Journal of Biomedical Informatics 58 (2015) S211–S218

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Textual inference for eligibility criteria resolution in clinical trials

Chaitanya Shivade [a,*], Courtney Hebert [b], Marcelo Lopetegui [b,c], Marie-Catherine de Marneffe [d], Eric Fosler-Lussier [a], Albert M. Lai [b]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[b] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA
[c] Clínica Alemana de Santiago, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile
[d] Department of Linguistics, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

*Article history:*
Received 11 February 2015
Revised 2 September 2015
Accepted 4 September 2015
Available online 14 September 2015

*Keywords:*
Natural language processing
Electronic health records
Textual inference
Clinical trials

ABSTRACT

Clinical trials are essential for determining whether new interventions are effective. In order to determine the eligibility of patients to enroll into these trials, clinical trial coordinators often perform a manual review of clinical notes in the electronic health record of patients. This is a very time-consuming and exhausting task. Efforts in this process can be expedited if these coordinators are directed toward specific parts of the text that are relevant for eligibility determination. In this study, we describe the creation of a dataset that can be used to evaluate automated methods capable of identifying sentences in a note that are relevant for screening a patient's eligibility in clinical trials. Using this dataset, we also present results for four simple methods in natural language processing that can be used to automate this task. We found that this is a challenging task (maximum *F*-score = 26.25), but it is a promising direction for further research.



## 1. Introduction

Clinical trials play an important role in medical research. The successful completion of a trial is dependent on achieving a significant sample size of patients enrolled into the trial within a limited time period. However, the process of eligibility determination is extremely challenging and time-consuming, often mandating manual chart review [1,2]. Such reviews can involve repeated readings of the patients' electronic health record (EHR) for multiple trials, across every visit. This limits the number of patients that can be evaluated. Although institutions participating in clinical trials spend significant resources in conducting eligibility screening, the cost of screening is generally not fully compensated through the contracts supporting the trials [3].

The eligibility requirements of a patient, for a clinical trial, are specified in the form of inclusion and exclusion criteria. These are detailed descriptions of the characteristics a patient must or must not have in order to participate in the trial. Patients can be pre-screened for eligibility by referring to either structured or unstructured data in the EHR, or a combination of both. While structured data such as diagnosis codes, laboratory results, medication orders, procedure information, and problem lists are useful for eligibility determination, clinical notes still remain the preferred means of documentation for physicians [4]. These notes frequently contain nuances of clinical presentation and care that are critical for making an eligibility screening determination that the structured data does not. Moreover, the criteria are specified in natural language. Hence, not all criteria can be translated into queries that leverage structured data. Köpcke et al. [5] found that there was a significant gap (65%) between the structured data documented for patient care and the data required for eligibility assessment.

Thus, the use of clinical notes from the EHR is imperative for eligibility determination and clinical trial coordinators often undertake lengthy reviews of the EHR, specifically clinical notes, to assist with determining patient eligibility. Since this process is expensive in terms of time and effort, it would be desirable to speed up this step in the eligibility screening workflow. The NLP community has developed techniques for problems such as question answering [6], textual entailment [7], textual inference [8] and information retrieval [9], which can be used for identifying text of interest from the entire document.

In this paper, we describe the creation of a dataset and evaluate baseline methods for identifying text in clinical notes that is pertinent to a given eligibility criterion for a trial. We compare simple

* Corresponding author at: 395 Dreese Laboratories, 2015 Neil Avenue, Columbus, OH 43210, USA.
*E-mail address:* shivade@cse.ohio-state.edu (C. Shivade).
*URL:* http://cse.ohio-state.edu/~shivade/ (C. Shivade).

http://dx.doi.org/10.1016/j.jbi.2015.09.008
1532-0464/

methods in natural language processing (NLP) that can identify specific sentences in clinical notes that are relevant to a trial's eligibility criteria. Such methods may be used to direct trial coordinators to specific parts of the notes, making the process of manual review easier, in the future. Recently, Köpcke et al. [10] reviewed 79 Clinical Trial Recruitment Support Systems (CTRSS) across 101 publications and found that success of a CTRSS depends on its successful workflow integration, rather than on sophisticated reasoning and processing algorithms. Although it needs formal validation, we believe that our proposed approach can be incorporated into the existing workflow seamlessly and will therefore help expedite the eligibility determination process.

## 2. Dataset creation

In order to evaluate the performance of an automated system that can identify text relevant to eligibility criteria in clinical notes, we created a new gold standard dataset, as outlined below. The 2014 i2b2 challenge [11] had two tracks with a training set of 790 notes, with annotations for protected health information and risk of heart disease, respectively. The third track evaluated the usability of systems developed for the i2b2 challenges. The fourth track allowed participants to demonstrate novel use of the data made available through the challenge. This work was a submission to this fourth track, proposing the use of NLP for expediting clinical trial eligibility screening. Since a limited amount of time was available to attempt the challenge, 80 notes were annotated for four criteria (20 notes per criterion) by our team of annotators (described in Section 2.3). Every annotator marked all full sentences in a note that were relevant to an eligibility criterion. These annotations are similar to other NLP shared tasks [7,12] such as the Recognizing Textual Entailment (RTE) challenges in the open domain. Therefore, systems developed for these annotations can leverage NLP research from such shared tasks for a challenging domain-specific real world problem and identify interesting research questions. There is at present no other publicly available dataset in the clinical domain with such annotations.

### 2.1. Criteria selection

The National Library of Medicine and the National Institutes of Health maintain the website www.clinicaltrials.gov, a registry and a database of publicly and privately supported clinical studies across the globe. The 2014 i2b2 challenge provided notes for coronary artery disease (CAD) patients from three cohorts: (1) patients with no CAD, (2) patients who developed CAD over the course of provided notes for that patient, and (3) patients who had CAD from beginning of their record. Therefore, we downloaded from clinicaltrials.gov study record data for all trials associated with the search term "coronary artery disease" as XML files (5054 trials in the May 2014 download). We extracted eligibility criteria for these trials and segmented them into individual sentences using a combination of the Ling-pipe [13] sentence chunking module trained on MEDLINE biomedical abstracts and user defined rules to identify sentence boundaries. Each sentence was assumed to be a single criterion. The first author (CS) and the physician (CH) in the team considered the following factors while selecting the criteria: (1) the resolution of the criterion would require a healthcare professional to refer to the notes of the patient, (2) it would be hard to assess patient eligibility for the criterion by referring only to the structured data associated with the patient's EHR, and (3) the criterion is commonly used across clinical trials for CAD. The criteria were selected to reflect realistic situations where clinical trial coordinators spend substantial time reading notes to assess eligibility, and would benefit from a NLP system such as the one proposed in this study.

