Journal of Biomedical Informatics 46 (2013) 1145–1151

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Unsupervised mining of frequent tags for clinical eligibility text indexing

Riccardo Miotto [a], Chunhua Weng [a,b,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[b] The Irving Institute for Clinical and Translational Research, Columbia University, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Accepted 29 August 2013
Available online 10 September 2013

Keywords:
Information storage and retrieval
Clinical trials
Tags
Information filtering
Eligibility criteria
Controlled vocabulary

ABSTRACT

Clinical text, such as clinical trial eligibility criteria, is largely underused in state-of-the-art medical search engines due to difficulties of accurate parsing. This paper proposes a novel methodology to derive a semantic index for clinical eligibility documents based on a controlled vocabulary of frequent tags, which are automatically mined from the text. We applied this method to eligibility criteria on ClinicalTrials.gov and report that frequent tags (1) define an effective and efficient index of clinical trials and (2) are unlikely to grow radically when the repository increases. We proposed to apply the semantic index to filter clinical trial search results and we concluded that frequent tags reduce the result space more efficiently than an uncontrolled set of UMLS concepts. Overall, unsupervised mining of frequent tags from clinical text leads to an effective semantic index for the clinical eligibility documents and promotes their computational reuse.

1. Introduction

Parsing clinical eligibility text is important to leverage the reuse of clinical information for automatic decision support [1,2]. Following this assumption, various methods and techniques have been recently developed to transform clinical trial protocol text into computable representations that can benefit automated tasks (e.g., classification, clustering, discovery [3–9]). Several efforts specifically focused on clinical trial eligibility criteria, which define the characteristics that a research volunteer must possess to qualify for a clinical trial study. Some of these techniques index eligibility criteria using template-based semantic patterns or formal ontologies (e.g., [10,11]), whereas others extract from the text terms covered by the Unified Medical Language System (UMLS) lexicon [12] (e.g., [13]). Nevertheless, eligibility criteria generally remain as free text and underused in modern computational tasks such as search. As an example, ClinicalTrials.gov [14] does not process the eligibility criteria text when ranking trials in response to user queries [15]. A major reason is that a standardized and widely accepted parser for clinical trial eligibility criteria is not yet defined [16,17].

The indexing methods proposed in the literature generally parse each clinical trial separately without considering textual similarities among them. This results in an ever-expanding index with high dimensionality and high likelihood of presenting too specific, redundant, or irrelevant concepts for individual documents, which is not amenable for automated processing. To address this issue, we propose an alternative approach based on cross-processing eligibility criteria from multiple studies to mine a finite vocabulary of tags frequently shared by these trials, thus serving as a semantic index for eligibility text.[1]

In the information retrieval literature, the problem of document indexing and tagging has been robustly studied in different application scenarios as well as in terms of information theory [18–21]. Tags are generally used in exploratory retrieval, in which users engage in iterative cycles of document refinement and exploration of new information (as opposed to standard free-text retrieval). A controlled vocabulary of tags defines an interpretative layer of semantics over the text and its parsed representation, and generally leads to more effective retrieval than uncontrolled annotations [22]. For example, the use of a controlled vocabulary benefited different text search applications (e.g., [23–26]) as well as multimedia retrieval (e.g., [27,28]).

We hypothesize that (1) an unsupervised, fully automated data mining approach applied to the clinical trial repository can produce a finite set of tags that is frequently shared among all trials and (2) these frequent tags can lead to a general and stable index of eligibility text, which can leverage the automated processing of clinical trials. This method is potentially superior to other approaches by balancing and minimizing the sparseness of the index and increasing efficiency and specificity of retrieval operations [29]. Moreover, because of their high frequency, tags extracted

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, USA. Fax: +1 212 305 3302.
E-mail address: cw2384@columbia.edu (C. Weng).

[1] In this domain, tags are defined as meaningful multi-word semantic concepts automatically extracted from the text, such as, e.g., "breast carcinoma", "diabetes", "active malignancy".

