HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2017 December 18.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2016 December ; 2016: 717–722. doi:10.1109/BIBM.2016.7822607

Analysis of Temporal Constraints in Qualitative Eligibility Criteria of Cancer Clinical Studies

Zhe He,
School of Information, Florida State University, Tallahassee, FL, USA

Zhiwei Chen, and
Department of Computer Science, Florida State University, Tallahassee, FL, USA

Jiang Bian
Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, USA

Abstract

Clinical studies, especially randomized controlled trials, generate gold-standard medical evidence. However, the lack of population representativeness of clinical studies has hampered their generalizability to the real-world population. Overly restrictive qualitative criteria are often applied to exclude patients. In this work, we develop a lexical-pattern-based tool to structure qualitative eligibility criteria with temporal constraints, with which we analyzed over 10,800 cancer clinical studies. Our results showed that restrictive temporal constraints are often applied on qualitative criteria in cancer studies, limiting the generalizability of their results.

Keywords

clinical trial; generalizability; health informatics

I. Introduction

Clinical studies are conducted for testing the efficacy and safety of a treatment (e.g., medication, device, and procedure) for one or more medical conditions. Even though clinical trials have been widely-accepted as a gold-standard of modern medical research [1], many of them failed to balance the internal validity and external validity, thereby limiting the applicability of the trial results to the real-world population. The lack of population representativeness is one of the major issues that lead to poor generalizability [2]. A study of the Southwest Oncology Group (SWOG) reported that although about 60% of new cases of cancer occur among older adults, they only comprise 25% of participants in cancer clinical trials [3]. In developed countries such as the United States, more than 40% of all newly diagnosed breast cancer patients are over 65 years old [3]. However, older adults are systematically underrepresented in clinical studies across major types of cancer [3] as well as other chronic conditions such as dementia [4], and diabetes [5]. This practice may significantly hamper the generalizability of the efficacy and safety findings of clinical studies to broader patient populations [6]. As a consequence, some drugs were later withdrawn from the market after serious adverse drug reactions (e.g., death, organ damage, and toxicity) were observed when they were administered to a broad patient population [7].

Moreover, there is a lack of evidence for assessing quality of care in elderly patients with multiple comorbidities as clinical practice guidelines commonly only focus on a single disease [8]. To provide the evidence for assessing the efficacy and safety of a medication that will be used in elders with multiple comorbidities, these patients should be appropriately represented in clinical trials [9].

Cancer clinical studies usually do not explicitly exclude older adults. Only 794 (4.4%) of all the 17,858 interventional clinical studies in ClinicalTrials.gov on lung neoplasms between 2000 and 2014 explicitly excluded patients >= 65 years old. Our recent study also shows that the quantitative eligibility criteria in cancer studies did not pose notable restrictions in the target population [10]. However, the qualitative eligibility criteria with moderate or strict restrictions were frequently observed in these trials. In particular, patients with clinically evident congestive heart failure and/or myocardial infarction were often excluded by colorectal cancer treatment trials. In this study, we present the prototype of a parsing tool called q**U**alitative e**L**igibili**T**y c**R**iteri**A** parser (ULTRA), which leverages pre-defined lexical patterns to structure qualitative eligibility criteria with temporal constraints. With this tool, we analyze the temporal constraints of exclusion criteria in cancer clinical studies to understanding how cancer clinical studies exclude patients with restrictive exclusion criteria. Based on the tool, we can answer questions such as "*what temporal constraints do lung cancer clinical studies use to exclude patients with myocardial infarction?"*

II. Method

A. General Analytical Pipeline

The goal of this work is to analyze the restrictions of qualitative eligibility criteria of cancer studies with respect to temporal constraints. In a clincial study summary, free-text eligiblity criteria define the characteristics of the eligible patients, i.e., the target population of the study. In order to analyze the pattern of eligibility criteria at large scale, we need to formalize eligibility criteria in a computer-recognizable format. We adapted a set of lexical patterns identified by Milian et al. [11] to recognize free-text eligibility criteria with temporal constraints, such as "*no history of myocardial infarction within the last five years*." Using these patterns, we can easily identify the temporal constraint and the candidate text region for named entity recognition (NER). Further, we are only interested in the concepts of the Unified Medical Language System (UMLS) about *disorders*, *procedures*, and *clinical drugs* [12]. Using the UMLS, we can normalize different terms with the same meaning, e.g., *myocardial infarction* and *heart attack*, into the same concept. Due to the complexity of eligibility criteria text, state-of-the-art methods for representing eligibility criteria with compound semantics have to be manually processed before automated processing [13]. In this work, we used dependency parsing and syntax tree to decompose complex criteria and identify noun phrases for NER. Figure 1 shows the overall analytical pipeline of this study, which consists of four major components: (1) eligibility criteria data collection, (2) basic nature language processing, (3) pattern matching and NER based on dependency parsing and syntactic parsing results, and (4) temporal constraint analysis. We will discuss the details of the pipeline in following sections.