The segmented criteria sentences described above were analyzed using MetaMap [14] to identify concepts from the Unified Medical Language System (UMLS). Ignoring concepts such as "patient," "study," "trials," and "subject," we found that medical history (C0262926) and lesion (C0221198) were the most common concepts in these criteria. We chose one eligibility criterion associated with each of these two concepts (Criteria 1 and 2). We also found that angina was the most common concept in the semantic category 'Sign or Symptom' across the eligibility criteria for CAD trials. It is a common practice among physicians to specify angina using grades as specified by the Canadian Cardiovascular Society Angina Grading Scale [15]. We chose classification of patients into Class 1 (Criterion 3) and Class 3 angina (Criterion 4) as the other two eligibility criteria in our study. These criteria are shown in Table 1.

### 2.2. Annotation process

Many studies have used physicians to create gold standard datasets in the clinical domain. However, Raghavan et al. [16] showed that individuals with varying level of clinical expertise could also generate high quality annotations. Our annotation team consisted of two senior undergraduate nursing students, a second-year graduate-entry nursing student, and a physician. We developed detailed annotation guidelines for the annotators for each criterion. The annotation process was carried out using **CLIN**ical Tr**I**als **CrI**teria **AN**notator (CLINICIAN), a tool developed at our institution. CLINICIAN is a secure web-based tool and has a simple user interface.

After having logged in, annotation for a single note has the following workflow: the user selects a criterion of choice (screenshot shown in Fig. 1a) to annotate notes for. This brings up a webpage (screenshot shown in Fig. 1b) that displays the criterion statement at the top and a note beneath it. The user first determines whether the note has any text that is relevant to the criterion and marks the note as "Not relevant" otherwise. If found to be relevant, the user highlights all complete sentences in the note that contribute toward determining, whether the patient meets the criterion, based on that particular note. After selecting all relevant sentences, the user clicks on one of the three buttons "Yes," "No," or "Maybe," indicating an assessment of whether the patient meets the criterion. This brings up the next note and the user repeats the workflow. Table 2 summarizes the distribution of these annotations in terms of mean ($\mu$) and standard deviation ($\sigma$) across 80 notes in the dataset and the appendix lists the details.

### 2.3. Annotation agreement

In order to ensure that these guidelines covered all cases, we carried out four training rounds. Each round involved annotations of ten notes per criterion (40 in total). All four annotators worked on the same set of 40 notes. This was followed by a discussion about the correctness of these annotations. Decisions associated

**Table 1**
Criteria used for annotations.

| Id | Statement |
|---|---|
| 1 | Known location of a coronary artery lesion |
| 2 | History of revascularization procedure |
| 3 | Ordinary physical activity such as walking and climbing stairs does not cause angina but angina with strenuous or rapid or prolonged exertion at work or recreation |
| 4 | Marked limitation of ordinary physical activity but walking one or two blocks on the level and climbing one flight of stairs in normal conditions at normal pace |

with correctness of annotations were made through group discussions led by the physician. The annotation guidelines were updated after every training round to restrict differences amongst annotators. After completing the training, we calculated reliability of agreement amongst the four annotators based on annotations obtained for 20 notes per criterion (80 in total). We ensured that notes from the training rounds were not used in the agreement testing set. We used Fleiss' kappa [17] as a metric of agreement, since more than two annotators were involved.

We calculated three metrics of agreement. As outlined in the previous section, if a note was found to be relevant, it was annotated for criterion eligibility. Thus, every note could be relevant ("Yes," "No," "Maybe") or "Not Relevant." The first metric calculated agreement on whether a note was relevant. The second metric calculated agreement for assessment of criterion eligibility criteria across all four answers, "Yes," "No," "Maybe," and "Not Relevant." All notes were segmented into constituent sentences using the Ling-pipe [13] sentence chunking module trained on MEDLINE biomedical abstracts. We calculated agreement on sentences found to be relevant by the annotators. A sentence was considered relevant only if the annotator highlighted it. We can observe in Table 3 that the annotators achieve high agreement considering overall performance across all three metrics.

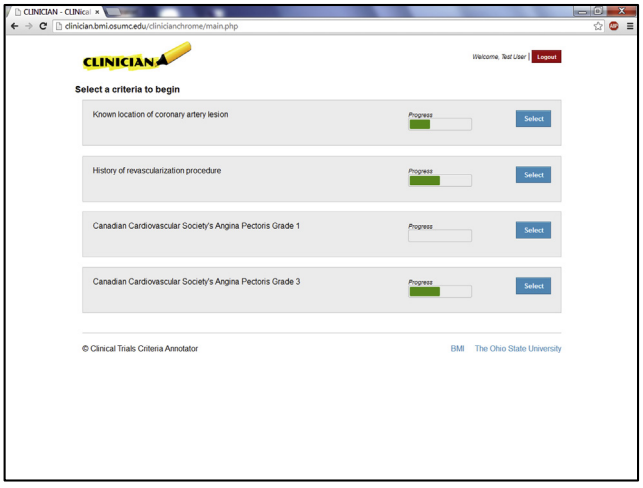


**Fig. 1a.** Criterion selection screen.

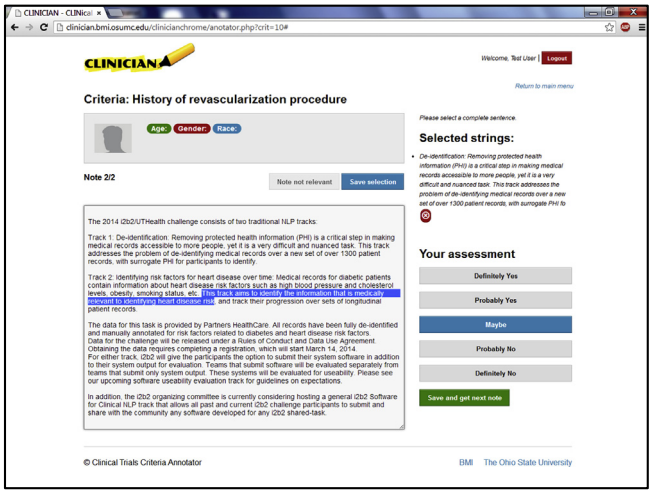


**Fig. 1b.** Working screen for note annotation.

**Table 2**
Overview of annotations in the dataset.

| Criterion | Relevant sentences, $\mu$ | Relevant sentences, $\sigma$ | Total number of sentences, $\mu$ | Total number of sentences, $\sigma$ |
|---|---|---|---|---|
| 1 | 1.05 | 1.23 | 49.45 | 44.59 |
| 2 | 1.65 | 2.39 | 44.30 | 26.29 |
| 3 | 1.35 | 1.59 | 41.45 | 33.18 |
| 4 | 1.75 | 2.12 | 52.55 | 30.42 |

