experiment

Due to the significant manual effort required for the evaluation, we decided to focus on a selected subset of clinical trials and patterns, described next.

A. Subset of clinical trials

We tested our method with clinical trials criteria from the large public repository ClinicalTrials.gov[1], using trials that specify breast cancer as a study condition. Our main focus of interest lies here because of our clinical collaborators, whose expertise will be crucial in further steps of the research. From the available clinical trials we randomly selected 1% (33) of cases.

B. Subset of patterns

We have selected 20 patterns out of the 165 for the evaluation: the 10 most frequent and the 10 most complex. The selection of most frequent patterns was based on the number of their occurrences in the eligibility criteria in the total corpus of over 3 thousand breast cancer clinical trials (described below). The selection of the most complex patterns was based on the number of pattern variables (i.e. pattern "no ()" has 1 variable, pattern "no () within () except for ()" 3) and their availability in the selected subset of clinical trials. We distinguished the most complex patterns to verify whether the performance of the pattern detection algorithm depends on the complexity of the patterns. As a result, the evaluation covered the following patterns:

[1]http://clinicaltrials.gov/

- 10 most frequent:
  History of (), () greater than or equal (), Female, No (), Required (), History of () within (), () normal, () negative, No history of (), Prior () for ()
- 10 most complex:
  Value () in range() - (), At least () since prior (), History of () within () prior to (), No prior () for (), No history of () within (), () allowed if (), If () then (), No concurrent () for (), No () within () except for (), Negative () within () prior to ()

C. Evaluation of the pattern detection algorithm

We evaluated the pattern detection algorithm in terms of precision and recall and analyzed the results of the annotation of sentences of the selected set of eligibility criteria. We manually verified whether the patterns detected by the algorithm were indeed the best match from our set, and whether the algorithm has found all of them.

D. Evaluation of implications of using selected patterns

The next part of the evaluation of our patterns and regular expressions is carried out from the perspective of generating relevant queries, enabling the identification of suitable candidates for clinical trials. In order to quantify the consequences of using the selected patterns for a given criterion to generate a query we introduce a scoring formula. It evaluates to what extent a pattern covers the meaning of the sentence, by indicating the correct classification:

- Temporal status (prior, current, planned) (TS)
- Time independent status TIS (present, absent, allowed)
- Specification type (exception, inclusion, purpose of a treatment, temporal or confirmation constraint, a limit for the number of occurrences). These constraints are further called a weakening (C_w) or strengthening condition (C_s), depending on the context.

The dimensions describe criteria content from various perspectives. A criterion can refer to patient history or current situation (TIS), can require or allow a presence/absence (TIS) of some medical parameter. In addition, a criterion can define various types of weakening or strengthening constraints (C_w), (C_s).

We calculate the score of a pattern P for a sentence S taking into account the above components, according to the formula:

$$Score(P(S)) = \frac{TS + TIS + C_w + 0.5C_s}{n}, \qquad (1)$$

where TS, TIS, C_w and C_s are a fraction of correctly identified elements of dimensions; n is the weighted number of specified elements in the formula. The score takes values between 0 and 1. A pattern does not need to specify all components, but some patterns may cover several dimensions and several values in a dimension. For instance a pattern ”No () prior ()” only has temporal status (TS = prior) and time independent status (TIS = absence), while pattern ”No () prior () within () except for ()” has two additional weakening conditions (a temporal constraint and an exception).

We consider all components equally important with the exception of strengthening conditions. The latter has the lowest weight because from the perspective of our application, suggesting some irrelevant candidates is less serious than missing eligible patients, as the final decision about the recruitment is made by the physician.

For each sentence we calculate the score of the pattern selected by the algorithm and compare it with the classification of the sentence implied by the manually selected best matching pattern. If the selected pattern is equal to the best match it receives the maximal score 1, otherwise a value between 0 and 1 is computed based on the weighing formula. This measure tells how well the selected pattern reflects the intended meaning of a criterion, in contrary to evaluation of precision, which only evaluates whether the selected pattern is the best match.

V. Analysis of results

Our approach to pattern detection is based on regular expressions. Restrictiveness and the variety of synonym forms covered by regular expressions influence the rate of correctly interpreted criteria. We evaluated the algorithm in terms of precision and recall and described previously measure, the results are given in Table I.

The average recall is higher then precision, which indicates that the restrictiveness of regular expression can be improved, especially concerning patterns: ”() normal”, ”Prior () for ()”, ”History of ()”. Lower recall of patterns: ”if () then ()”, ” () allowed if ()”, ”No history of () within ()” indicates that more synonym forms can be added.

The average precision for the group of most complex patterns is significantly higher then for the group of most frequent ones (0.98 vs. 0.83), while recall is lower (0.86 vs. 0.99). This finding confirms our intuition that the algorithm performs better in the correct identification of complex phrases. It should be noticed that the most frequent patterns account for almost 40 % of all defined patterns, therefore the focus should be placed on preventing errors related to them, unless we develop an application focused on particular kinds of eligibility criteria.