B. Data Collection

ClinicalTrials.gov is a clinical study and results registry created and maintained by the U.S. National Library of Medicine (NLM). Previously, we transformed the clinical studies in ClinicalTrials.gov into a relational database called COMPACT [14]. From COMPACT, we retrieved all the free-text eligibility criteria for clinical studies on four major types of cancer, namely lung cancer (n = 2,880), breast cancer (n = 4,267), prostate cancer (n = 2,123), and colorectal cancer (n = 1,578) for studies with a start date between January 2000 and December 2014.

C. Eligilbity Criteria Parsing

After collecting the free-text eligibility criteria of cancer studies from the COMPACT database, we parsed the eligibility criteria for each study as follows. First, we identified the subsections of eligibility criteria. Normally, the eligibility criteria section is divided into '*Inclusion Criteria*' and '*Exclusion Criteria*' subsections, characterizing patients who are or are not eligible for the study. However, in the study summaries on ClinicalTrials.gov, subsection names of the eligiblity criteria may also be '*Patient Characteristics*,' '*Disease Characteristics*,' and '*Prior Concurrent Therapy*.' After scrutinizing all the possble section names, we considered all the sections except '*Exclusion Criteria*' as inclusion criteria. The eligiblity criteria section in the XML-format summaries uses hashtags (i.e., "#") to separate each individual eligiblity criterion. Based on the hashtags, we split the criteria subsection into a set of individual criteria. Nevertheless, an criterion may include more than one sentence. For each criterion, we employed the Stanford CoreNLP [15] tool to perform tokenization, sentence spliting, Part-of-Speech (POS) tagging, duration recognition, syntax parsing, and dependency parsing. With pre-defined lexical rules, we identified qualitative criteria with temporal constraints. We then used the Stanford CoreNLP tool to extract the temporal constraints and candidate text region for concept recognition through NER. In the NER task, we also used dependency parsing and synatic parsing results to identify UMLS concepts from eligibility criteria of varying complexity. The details of pattern matching and UMLS concept recognition are explained as follows:

1) Recognizing Criteria with Temporal Constraints—We extracted contextual information (i.e., temporal constraints in the criteria) from eligibility criteria with a list of pre-defined lexical patterns. Milian et al. previously used a set of lexical patterns to structure eligibility criteria [11]. We selected a subset of the lexical patterns that contain temporal constraints. Table V lists the lexical patterns with more than 100 criteria sentences. After matching a criterion with the pattern, we then formalized the temporal phrases (e.g., '*within the last five years*') in TIMEX3 format [16]. We used a Java library called time4j to normalize the units of the temporal constraints to 'month,' where the granularity is sufficient for our purpose.

TokensRegex in Stanford CoreNLP is a cascading regular expression language that can structure a sequence of tokens according to predefined lexical patterns in a cascading manner [17]. TokensRegex, which describes text as a sequence of tokens (e.g., words and punctuations), can apply writing patterns over the tokens rather than matching patterns merely at the character level. With TokenRegex, we were able to easily recognize and

structure the temporal information. Table I gives an example illustrating how we used a lexical rule in TokensRegex to structure a criterion with a temporal constraint. We identified UMLS terms from the words labeled as *TEXT_NER*. In addition, the *stage* of the pattern specifies the priority of pattern, i.e., which pattern should be matched first. Figure 2 gives an example sentence that matches this pattern: *Any history of hemoptysis within the past 12 months.* In this example, the term *history* was labeled as KEY (of the pattern). The term *hemoptysis* was labeled as TEXT_NER. The phrase *12 months* was labeled as DURATION.