## 3. Related work

There has been a large body of work investigating automated methods identifying eligible patients for clinical trials. Many articles investigate methods to standardize eligibility criteria for trials [18–20]. These methods facilitate the automation by representing natural language eligibility criteria into a computable format. The medical records tracks of the annual Text Retrieval Challenge (TREC) [21] for 2011 and 2012 aimed at identifying specific cohorts of patients based on the content of clinical notes. The dataset consisted of a variety of clinical notes (radiology reports, progress notes, discharge summaries, etc.) mapped to multiple patient visits. Participants of these challenges had to develop an information retrieval system to identify visits relevant to an eligibility criterion. However, these criteria were synthetically created and the goal was to identify relevant notes at a coarse level of granularity, namely patient visits (a single visit could constitute up to 50 notes). Unfortunately this dataset is no longer available [21]. Ni et al. [22] evaluate a system working on real-world trials with a goal similar to the TREC challenge and retrieve patient encounters relevant to trial. Their system uses a combination of NLP, information extraction and machine learning methods.

This work focuses on a task that involves retrieving information at a finer level of granularity. The goal is to identify specific sentences within a note that are relevant to a trial eligibility criterion. It is similar to the larger problem of question–answering in NLP. Many studies [6,23,24] have explored clinical question–answering in the past. However, all of these systems look into scientific articles for answers. McKeown et al. [25] describe a system that can customize keyword search results over biomedical literature by re-ranking them using concepts identified in a patient's EHR. As stated in Patrick and Li [26], no question–answering system investigates use of clinical notes for finding answers. Athenikos and Han [27] discuss in their extensive survey of clinical question answering systems, that current systems are limited in the type of questions they can handle. Very few systems can handle questions in natural language, and those that can, are limited to handling a set of restricted types such as definitional questions. This limits the utility of such systems for trial recruitment where criteria are natural language statements, not even questions, with varying degree of specificity and complexity [28].

The Recognizing Textual Entailment (RTE) [7] challenges have also been very popular with the NLP community. The problem statement for these challenges is closest to our work: given two

**Table 3**
Fleiss' kappa for agreement metrics.

| Metric | Overall | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 |
|---|---|---|---|---|---|
| Note relevance | 0.858 | 0.751 | 0.883 | 0.810 | 0.948 |
| Criterion assessment | 0.795 | 0.769 | 0.818 | 0.611 | 0.846 |
| Sentence relevance | 0.799 | 0.758 | 0.784 | 0.825 | 0.813 |

fragments of text (called *hypothesis* and *text*), decide whether the meaning of the *hypothesis* can be inferred from the *text*. In the first three RTE challenges, automated systems were developed for answering "Yes" or "No" to the question of entailment. For example: The hypothesis "A Filipino hostage was freed in Iraq" should be inferred to be true from the text "A Filipino hostage in Iraq was released." Similarly, the hypothesis "Time Warner is the world's largest company" should be inferred to be false from the text "Time Warner is the world's largest media and Internet company." The later challenges, RTE-4 [29] to RTE-7 [30] added a third category of "Unknown." From RTE-1 [7] to RTE-5 [31], while the size of the hypothesis, in terms of number of words, remained almost the same (7–10 words), the average size of the text increased steadily from 24 words in RTE-1 to 99 words in RTE-5. The RTE-5 challenge had a pilot task that was corpus-based: it was a search task aimed at finding all sentences that entail a given hypothesis, in a given set of documents, about a topic. This is analogous to the task of finding all sentences that are relevant to a given eligibility criterion, in a given set of clinical notes, for a particular diagnosis.

The RTE tasks have been known to be hard problems in NLP. None of systems in RTE-1 could perform better than the baseline and accuracy was between 0.5 and 0.6 on a balanced dataset of positive and negative entailments, which steadily increased from 0.75 for RTE-2 to 0.8 for RTE-3. RTE-4 and RTE-5which had a 3-way task to determine entailment between text-hypothesis pairs as "Yes," "No" and "Unknown," noticed a fall in accuracy, with the best accuracy being 0.68 for both challenges. The RTE-5 pilot, RTE-6 and RTE-7 challenges had a different setting with a search-based task. The best *F*1 scores for these tasks were 0.455, 0.48 and 0.48 respectively.

There are a few important differences between the RTE challenges and our task. First, the *hypotheses* in the RTE challenges were manually curated based on the *text* against which entailment was to be ascertained. Care was taken to ensure that the hypotheses were explicit, thus limiting ambiguities, as well as concise and easy to interpret in terms of spatial and temporal descriptions. In our case, the criteria and the notes reflect data in the real world. The criteria therefore do not obey the above desirable properties. Second, the *texts* of interest in RTE challenges were newswire documents, which are fairly consistent and well formed in terms of English grammar, spellings and punctuations. Our text of interest originates from clinical notes, which are characterized by misspellings, abbreviations, incomplete sentences, inconsistent document structure and other undesirable properties of poor text quality for any NLP task. Third, there is a significant word overlap between the text and the hypothesis in these challenges and this increased steadily from 69.25% in RTE-1 to 77.14% in RTE-5 for the positive entailment relation. Thus, it was observed that systems relying on lexical overlap performed well [31]. For the search-based tasks, this overlap dropped to 47.59% for RTE-5 pilot and then increased from 54.66% for RTE-6 to 58.89% for RTE-7. Therefore, it was observed that systems taking word overlap into account had better chances of performing well. Our data collection shows a very low word-overlap between the criteria and the relevant sentences ($\mu = 17\%$) making the task much more challenging. Lastly, the RTE challenges aimed at open-domain textual inference, while our task is very domain-specific. Systems in the RTE challenges used external knowledge sources such as WordNet [32], Wikipedia extracts, and FrameNet [33]. to enrich their systems. Ablation tests of participating systems indicated that they had a significant impact on the system performance. However such resources are domain-independent and hence have very little coverage in the biomedical domain [34]. Given the quality of data and domain-specific nature of our task, these resources are likely to fall short.