The score of annotation, indicates the average extent to which a pattern chosen by the algorithm covers the details of the best matching pattern corresponding to the criteria. Detailed inspection of results, the detected elements of criterion showed that in some cases even using suboptimal pattern can lead to correct filtering of patients. However, the opposite also can happen. The majority of mistakes is caused by failing to recognize the broader context, strengthening conditions, time independent status (TIS) and than weakening conditions. An example of misinterpreted context is

Table I
EVALUATION

Pattern	Precision	Recall	Score
AVG	0. 9	0.91	0.92
AVG Most frequent	0.83	0.96	0.86
History of ()	0.76	1	0.83
() greater than or equal ()	0.82	1	-
No ()	0.94	0.95	0.97
Required ()	0.87	0.95	0.92
History of () within ()	0.88	0.96	0.94
Female	1	1	1
Normal ()	0.5	1	0.5
No history of ()	0.8	0.86	0.86
() negative	1	1	1
Prior () for ()	0.7	1	0.76
AVG Most complex	0.98	0.85	0.99
Value in range() - ()	0.9	1	0.9
At least () since prior ()	0.92	0.92	0.97
No prior () for ()	1	0.83	1
History of () within () prior to ()	1	1	1
No history of () within ()	1	0.8	1
() allowed if ()	1	0.64	1
If () then ()	1	0.33	1
No concurrent () for ()	1	1	1
No () within () except for ()	1	1	1
Negative () within () prior to ()	1	1	1

recognition of a pattern: "History of ()" in a sentence "Prior systemic therapy in the adjuvant setting is not considered a regimen." which has only an explanatory role.

The focus needs to be put on preventing errors connected to misinterpreting the context and TIS, which would deteriorate both precision and recall of finding eligible candidates.

The evaluation let us analyze the strong and weak points of patterns detection. We conclude that the precision and recall of our algorithm is sufficient to facilitate the process of formalization of eligibility criteria.

VI. CONCLUSION

In this paper we presented the evaluation of our approach to the formalization of eligibility criteria which aims to support semi-automatic verification of patient eligibility for clinical trials. The method is based on the detection of predefined patterns, which provide the contextual information, in which eligibility criteria mention findings, treatments or lab results. The set contains 165 items. Correct identification enables distinguishing cases when a mentioned treatment disqualifies a patient, or, on the contrary, it is a necessary entry condition, or is allowed only under specific circumstances. The patterns can be further linked to incomplete queries, and their execution will assess patient eligibility.

The experiment performed in this paper aimed to evaluate our approach from the perspective of the pattern detection algorithm. The correct identification of patterns is crucial to accurate interpretation of the criteria meaning and for effective fulfillment of the task of supporting recruitment. With our approach based on regular expressions, we obtained an average precision and recall calculated for the selected 20 patterns equal to 0.90 and 0.91respectively. The results for the particular patterns vary from 1 (e.g. for pattern "no history of () within ()") to 0.5 in case of pattern "() normal". The information about strong and especially weak points should be taken into account in the next step of generating queries. However, the analysis of the impact of using the patterns proposed by the algorithm shows that even using suboptimal patterns can lead to correctly filtering potential candidates.

We conclude that the algorithm of pattern detection has sufficient precision and recall to support, in most cases, the generation of correct corresponding queries and to determine patient eligibility. The broad range of defined patterns allows automatic interpretation of even complex eligibility criteria. In future work, we will investigate other NLP techniques, e.g. parse trees or grammars to improve the obtained results of pattern detection. The planned experiments with patient data in real clinical context will ultimately verify the feasibility of the presented method.

REFERENCES

[1] 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.

[2] S. Wang, L. Ohno-Machado, P. Mar, A. Boxwala, and R. Greenes, "Enhancing arden syntax for clinical trial eligibility criteria," *Proceedings AMIA Symposium*, 1999.

[3] S. Tu, M. Peleg, S. Carini, D. Rubin, and I. Sim, "Ergo: A templatebased expression language for encoding eligibility criteria," Tech. Rep., 2009.

[4] M. Sordo, O. Ogunyemi, A. A. Boxwala, and R. A. Greenes, "Gello: An object-oriented query and expression language for clinical decision support," *Proceedings AMIA Symposium*, vol. 2003, no. 5, p. 1012.

[5] C. Weng, S. W. Tu, I. Sim, and R. Richesson, "Formal representations of eligibility criteria: A literature review," *Journal of Biomedical Informatics*, 2009.

[6] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim, "A practical method for transforming free-text eligibility criteria into computable criteria," *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 239 – 250, 2011.

[7] K. Milian, A. ten Teije, A. Bucur, and F. van Harmelen, "Patterns of clinical trial eligibility criteria," in *Proceedings of the AIME'11 workshop on Knowledge Representation for Healthcare (KR4HC11)*, ser. lecture notes AI. Springer, 2011.

[8] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," in *Proceedings AMIA Symposium*, 2001.

[9] H. Cunningham, D. Maynard, K. Bontcheva, and et al., *Text Processing with GATE (Version 6)*, 2011. [Online]. Available: http://tinyurl.com/gatebook