2) Named Entity Recognition (NER)—In our previous work [18], we employed a well-known technique called *n-gram* to extract terms from a word sentence for UMLS concept matching. Specifically, after sentence detection, tokenization, and POS tagging, we applied a fuzzy matching method to match an n-gram to a UMLS term. The fuzzy matching method would check if a sequence of words can match a UMLS term when the case, stop words and minor lexical variations are ignored. However, eligibility criteria often have complex syntax. For example, the criterion '*History of drug, alcohol, or substance abuse within the 6 months before screenin*' contains three major criteria concepts: *drug abuse*, *alcohol abuse*, and *substance abuse*. It is thus not feasible to use n-grams, which consist of a contiguous sequence of n words, to identify concpets in criteria with conjunctions. Therefore, in this study, we used the Stanford CoreNLP toolkit to perform dependency parsing and syntax tree analysis to decompose complex criteria with conjunction, aiming at improving the accuracy and recall of the NER task. Note that we also used the dependency parsing results to identify nouns or noun phrases from simple criteria. Since most qualitative criteria with temporal constraints in our dataset are relevant to disorders, procedures, and drugs, we only included UMLS concepts of the semantic types shown in Table II. The details on dependency parsing and syntax tree analysis are explained below.

a) Dependency Parsing: Dependency parsing provides a representation of grammatical relations between words in a sentence. Therefore, it allows us to identify a meaningful term with adjacent or non-adjacent words in a sentence. For example, in the sentence '*Patients who received chemotherapy, steroid or biologic treatment within 4 weeks prior to enrollment*', the term *steroid* and *treatment* are not adjacent, but they are connected by the syntactic relationship 'conjunction' in the dependency parsing. As such, we can recognize the UMLS concept *steroid treatment* (C0149783) in this sentence. We only considered the relations in the dependency parsing results that can be used to construct noun phrases, including *amod* (adjectival modifier), *acomp* (adjectival complement), *vmod* (reduced, non-finite verbal modifier), *nn* (noun compound modifier), *compound* (general compound modifier) and *nmod* (nominal modifier). With 'conj_or' and 'conj_and' relations, we can automatically construct meaningful terms based on frequently occurring syntactic structures, such as A (adjective), B (adjective) or C (adjective) N (noun), by combining AN, BN, CN to extract possible UMLS terms. We extracted words associated by the above relations and labeled them as TEXT_NER. Figure 3 shows an example of dependency parsing for the sentence '*No history of drug, alcohol, or substance abuse within the 6 months before screening.*' Based on the dependency parsing results, we decomposed the sentence and matched the terms *substance abuse, alcohol abuse, and drug abuse* to the UMLS concepts C0740858, C0085762, and C0013146, respectively.

We also detected negation in the sentence with dependency parsing by extracting the negation relation before the first word of our pre-defined lexical patterns. The negation relation is represented as 'neg' in Figure 3, i.e., the word '*No*' has a negation relation with word '*history*,' the first word of the pattern HISTORY_WITHIN.

b) Syntax Tree: The output of syntax parsing is a syntax tree, which is a tree representation of the syntactic structure of a sentence. We traversed the syntax tree to find the longest noun phrase that can match a UMLS term. We only extracted UMLS terms from a word sequence that satisfies the following three criteria: (a) it is labeled as TEXT_NER in the pattern matching process; (b) it is a single-word noun or a noun phrase with no more than five words, including *NN* (single noun), *NNS* (plural noun), *NNP* (proper single noun), *NNPS* (proper plural noun) and *NP* (noun phrase); (c) none of the words in the sequence match any UMLS terms in the dependency parsing phase. If a longer noun phrase matches a UMLS term, its subsequence will not be matched again. Figure 4 shows the syntax tree for the sample sentence '*Patients must not have a history of bleeding diathesis or have used anticoagulant therapy within 10 days of study entry*,' which follows the Penn Treebank Project annotations [19].

Using the pattern matching, we labeled the word sequence '*bleeding diathesis or have used anticoagulant therapy*' as TEXT_NER. In this word sequence, '*bleeding diathesis*' and '*anticoagulant therapy*' are noun phrases. '*Anticoagulant therapy*' has been already recognized as a UMLS concept (C0150457) with the dependency parsing, thus was not further processed. '*Bleeding diathesis*' was not recognized as a noun phrase by the dependency parsing because '*bleeding*' was erroneously tagged as a verb. Using the syntax tree, we identified it as a UMLS concept (C0005779).