## 4. Problem definition

Textual Entailment systems can be applied in three different modes [35]. In the *recognition* mode, given a text and hypothesis pair as input, the system needs to classify whether entailment holds for the pair or not. This was the focus of challenges RTE-1 to RTE-5. In the *search* mode, the system is given a hypothesis and a corpus, and needs to find all text fragments in the corpus that entail the hypothesis. RTE-5 had a pilot task and exploring this mode and was the main task for RTE-6 and RTE-7. In the *generation* mode, the system is given a text, and needs to generate statements which are entailed by the text. Although, we have collected annotations that can be used for either of these modes, the focus of this study is to evaluate a system in the search mode. Thus, given an eligibility criterion of interest, the goal is to automatically identify all sentences in a note that are relevant to that criterion. The system is evaluated using standard metrics of precision, recall and *F*1 which are computed by comparing the system output with the gold standard annotations.

## 5. Methods

In order to develop an understanding of the task, we implemented two lexical methods considered as baselines in the RTE literature. We also implemented two semantic methods that are adaptations of these baselines to the clinical domain that are informed by specialized knowledge-sources. These implementations develop an understanding of the challenges associated with the task and serve as a direction for further research. These algorithms are applied at a sentence level in every clinical note to determine a relevance score of every sentence with a criterion statement. In terminologies used by the RTE community, these algorithms were applied to pairs of text and hypotheses, where text is a sentence in the note (denoted as *N*) and hypothesis is a sentence defining the criterion (denoted as *C*). It should be noted that these methods ignore document-level features such as co-references, negations, and document structure. All sentences in a note having a positive score are considered to be relevant to the criterion.

### 5.1. Lexical methods

#### 5.1.1. Lexical matching

This is one of the simplest algorithms for textual inference that computes a score based on lexical overlap obtained by matching the lemma for each word in *C* with the lemma for some word in *N*, ignoring stop words.

#### 5.1.2. Lucene-based matching

Apache Lucene [36] is an open source text search engine; given a query of search terms, it returns a set of documents ranked by score of relevance. This system served as a pure information retrieval baseline in the search-based RTE challenges (RTE-5 to RTE-7), where every sentence is a document and the search terms in the query are all words in the hypothesis. The Lucene search engine can be configured using parameters to analyze the query terms and the text to be searched, in different ways. We followed parameters as described in the RTE challenge guidelines, namely, the StandardAnalyzer (tokenization, lowercase, stop-word filtering and basic clean-up of words), a Boolean "OR" query for all words in the criterion sentence and the default document scoring function. Unlike the RTE challenge, which uses an older version (2.9.1) of Lucene, we used version 4.4 for our experiments.

As in the RTE challenge, we evaluated this approach by considering a limited number of sentences top-ranked by Lucene, as the relevant sentences for a criterion, across all notes for that criterion. We found that, on average, there were between two to three sentences annotated as relevant to a criterion in our dataset. As mentioned earlier, since 20 notes were annotated for each criterion, we considered different thresholds in the range of 30–50 sentences (in increments of 5) for each criterion.

### 5.2. Semantic methods

#### 5.2.1. Concept matching

The UMLS brings together many biomedical vocabularies and standards. It is the largest thesaurus of biomedical concepts, mapping them to semantic types and associating them through hierarchical and non-hierarchical relations. We used MetaMap [14] to identify UMLS concepts in every pair of *C* and *N*. A score was computed for every sentence *N*, as the number of concepts in *N* that were exactly identical to a concept in *C*.

#### 5.2.2. UMLS similarity-based matching

UMLS::Similarity is a freely available open source tool [37] that can be used to obtain similarity or relatedness between any two concepts from the UMLS. Given one or more ontologies and one or more hierarchical relations of interest, the tool treats the UMLS as a graph, with concepts as nodes and relations as edges. The tool has a library of measures to compute similarity between two concepts using different graph-based properties. As in the previous approach, we identified UMLS concepts in every pair of *C* and *N* using MetaMap. A similarity score was computed between every pair of concepts in a given pair of *C* and *N*. The score for a sentence *N* was the sum of similarity scores its constituent concepts share with the concepts in a criterion *C*.

Systematized Nomenclature of Medicine–Clinical Terms (SNOMED–CT) is the most comprehensive healthcare terminology in the world. Pedersen et al. [38] demonstrated that similarity measures between clinical concepts, computed using different measures, had high correlation with physicians and human coders. They used parent–child relationships between concepts in SNOMED–CT to define the graph and computed similarity scores. We used the same relations, but on the version of SNOMED–CT (2013_01_31) included in the 2013AA release of the UMLS.

### 5.3. Comparison of similarity measures

The UMLS::Similarity tool provides implementations of a number of similarity measures capturing different relationships between two concepts. This includes path-based measures, information content-based measures and corpus-based measures. The simplest ones are based on the path information between two concepts in the UMLS graph. The *path* implementation is simply the inverse of path length between two concepts. The *cdist* is an implementation of the measure proposed by Rada et al. [39] that computes the number of edges along the shortest path between two concepts. Wu and Palmer [40] proposed a measure, *wup* incorporating depth of the Least Common Subsumer (LCS) of the two concepts into the similarity calculations. The *lch* measure proposed by Leacock and Chodorow extends the *path* measure by incorporating depth of the taxonomy. Finally, Nguyen and Al-Mubaid [41] incorporate both depth and LCS in their measure *nam*.

Information content (IC) of a concept is defined as the negative log of probability of the concept. Implementation of the simplest measure *res* proposed by Resnik [42] computes the IC of the LCS of two concepts. The *jcn* and *lin* implementations of measures proposed by Jiang and Conrath [43] and Lin [44] respectively propose using IC of the two concepts and their LCS in the similarity calculation.

Finally, the implementation of a corpus based method *vector* is also available. This method proposed by Patwardhan and Pedersen [45] computes word vectors for each word in the definition of a concept and further captures distributional similarity of words co-occurring with the definition word in the corpus. These vectors are averaged to create a single co-occurrence vector for the concept. Similarity between two concepts is the cosine of the two vectors. A detailed description for each of these metrics can be found in a study [37] published by the authors of the tool.

### 5.4. Effect of similarity threshold

While some of the similarity measures discussed above are bounded between 0 and 1, some of them are not. We varied the threshold value for considering two concepts to be similar using 10-fold cross validation. Since each criterion was annotated for 20 notes, a fold consisted of two documents. Thus, training was carried out on 18 notes and testing was carried out on 2 notes. An optimal similarity threshold was determined in the training and used in the testing for evaluation.

### 5.5. Effect of the amount of data on similarity computations

As mentioned earlier, in order to determine if a note sentence *N* was relevant to criterion *C*, the score for a sentence was calculated as the sum of similarity scores its constituent concepts share with the concepts in a criterion *C*. This similarity score was calculated using the different metrics described above. Path-based measures rely only on the structure of the UMLS graph. However, similarity using the IC measures, *res*, *lin* and *jcn*, require a corpus for the IC calculation of a concept. Similarly the corpus-based measure *vector* requires a corpus for the distributional similarity calculations. We varied the amount of data used for computing these measures. The data for the i2b2 challenge was made available to the participants in increments. The first batch of training data consisted of 521 notes, followed by a second and final training batch of 269 notes. A test set of 514 notes was also made available. We varied the data used for computing similarity reflecting these increments. First, we used only the 80 notes in our dataset. This was followed by notes from the first training batch (521 notes), the full training set (790 notes) and finally the full data set (1304 notes) comprising both the training and testing sets. It should be noted that although the amount of data for computing IC and distributional similarity was varied in these experiments, the 10-fold cross validation experiments described above always used the same amount of data for training (18 notes) and testing (2 notes) in each fold.

## 6. Evaluation

We evaluated the performance of all four approaches by using standard metrics of precision, recall and *F*1-score on a per sentence basis. A prediction was assessed as correct if a method identifies a sentence in the note to be relevant to the criterion, as per the gold standard annotations.

The lexical matching and concept matching methods do not involve any parameters. We simply evaluated them on all notes in our dataset. Table 9 summarizes the performance of these two methods across all four criteria in terms of the *F*1-score. The last column in the table reflects the overall performance of a method, which is the average *F*1-score across all four criteria.

As stated in Section 5.1, the Lucene-based method was evaluated for different thresholds of top scored sentences considered as relevant to the criterion. The performance remained more or less

**Table 4**
Performance of Lucene at varying thresholds.

| Threshold | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 | Average |
|---|---|---|---|---|---|
| 30 | **0.4** | 0.000 | 0.140 | **0.062** | 0.150 |
| 35 | **0.4** | 0.000 | **0.161** | 0.057 | 0.155 |
| 40 | **0.4** | 0.027 | **0.161** | 0.053 | 0.161 |
| 45 | **0.4** | **0.051** | **0.161** | 0.050 | **0.166** |
| 50 | **0.4** | 0.048 | **0.161** | 0.047 | 0.164 |

constant. The best performance was achieved at a threshold of 45 sentences being considered as relevant from the set of 20 notes for each criterion. Table 4 summarizes these results. Values in bold indicate the highest value in that column.

The evaluation for UMLS::Similarity-based matching methods was carried out using 10-fold cross validation. We calculated micro-averaged and macro metrics. The micro-averaged $F1$ was calculated by averaging the $F1$ score across the 10-folds. It should be noted that although a fold is defined at a note level, the $F1$ score is calculated at a sentence level. Thus, while calculating this metric, the total number of predictions for a fold equals the sum of number of sentences in the two test notes for that fold. In contrast, a macro-$F1$ is calculated by considering all sentence level predictions across the 20 notes together. In this calculation, the total number of predictions equals the total number of sentences across all 20 notes. Table 5 summarizes the micro-averaged performance of different similarity measures with their standard deviations ($\sigma$).

The standard deviation across the folds is very large. This is possibly because there is a large variation in the number of sentences occurring in a note ($\mu = 46.97$, $\sigma = 33.70$). Similarly there is a large variation in the number of criterion-relevant sentences ($\mu = 2.46$, $\sigma = 1.84$). A large number of notes (33 = 41.25%) in the dataset do not have any sentences that are relevant to a criterion. Therefore we also calculated a macro $F1$ score, which considers predictions for all sentences in the 20 notes during the test folds for performance calculations. These evaluations are summarized in Table 6.

We also evaluated the IC-based measures and the corpus-based method by varying the amount of data they use in calculating IC and distributional similarity respectively. The results shown in Tables 5 and 6 use only 80 notes from our dataset. The variation in performance due to the amount of data was calculated both using micro-$F1$ and macro-$F1$. These results are summarized in Tables 7 and 8 respectively. Contrary to our expectations, we observed that performance of most measures worsened with an increase in data. To understand the reason behind these observations, we compared the precision and recall of predictions in experiments with increasing amounts of data. On comparing the precision and recall values for predictions using only 80 notes with those using the full dataset, we found that a drop in the macro-$F1$ was associated with the decrease in precision of the predictions. It is possible that adding more data is contributing to adding noise than capturing similarity in the IC-based measures.

**Table 6**
Performance comparison of different similarity measures based on macro-$F1$.

| Similarity measure | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 | Average |
|---|---|---|---|---|---|
| path | 0.162 | 0.089 | 0.369 | 0.080 | 0.175 |
| cdist | 0.162 | 0.089 | 0.369 | 0.080 | 0.175 |
| wup | 0.162 | 0.152 | 0.367 | 0.089 | 0.193 |
| lch | 0.162 | 0.094 | 0.369 | 0.075 | 0.175 |
| Zhong | 0.158 | 0.169 | 0.342 | 0.090 | 0.189 |
| nam | 0.154 | 0.103 | **0.667** | 0.070 | **0.248** |
| jcn | 0.133 | 0.145 | 0.640 | 0.031 | 0.237 |
| lin | **0.229** | 0.261 | 0.354 | 0.048 | 0.223 |
| res | 0.061 | **0.367** | 0.304 | 0.062 | 0.198 |
| vector | 0.140 | 0.075 | 0.308 | **0.109** | 0.158 |