III. Results

A. General Descriptive Statistics

Table III shows the number of clinical studies for each type of cancer included in this study. The majority of the studies are interventional studies (i.e., clinical trials).

Table IV summarizes the characteristics of eligibility criteria of cancer studies. The majority of the eligibility criteria sentences have no negation, regardless of the subsection in which they appear.

B. Evaluation of the Pre-defined Lexical Patterns

We performed a preliminary evaluation to assess the quality of the lexical patterns used in ULTRA. We randomly selected 500 eligibility criteria, and manually reviewed the parsing results. The task to manually identify qualitative criteria with temporal constraints is trivial and thus does not require cancer domain knowledge. We define that the qualitative criteria with temporal constraints are positives; others are negatives in our manually annotated corpus. Qualitative criteria with temporal constraints that can be correctly matched to one of our pre-defined lexical patterns are true positives (TP = 81). Qualitative criteria with temporal constraints that cannot be matched to any of our pre-defined lexical patterns are false negatives (FN = 20). These criteria are mostly with compound semantics with joint use

of parentheses and conjunctions. Qualitative criteria without temporal constraints and quantitative criteria such as 'ANC >= 1500 per uL' that did not match any pattern are true negatives (TN = 394), otherwise false positives (FP = 5). FP criteria are mostly about age. The recall (i.e., TP/(TP+FN)) is 80.2%; the precision (i.e., TP/(TP+FP)) is 94.2%; the F-score is 86.6%.

C. Frequency of Lexical Patterns in Cancer Studies

Table V lists all the lexical patterns used in this study to recognize and structure qualitative criteria with temporal constraints, their frequency in inclusion and exclusion criteria, and a sample criterion sentence for each pattern. The keywords of the patterns are in bold typeset in the sample sentences. The temporal constraints are in italic typeset in the sample sentences. The most frequent pattern is "WITHIN" followed by "MORETHAN," "WITHIN_PT." It is reasonable that the majority of the qualitative criteria do not follow any patterns (n = 162,132), because the majority of eligibility criteria do not have temporal constraints.

D. Semantic Types of the UMLS Concepts in the Criteria

Figure 5 shows the distribution of semantic types of the UMLS concepts identified from eligibility criteria of each type of cancer studies. The numbers in the bar graph represent the percentage of criteria in the studies of the corresponding cancer type. It is clear that most criteria are about diseases or syndromes, therapeutic or preventive procedures, diagnostic procedures, and laboratory procedures. It is worth noting that breast cancer studies have more criteria about therapeutic or preventive procedures than other types, whereas prostate cancer studies have much fewer such criteria. Meanwhile, prostate cancer studies have more criteria about diagnostic procedures than other types. All the semantic types have at least 10 occurrences.

E. Temporal Constraint Distribution of Frequent Concepts

Figure 6 shows the temporal constraint distribution for frequent qualitative eligibility criteria. We included those criteria with negation in the inclusion criteria as well as those without negation in the exclusion criteria. The x-axis represents the number of months used as the temporal constraint in the eligibility criteria. Note that "0" represents the temporal constraint of less than one month. The numbers in the cell represent the number of occurrences of the criterion with the corresponding temporal constraint. The criteria are ranked by their frequency from bottom up. It is clear that the criteria about therapeutic procedures with temporal constraints are more frequently used than those about disorders. In addition, temporal constraints applied on exclusion criteria about therapeutic procedures are systematically shorter (<= two months) than those applied on exclusion criteria about disorders (>= six months).

Figure 7 illustrates the temporal constraint distribution for the criterion on myocardial infarction in four types of cancer. The x-axis represents the number of months in the temporal constraint. The y-axis represents the number of studies. We included those criteria with negation in the inclusion criteria as well as those without negation in the exclusion criteria. The majority of the criteria about congestive heart failure do not have temporal

constraints. For exclusion criteria about myocardial infarction and cerebrovascular accident, ‘six months’ is the mostly used temporal constraints, followed by ’12 months,’ both of which implying moderate restriction of the target population [20].