**Table 7**
Effect of the amount of data on performance of similarity measures, measured in terms of micro-$F1$.

| | 80 notes | | First train batch | | All training data | | All data | |
|---|---|---|---|---|---|---|---|---|
| | $F1$ | $\sigma$ | $F1$ | $\sigma$ | $F1$ | $\sigma$ | $F1$ | $\sigma$ |
| *jcn* | | | | | | | | |
| Criterion 1 | **0.033** | 0.105 | **0.033** | 0.105 | 0.031 | 0.097 | 0.031 | 0.097 |
| Criterion 2 | **0.091** | 0.121 | 0.060 | 0.085 | 0.055 | 0.089 | 0.087 | 0.115 |
| Criterion 3 | **0.591** | 0.335 | 0.591 | 0.335 | 0.417 | 0.297 | 0.417 | 0.297 |
| Criterion 4 | 0.031 | 0.034 | **0.045** | 0.046 | **0.045** | 0.046 | **0.045** | 0.046 |
| Overall | **0.186** | 0.149 | 0.182 | 0.143 | 0.137 | 0.132 | 0.145 | 0.139 |
| *lin* | | | | | | | | |
| Criterion 1 | **0.125** | 0.213 | 0.115 | 0.194 | 0.075 | 0.169 | 0.075 | 0.169 |
| Criterion 2 | **0.154** | 0.169 | 0.113 | 0.110 | 0.113 | 0.148 | 0.113 | 0.148 |
| Criterion 3 | 0.400 | 0.294 | **0.406** | 0.301 | **0.406** | 0.301 | **0.406** | 0.301 |
| Criterion 4 | 0.057 | 0.057 | **0.062** | 0.078 | 0.042 | 0.072 | 0.042 | 0.072 |
| Overall | **0.184** | 0.183 | 0.174 | 0.171 | 0.159 | 0.172 | 0.159 | 0.172 |
| *res* | | | | | | | | |
| Criterion 1 | 0.022 | 0.070 | 0.019 | 0.060 | **0.102** | 0.255 | 0.022 | 0.070 |
| Criterion 2 | **0.225** | 0.258 | 0.194 | 0.210 | **0.225** | 0.258 | **0.225** | 0.258 |
| Criterion 3 | 0.333 | 0.211 | **0.511** | 0.278 | 0.485 | 0.274 | 0.485 | 0.274 |
| Criterion 4 | 0.058 | 0.074 | 0.059 | 0.075 | **0.074** | 0.077 | **0.074** | 0.077 |
| Overall | 0.160 | 0.153 | 0.196 | 0.156 | **0.222** | 0.216 | 0.202 | 0.170 |
| *vector* | | | | | | | | |
| Criterion 1 | 0.097 | 0.160 | 0.132 | 0.267 | 0.142 | 0.268 | **0.158** | 0.259 |
| Criterion 2 | **0.047** | 0.066 | 0.031 | 0.066 | 0.035 | 0.049 | 0.022 | 0.047 |
| Criterion 3 | **0.322** | 0.190 | 0.279 | 0.193 | 0.276 | 0.195 | 0.260 | 0.194 |
| Criterion 4 | **0.108** | 0.120 | 0.079 | 0.078 | 0.091 | 0.096 | 0.027 | 0.058 |
| Overall | **0.144** | 0.134 | 0.130 | 0.151 | 0.136 | 0.152 | 0.117 | 0.139 |

The *res* implementation of the information content-based similarity measure proposed by Resnick [42] gave the best macro-$F1$ score using the full training corpus for information content calculations. The best result from Lucene and UMLS::Similarity-based matching is summarized along with the other two methods in Table 9.

While differences in overall (considering all four criteria together) predictions, between lexical matching and the Lucene based matching are not significant, all other differences are statistically significant ($p < 0.05$) as per McNemar's test [46]. We inves-

**Table 5**
Performance comparison of different similarity measures based on micro-$F1$.

| Similarity measure | Criterion 1 | | Criterion 2 | | Criterion 3 | | Criterion 4 | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | $F1$ | $\sigma$ | $F1$ | $\sigma$ | $F1$ | $\sigma$ | $F1$ | $\sigma$ | $F1$ |
| path | 0.075 | 0.169 | 0.024 | 0.051 | 0.414 | 0.292 | 0.041 | 0.071 | 0.138 |
| cdist | 0.075 | 0.169 | 0.024 | 0.051 | 0.406 | 0.292 | 0.041 | 0.071 | 0.136 |
| wup | 0.062 | 0.131 | 0.103 | 0.110 | 0.406 | 0.293 | 0.044 | 0.075 | 0.154 |
| lch | 0.075 | 0.169 | 0.03 | 0.063 | 0.414 | 0.292 | 0.050 | 0.085 | 0.142 |
| Zhong | 0.059 | 0.125 | 0.1245 | 0.101 | 0.378 | 0.281 | 0.049 | 0.083 | 0.153 |
| nam | 0.083 | 0.18 | 0.0276 | 0.059 | **0.648** | 0.268 | 0.074 | 0.092 | **0.208** |
| jcn | 0.033 | 0.105 | 0.0907 | 0.121 | 0.590 | 0.335 | 0.031 | 0.034 | 0.186 |
| lin | **0.125** | 0.212 | 0.1542 | 0.169 | 0.400 | 0.087 | 0.057 | 0.057 | 0.184 |
| res | 0.022 | 0.070 | **0.2254** | 0.258 | 0.333 | 0.211 | 0.058 | 0.074 | 0.160 |
| vector | 0.097 | 0.159 | 0.0471 | 0.066 | 0.322 | 0.190 | **0.108** | 0.119 | 0.144 |

**Table 8**
Effect of the amount of data on performance of similarity measures, measured in terms of macro-*F*1.

| | 80 notes *F*1 | First train batch *F*1 | All training data *F*1 | All data *F*1 |
|---|---|---|---|---|
| *jcn* | | | | |
| Criterion 1 | **0.133** | 0.087 | 0.129 | 0.129 |
| Criterion 2 | **0.145** | 0.097 | 0.091 | 0.141 |
| Criterion 3 | **0.640** | **0.640** | 0.370 | 0.370 |
| Criterion 4 | 0.031 | 0.050 | **0.051** | **0.051** |
| Overall | **0.237** | 0.219 | 0.161 | 0.173 |
| *lin* | | | | |
| Criterion 1 | **0.229** | 0.195 | 0.167 | 0.167 |
| Criterion 2 | **0.261** | 0.167 | 0.191 | 0.191 |
| Criterion 3 | **0.354** | 0.351 | 0.351 | 0.351 |
| Criterion 4 | 0.048 | 0.067 | **0.082** | **0.082** |
| Overall | **0.223** | 0.195 | 0.198 | 0.198 |
| *res* | | | | |
| Criterion 1 | **0.061** | 0.032 | 0.150 | 0.036 |
| Criterion 2 | **0.367** | 0.317 | **0.367** | **0.367** |
| Criterion 3 | 0.304 | **0.507** | 0.475 | 0.475 |
| Criterion 4 | 0.062 | 0.064 | **0.075** | **0.075** |
| Overall | 0.198 | 0.230 | **0.267** | 0.238 |
| *vector* | | | | |
| Criterion 1 | 0.140 | 0.159 | **0.177** | 0.151 |
| Criterion 2 | 0.075 | **0.118** | 0.064 | 0.079 |
| Criterion 3 | **0.308** | 0.265 | 0.260 | 0.243 |
| Criterion 4 | **0.109** | 0.072 | 0.085 | 0.109 |
| Overall | **0.158** | 0.153 | 0.147 | 0.145 |

**Table 9**
Comparative performance all four methods.

| Method | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 | Average |
|---|---|---|---|---|---|
| Lexical matching | 0.368 | 0.086 | 0.182 | 0.070 | 0.176 |
| Concept matching | 0.182 | 0.101 | 0.156 | **0.088** | 0.132 |
| Lucene | **0.400** | 0.051 | 0.161 | 0.050 | 0.166 |
| UMLS similarity matching | 0.150 | **0.367** | **0.475** | 0.075 | **0.267** |

tigated the results and found that semantic approaches performed better in criteria where the relevant sentences in a note described the condition in related, but different words. This was the primary reason why the method leveraging UMLS::Similarity performed well. For example, relevant sentences for criterion 2: "History of revascularization procedure" are framed using specific names of procedures such as "CABG (Coronary Artery Bypass Graft)" which are impossible to capture without use of an external knowledge base. This strengthens our intuition behind why textual inference is harder in the clinical domain than open domain due to a low lexical overlap as stated in Section 3. Instances where UMLS Similarity Matching falls short are those where concepts are identified as relevant by human annotators but, the similarity measure fails to capture them. For example: the concepts coronary lesion (C3272304) and NSTEMI (C3537184) are found to be related by the annotators, but the similarity metrics fail to capture this relation. This is most dominant in criterion4 where all methods perform the worse. Although, the largest number of criterion-relevant sentences ($n = 35$) are found for this criterion, there is variability in language that is not captured by the current methods. We believe this is problem can be addressed if more labeled data is available.

## 7. Limitations and future work

The primary limitation of this study is the size of data used in these experiments. We used only 20 notes per criterion (80 notes in total) in our experiments. A larger dataset with annotations for relevant sentences can possibly yield more accurate and robust results. Our annotation team consists of nursing students led by a physician. Having physicians or experts from the domain (in this case, cardiologists) as annotators could improve upon the quality of annotations. The UMLS::Similarity tool is used in conjunction with a reference terminology to compute similarity between UMLS concepts. In this work, we followed previous work [38] and used SNOMED–CT as the reference terminology. Recent studies [47,48] have investigated the problem of choosing the right UMLS terminology for NLP tasks clinical notes. Determining the right terminology or a combination of them to conclude similarity between concepts is an unexplored topic of research. The eligibility criteria used in this study were manually picked by the authors as explained in Section 2.1. Automating this step will be a useful direction of research for textual entailment in clinical text. Although we evaluate a search-based textual entailment task, the annotations described in this study are complete for recognition-based tasks too. Our future work will include investigating machine-learning methods for improving performance of the current system.

## 8. Conclusion

In this paper, we describe the problem of textual inference in the context of clinical trial recruitment. The objective of this work was to evaluate baseline methods that can identify sentences from clinical notes that are relevant to different clinical trial eligibility criteria. To do this, we constructed a dataset and showed that a high agreement was achieved between annotators with varying level of clinical expertise. This suggests that there is systematic reasoning employed by humans for performing this task. We evaluated baseline computational methods by considering previous work for similar tasks and found that there is significant room for improvement. We found that semantic methods gave better results than lexical methods and appear to be a promising direction for improvement. The biomedical domain has rich knowledge sources such as the UMLS, which can be leveraged to perform intelligent inferences in clinical text. In the future, we plan to explore such resources and linguistic properties of text that are specific to the clinical domain to achieve better results in this task.

## Conflict of interest statement

The authors declare that they have no competing interest.

## Acknowledgments

We would like to thank our annotators Alissa Schultz, Jennifer Fox and Jessica Schellenbach for their help in the annotation effort. Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011116. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.09.008.

## References

[1] L. Penberthy, R. Brown, F. Puma, B. Dahman, Automated matching software for clinical trials eligibility: measuring efficiency and flexibility, Contemp. Clin. Trials 31 (2010) 207–217, http://dx.doi.org/10.1016/j.cct.2010.03.005.

[2] G. Joseph, D. Dohan, Recruiting minorities where they receive care: institutional barriers to cancer clinical trials recruitment in a safety-net hospital, Contemp. Clin. Trials 30 (2009) 552–559, http://dx.doi.org/10.1016/j.cct.2009.06.009.

[3] L.T. Penberthy, B.A. Dahman, V.I. Petkov, J.P. DeShazo, Effort required in eligibility screening for clinical trials, J. Oncol. Pract. 8 (2012) 365–370, http://dx.doi.org/10.1200/JOP.2012.000646.

[4] S.T. Rosenbloom, J.C. Denny, H. Xu, N. Lorenzi, W.W. Stead, K.B. Johnson, Data from clinical notes: a perspective on the tension between structure and flexible documentation, J. Am. Med. Inform. Assoc. 18 (2011) 181–186, http://dx.doi.org/10.1136/jamia.2010.00723.

[5] F. Köpcke, B. Trinczek, R.W. Majeed, B. Schreiweis, J. Wenk, T. Leusch, et al., Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence, BMC Med. Inform. Decis. Mak. 13 (2013) 37, http://dx.doi.org/10.1186/1472-6947-13-37.

[6] D. Demner-Fushman, J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Comput. Linguist. 33 (2007) 63–103, http://dx.doi.org/10.1162/coli.2007.33.1.63.

[7] I. Dagan, O. Glickman, B. Magnini, The PASCAL recognising textual entailment challenge, in: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. D'Alché-Buc (Eds.), Proc. First Int. Conf. Mach. Learn. Challenges Eval. Predict. Uncertain. Vis. Object Classif. Recognizing Textual Entailment, vol. 3944, Springer, Berlin, Heidelberg, 2006, pp. 177–190, http://dx.doi.org/10.1007/11736790.

[8] B. MacCartney, C.D. Manning, Natural logic for textual inference, in: RTE '07 Proc. ACL-PASCAL Work. Textual Entailment Paraphrasing, Association for Computational Linguistics, Stroudsburg (PA, USA), 2007, pp. 193–200.

[9] R.M. Grad, P. Pluye, Y. Meng, B. Segal, R. Tamblyn, Assessing the impact of clinical information-retrieval technology in a family practice residency, J. Eval. Clin. Pract. 11 (2005) 576–586, http://dx.doi.org/10.1111/j.1365-2753.2005.00594.x.

[10] F. Köpcke, H.-U. Prokosch, Employing computers for the recruitment into clinical trials: a comprehensive systematic review, J. Med. Internet Res. 16 (2014) e161, http://dx.doi.org/10.2196/jmir.3446.

[11] A. Stubbs, C. Kotfila, H. Xu, Ö. Uzuner, Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform. 58S (2015) S67–S77, http://dx.doi.org/10.1016/j.jbi.2015.07.001.

[12] E.M. Voorhees, D.M. Tice, Building a question answering test collection, in: Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. – SIGIR '00, ACM Press, New York (NY, USA), 2000, pp. 200–207. doi:10.1145/345508.345577.

[13] LingPipe 4.1.0 <http://alias-i.com/lingpipe/>.

[14] A.R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, in: Proc. Annu. AMIA Symp., 2001. pp. 17–21.