IV. Discussion and Conclusion

In this work, we developed a lexical-rule-based natural language processing tool called ULTRA to structure eligibility criteria with temporal constraints. We pre-selected 21 lexical patterns from a published paper on enhancing the reuse of structured eligibility criteria [11]. To facilitate automated decomposition of complex criteria with conjunction, we leveraged the dependency tree and syntax tree generated by the Stanford CoreNLP toolkit. Using ULTRA, we parsed free-text eligibility criteria of about 10,800 clinical studies on the four most prevalent types of cancer, including lung cancer, breast cancer, prostate cancer, and colorectal cancer. We found that frequently used exclusion criteria about procedures, e.g., major surgery, radiation therapy, and chemotherapy, are often applied with a short temporal constraint (e.g., one month). Meanwhile, frequently used qualitative criteria about disorders, e.g., myocardial infarction, stroke, and unstable angina, are often applied with longer temporal constraint such as six months, indicating moderate restrictions [20].

Cancer clinical studies often suffer from lack of participants and delayed enrollment [20]. Informatics has the potential to optimize clinical research participant selection during the design phase of a new study. As eligibility criteria represent the study population being sought after, computational techniques based on natural language processing and biomedical ontologies can be leveraged to identify the issues of population representativeness from the eligibility criteria during the design phase of a new study. By relaxing unnecessarily restrictive eligibility criteria, the study can be accessible to patients whose health conditions may not directly interact with the inventions tested in the studies and are thus safe to participate.

This study has a few limitations. First, we ignored sentences with more than 500 characters after sentence split, because some criteria were copied from the study protocol into a study summary without punctuations. Second, it is difficult to exhaustively list all the lexical patterns due to the freedom of natural language. As such, some overly simplified criterion such as “*1 year from pregnancy, lactation or chemotherapy*” were not captured by our patterns. Nevertheless, our preliminary evaluation of the lexical patterns showed satisfactory coverage of the qualitative criteria with temporal constraints. Third, we may have missed meaningful criteria concepts that cannot be covered by the UMLS. Nevertheless, we used fuzzy matching to allow a term to match a UMLS concept with minor variations. In the future, we will improve the ULTRA tool to identify meaningful criteria concepts that are not covered by the UMLS. In future studies, we will perform a comprehensive evaluation to evaluate the accuracy of named entity recognition, temporal constraints extraction, and negation detection. We will also enhance the named entity recognition with frequent syntax tree mining.

Acknowledgments

This study was supported by the Institute for Successful Longevity under Planning Grant and the Council on Research and Creativity of the Florida State University under First Year Assistant Professor Grant (PI: He). This work was also partially supported by the US National Center for Advancing Translational Sciences under the Clinical and Translational Science Award UL1TR001427 (PIs: Nelson & Shenkman). The content is solely the responsibility of the authors and does not represent the official view of the National Institutes of Health.

References

1. Reichert JM. Trends in development and approval times for new therapeutics in the United States. Nat Rev Drug Discov. Sep.2003 2:695–702. [PubMed: 12951576]
2. From the NIH Director: The Importance of Clinical Trials. Apr 9. 2014 Available: http://www.nlm.nih.gov/medlineplus/magazine/issues/summer11/articles/summer11pg2-3.html
3. Hutchins LF, Unger JM, Crowley JJ, Coltman CA Jr, Albain KS. Underrepresentation of patients 65 years of age or older in cancer-treatment trials. N Engl J Med. Dec 30.1999 341:2061–7. [PubMed: 10615079]
4. Schoenmaker N, Van Gool WA. The age gap between patients in clinical studies and in the general population: a pitfall for dementia research. Lancet Neurol. Oct.2004 3:627–30. [PubMed: 15380160]
5. He Z, Wang S, Bornanian E, Weng C. Assessing the Population Representativeness of Type 2 Diabetes Trials by Combining Public Data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform. 2015; 2015:569–73.
6. Beers E, Moerkerken DC, Leufkens HG, Egberts TC, Jansen PA. Participation of older people in preauthorization trials of recently approved medicines. J Am Geriatr Soc. Oct.2014 62:1883–90. [PubMed: 25283151]
7. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions. Arch Intern Med. Jun 27.2005 165:1363–9. [PubMed: 15983284]
8. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. Aug 10.2005 294:716–24. [PubMed: 16091574]
9. Shenoy P, Harugeri A. Elderly patients' participation in clinical trials. Perspect Clin Res. Oct-Dec; 2015 6:184–9. [PubMed: 26623388]
10. He, Z., Chen, Z., George, TJ., Lipori, G., Bian, J. Assessing the Population Representativeness of Colorectal Cancer Treatment Clinical Trials. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Orlando, FL. 2016. p. 2970-2973.
11. Milian K, Hoekstra R, Bucur A, Ten Teije A, van Harmelen F, Paulissen J. Enhancing reuse of structured eligibility criteria and supporting their relaxation. J Biomed Inform. 2015; 56:205–19. [PubMed: 26015310]
12. NLM. Semantic Group of the Unified Medical Language System. Jun 1. 2016 Available: https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt
13. Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, et al. A practical method for transforming free-text eligibility criteria into computable criteria. J Biomed Inform. Apr.2011 44:239–50. [PubMed: 20851207]
14. He Z, Carini S, Hao T, Sim I, Weng C. A Method for Analyzing Commonalities in Clinical Trial Target Populations. AMIA Annual Symp Proc. 2014:1777–86.
15. Manning CD, Surdeana M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014:55–60.
16. TimeML. Guidelines for temporal expression annotation for english for tempeval 2010. Jun 1. 2016 Available: http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf

17. Chang, A., Manning, CD. TokensRegex: Defining cascaded regular expressions over tokens. Jul 1. 2016 Available: http://nlp.stanford.edu/software/tokensregex.html
18. Park MS, He Z, Chen Z, Oh S, Bian J. Consumer's use of UMLS concepts on social media: diabetes-related textual data analysis in blog and social Q&A sites. JMIR Diabetes. 2016 In press.
19. The Penn Treebank Project. Jul 1. 2016 Available: https://www.cis.upenn.edu/~treebank/
20. Lewis JH, Kilgore ML, Goldman DP, Trimble EL, Kaplan R, Montello MJ, et al. Participation of patients 65 years of age or older in cancer clinical trials. J Clin Oncol. Apr 1.2003 21:1383–9. [PubMed: 12663731]

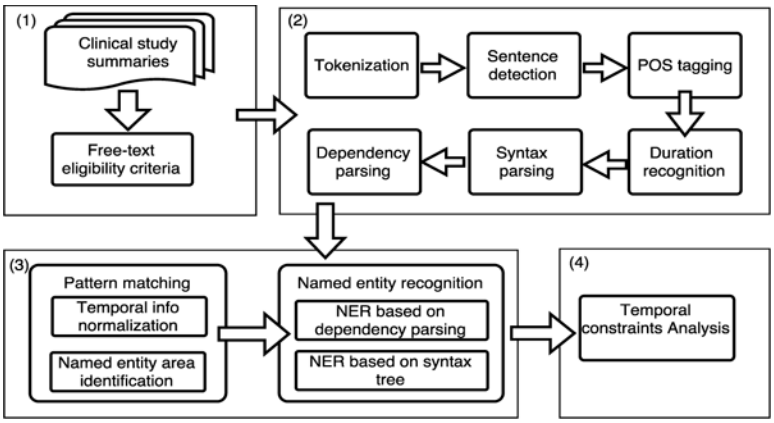

Fig. 1.
Analytical pipeline of this study

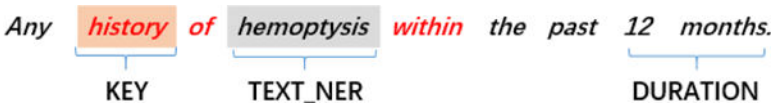

Fig.2.
An example sentence that matches the “HISTORY_WITHIN” pattern