[15] L. Campeau, Letter: Grading of angina pectoris, Circulation 54 (1976) 522–523.

[16] P. Raghavan, E. Fosler-Lussier, A.M. Lai, Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise, AMIA Annu. Symp. Proc. 2012 (2012) 1366–1374.

[17] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378–382, http://dx.doi.org/10.1037/h0031619.

[18] C. Weng, X. Wu, Z. Luo, M.R. Boland, D. Theodoratos, S.B. Johnson, EliXR: an approach to eligibility criteria extraction and representation, J. Am. Med. Inform. Assoc. 18 (Suppl. 1) (2011). i116-24. http://dx.doi.org/10.1136/amiajnl-2011-000321.

[19] C. Weng, S.W. Tu, I. Sim, R. Richesson, Formal representation of eligibility criteria: a literature review, J. Biomed. Inform. 43 (2010) 451–467, http://dx.doi.org/10.1016/j.jbi.2009.12.004.

[20] S.W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, et al., A practical method for transforming free-text eligibility criteria into computable criteria, J. Biomed. Inform. 44 (2011) 239–250, http://dx.doi.org/10.1016/j.jbi.2010.09.007.

[21] E.M. Voorhees, The TREC Medical Records Track, in: Proc. Int. Conf. Bioinf. Comput. Biol. Biomed. Inf. – BCB'13, ACM Press, New York (NY, USA), 2013, pp. 239–246, http://dx.doi.org/10.1145/2506583.2506624.

[22] Y. Ni, S. Kennebeck, J.W. Dexheimer, C.M. McAneney, H. Tang, T. Lingren, et al., Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department, J. Am. Med. Inform. Assoc. (2014). amiajnl – 2014–002887 – . http://dx.doi.org/10.1136/amiajnl-2014-002887.

[23] B.L. Cairns, R.D. Nielsen, J.J. Masanz, J.H. Martin, M.S. Palmer, W.H. Ward, et al. The MiPACQ clinical question answering system. AMIA Annu. Symp. Proc. 2011;2011:171–180.

[24] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J.J. Cimino, et al., AskHERMES: an online question answering system for complex clinical questions, J. Biomed. Inform. 44 (2011) 277–288, http://dx.doi.org/10.1016/j.jbi.2011.01.004.

[25] K.R. McKeown, N. Elhadad, V. Hatzivassiloglou, Leveraging a common representation for personalized search and summarization in a medical digital library, in: Proc. 3rd ACM/IEEE-CS Jt. Conf. Digit. Libr., IEEE Computer Society, 2003, pp. 159–170.

[26] J. Patrick, M. Li, An ontology for clinical questions about the contents of patient notes, J. Biomed. Inform. 45 (2012) 292–306, http://dx.doi.org/10.1016/j.jbi.2011.11.008.

[27] S.J. Athenikos, H. Han, Biomedical question answering: a survey, Comput. Methods Programs Biomed. 99 (2010) 1–24, http://dx.doi.org/10.1016/j.cmpb.2009.10.003.

[28] J. Ross, S. Tu, S. Carini, I. Sim, Analysis of eligibility criteria complexity in clinical trials, AMIA Summits Trans. Sci. Proc. 2010 (2010) 46–50.

[29] D. Giampiccolo, H. Trang Dang, B. Magnini, I. Dagan, E. Cabrio, B. Dolan, The fourth PASCAL recognizing textual entailment challenge, Text Anal. Conf. 2008. Gaithersburg, MD, USA, 2009.

[30] L. Bentivogli, C. Peter, I. Dagan, D. Giampiccolo, The seventh PASCAL recognizing textual entailment challenge, in: Proc. TAC, 2011.

[31] L. Bentivogli, I. Dagan, H.T. Dang, D. Giampiccolo, B. Magnini. The Fifth PASCAL Recognizing Textual Entailment Challenge, in: Proc. TAC, 2009.

[32] C. Fellbaum, WordNet: an electronic lexical database. Bradford Books, 1998. p. 10.

[33] C.F. Baker, C.J. Fillmore, J.B. Lowe, The Berkeley framenet project, Proc. 36th Annu. Meet. Assoc. Comput. Linguist., vol. 1, Association for Computational Linguistics, Morristown (NJ, USA), 1998, p. 86, http://dx.doi.org/10.3115/980845.980860.

[34] O. Bodenreider, A. Burgun, Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System, in: Proc. NAACL 2001 Work. WordNet other Lex. Resour. Appl. Extensions Cust., Pittsburgh, PA, 2001. pp. 77–82.

[35] I. Dagan, D. Roth, M. Sammons, Recognizing Textual Entailment, Morgan & Claypool Publishers, 2013.

[36] Apache Lucene <http://lucene.apache.org/> [accessed 12.08. 14].

[37] B.T. McInnes, T. Pedersen, S.V.S. Pakhomov, UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity, in: Proc. Annu. AMIA Symp., vol. 2009, 2009, pp. 431–435.

[38] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (2007) 288–299, http://dx.doi.org/10.1016/j.jbi.2006.06.004.

[39] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Trans. Syst. Man Cybern. 19 (1989) 17–30, http://dx.doi.org/10.1109/21.24528.

[40] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proc. 32nd Annu. Meet. Assoc. Comput. Linguist., Association for Computational Linguistics, Morristown (NJ, USA), 1994, pp. 133–138, http://dx.doi.org/10.3115/981732.981751.

[41] H. Al-Mubaid, H.A. Nguyen, Measuring semantic similarity between biomedical concepts within multiple ontologies, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 39 (2009) 389–398, http://dx.doi.org/10.1109/TSMCC.2009.2020689.

[42] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, 1995, pp. 448–453.

[43] J. Jay, D.W.C. Jiang, Semantic similarity based on corpus statistics and lexical taxonomy, n.d.

[44] D. Lin, An Information-Theoretic Definition of Similarity, 1998. pp. 296–304.

[45] S. Patwardhan, T. Pedersen, Using WordNet-based context vectors to estimate the semantic relatedness of concepts, in: Proc. EACL, 2006.

[46] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (1998) 1895–1923, http://dx.doi.org/10.1162/089976698300017197.

[47] C. Shivade, P. Malewadkar, E. Fosler-Lussier, A.M. Lai, Comparison of UMLS terminologies to identify risk of heart disease in clinical notes, J. Biomed. Inform. (2015).

[48] S.T. Wu, H. Liu, D. Li, C. Tao, M.A. Musen, C.G. Chute, et al., Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis, J. Am. Med. Inform. Assoc. 19 (2012) e149–e156, http://dx.doi.org/10.1136/amiajnl-2011-000744.