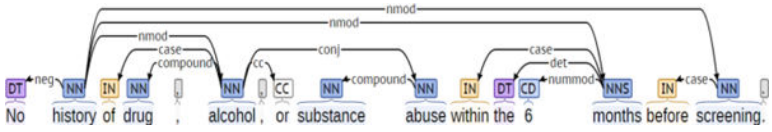

Fig. 3.
An exmaple of dependency parsing

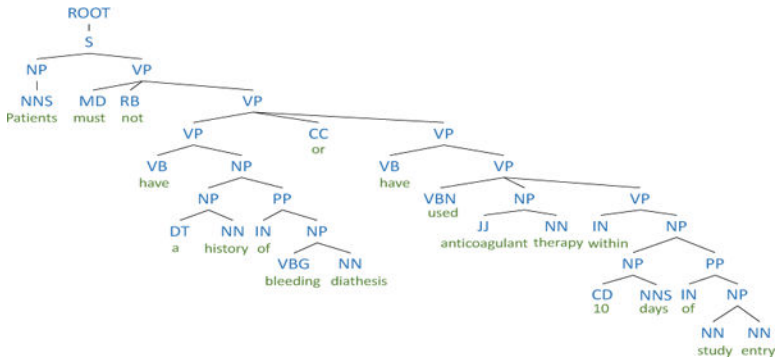

Fig. 4.
Example of syntax tree for an criterion

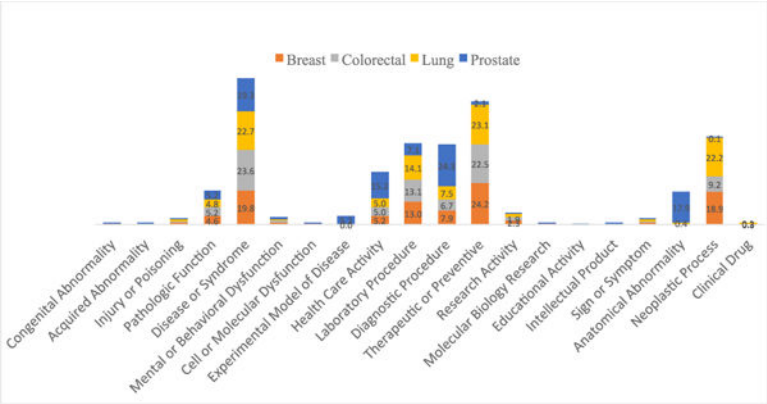

Fig. 5.
Distribution of semantic types of eligibility criteria

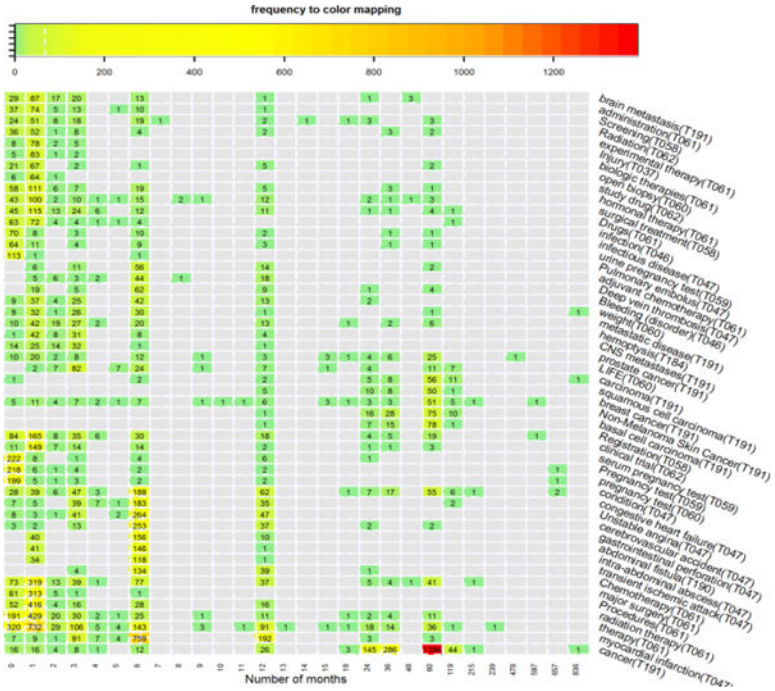

Fig. 6.

Temporal constraint distribution of frequent qualitative eligiblity criteria

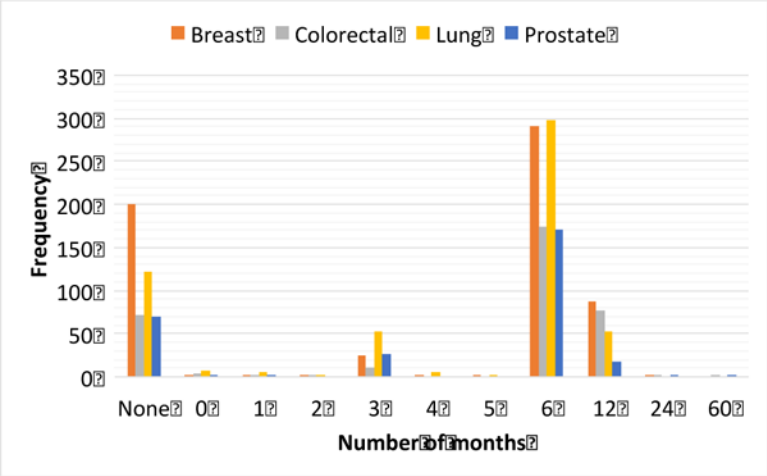

Fig. 7.
Distribution of temporal constraints for myocardial infarction

TABLE I

Example Representation of Contextual Criteria Pattern Using TokenRegex

Pattern	(“history”) “of” ([]{1,20}) (“within”|“in”|“less” “than”|“equals” “to”) ([]{0,5} [{ ner:DURATION }]+)
Action	Annotate($1, ner, “KEY”), Annotate($2, ner, “TEXT_NER”)
Stage	20
Pattern name	HISTORY_WITHIN

TABLE II

The UMLS Semantic Types Included in This Study

TUI	Semantic Type	TUI	Semantic Type
T020	Acquired Abnormality	T191	Neoplastic Process
T190	Anatomical Abnormality	T046	Pathologic Function
T049	Cell or Molecular Dysfunction	T184	Sign or Symptom
T019	Congenital Abnormality	T060	Diagnostic Procedure
T047	Disease or Syndrome	T058	Health Care Activity
T050	Experimental Model of Disease	T059	Laboratory Procedure
T037	Injury or Poisoning	T063	Molecular Biology Research Technique
T048	Mental or Behavioral Dysfunction	T061	Therapeutic or Preventive Procedure
T200	Clinical Drug		

TABLE III

Number of Clinical Studies

Cancer type	Number of studies	Interventional studies	Observational studies	Others
Lung	2,880	2,407 (83.6%)	452 (15.7%)	21 (0.7%)
Breast	4,267	3,457 (81.0%)	786 (18.4%)	24 (0.6%)
Prostate	2,123	1,820 (85.7%)	293 (13.8%)	10 (0.5%)
Colorectal	1,578	1,316 (83.4%)	255 (16.2%)	7 (0.4%)

TABLE IV

Characteristics of Eligiblity Criteria of Cancer Studies

Subsection of criteria	Negation?	Count
Exclusion Criteria	No	93,915
	Yes	12,409
Inclusion Criteria	No	133,828
	Yes	41,735

TABLE V

List of Lexical Patterns, Their Frequency in Eligibility Criteria, And A Sample Sentence for Each Pattern

Pattern name	Overall freq.	Freq. in inclusion criteria	Freq. in exclusion criteria	Sample criterion
WITHIN	10,298	4,501	5,797	No myocardial infarction **within** *the past 6 months*
MORETHAN	5,184	3,998	1,186	Patients must have been off previous anti-androgen therapy for **more than** *4 weeks*.
WITHIN_PT	2,931	1,284	1,647	2 metastases seen on standard imaging **within** *30 days* **prior to** registration
ATLEAST_PT	2,279	2,132	225	**At least** *5 years* since **prior** chemotherapy
MORETHAN_PT	1,439	1,273	166	**More than** *4 weeks* since **prior** immunotherapy
HISTORY_WITHIN	1,048	147	901	No **history** of nephrolithiasis **within** *the past 5 years*
PRIOR_TO	1,014	528	486	Completed treatment for breast cancer *a minimum of 1 year* **prior to** study enrollment.
HISTORY_WITHIN_PT	456	38	418	**History** of prostatic surgery **within** *4 weeks* **prior to** the screening visit
PRIOR_F	308	194	114	**Prior** radiation castration with amenorrhea **for** *at least 6 months*
HISTORY_MORETHAN	138	55	83	**History** of radiation castration and amenorrheic for >= *3 months*
HISTORY_WITHIN_EF	89	3	86	**History** of other malignant disease **in** the past 5 years **except** basal cell carcinoma
CONCURRENT_EF	85	54	31	No **concurrent** or planned chemotherapy or surgery **for** at least 2 months after radiotherapy
HISTORY_MORETHAN_PT	41	16	25	**History of** treatment by complete androgen blockade **for greater than** *3 months* **prior to** enrollment
CONFIRMED_WITHIN	37	37	0	**Histologically confirmed** adenocarcinoma of the prostate **within** *the past 120 days*
POSITIVE_WITHIN	34	17	17	At least one **positive** biopsy **within** *the previous 6 months*
PRIOR_MT	19	11	8	**Prior** use of hormonal therapy for prostate cancer **for more than** *2 months*.
PRIOR_FEF	10	5	5	**Prior** chemotherapy for prostate cancer **within** *12 months* before enrollment **except from** docetaxel