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.012

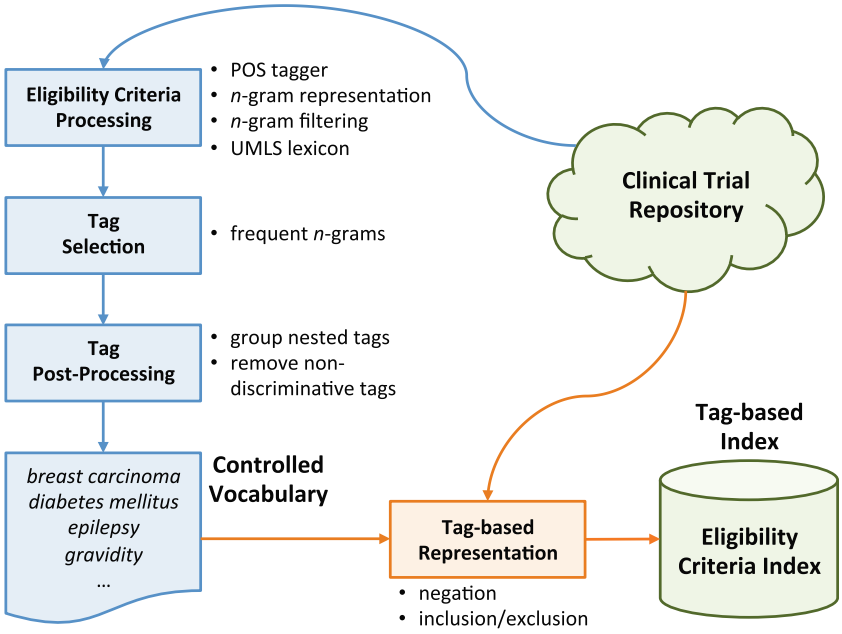

Fig. 1. Overview of tag mining and clinical trial eligibility criteria indexing.

from the text are likely to be more generic and intuitive than independent annotations, and therefore might be also more effective in helping users with interactive tasks [30].

The original contribution of this article is three-fold: we (1) present a novel method for mining a controlled vocabulary of frequent tags from clinical eligibility text; (2) apply this method to ClinicalTrials.gov and report statistics on tag distributions; and (3) propose to evaluate the effectiveness of frequent tags at filtering clinical trial search results when there is no gold standard available for comparing these tags against.

2. Material and methods

Fig. 1 depicts the proposed tag mining approach. First, the eligibility criteria in the repository are processed to extract potential tags individually. Then, only the tags meeting the frequency threshold are retained, post-processed, and used to index the trials.

2.1. Mining frequent tags

Eligibility text is often divided into two sections, one specifying whom to include (inclusion criteria) and the other whom to exclude (exclusion criteria). It might be listed explicitly and separately in a tabular format or expressed together as a general and vague text. Fig. 2 provides two examples of eligibility criteria with different structures, one tabular and the other free text. Classifying a criterion as inclusion or exclusion is straightforward when the division is explicitly reported, but more difficult when that division is implicit, since inclusion and exclusion criteria can be expressed in different ways. However, because tags are meant to identify high-level general concepts shared among clinical trials, the mining process can just focus on extracting tags regardless of their classification.

2.1.1. Eligibility text processing

The algorithm to process the eligibility criteria of a clinical trial relies on basic text processing techniques [31]. First, each criterion is automatically annotated with a part-of-speech (POS) tagger, defined in the *Natural Language Toolkit* [32], to identify the grammatical role of each word. In this application, the grammatical role of a word will be used only for noise reduction (e.g., to remove tags composed by only, e.g., verbs, adverbs); for this reason, we favored a general well-established solution rather than a more domain-related one [33]. The text is then processed to remove special characters and punctuation and to build all the possible *n*-grams (i.e., continuous sub-sequences of *n* words).[2] *N*-grams composed of only English stop words or irrelevant grammatical structures are removed.

Lastly, each *n*-gram is matched against the UMLS Metathesaurus and retained only if at least one substring of it is a recognizable UMLS concept. Moreover, we considered only those UMLS concepts appearing in semantic categories most relevant to the clinical trial domain [34] (i.e., 27 different semantic types out of 136, including, e.g., "Disease or Syndrome", "Individual Behavior", "Finding") in order to reduce the number of extraneous tags. As an example, "malignancy within the past 5 years" is considered a valid *n*-gram because at least one word, "malignancy", is present in the part of the UMLS lexicon considered, even if the entire sentence is not.[3] Each *n*-gram term found in the UMLS lexicon is also normalized according to its preferred Concept Unique Identifier (CUI) in order to reduce the sparseness of the concepts. Using the CUIs also enables the handling of synonyms, since similar concepts are aligned to the same preferred term because of the UMLS specification (e.g., "atrial fibrillation" and "auricular fibrillation" are both mapped to "atrial fibrillation"). This allows defining a vocabulary possibly composed by semantically unique tags. After this process, each clinical trial's eligibility criteria are summarized by a set of UMLS CUI-based *n*-grams representing the criteria's relevant concepts.

2.1.2. Frequent tag selection

Given a repository of clinical trials and their *n*-gram-based representations, the set of tags is obtained by retaining the *n*-grams

[2] We used *n*-grams with lengths ranging from 1 to 10. In fact, we observed in preliminary experiments that tags longer than 7 words were unlikely to appear frequently. Therefore, we used 10 as maximum length to handle potential outliers.

[3] If the regular UMLS lexicon were used, the previous example "malignancy within the past 5 years" would have had three terms correctly matched (i.e., "malignancy", "past", and "years").

Eligibility Criteria expressed in a Tabular Format

Inclusion Criteria:

- *must be 18 years of age or older*
- *newly diagnosed primary breast cancer*
- *must have a breast or breast/chest MRI ordered as part of standard of care*
- *must meet one or more of the following in Part 1:*
- *T1 carcinoma of the breast appropriate for breast conserving surgery diagnosis established by core biopsy*
- *Tis (ductal carcinoma in situ) carcinoma > 2 cm diameter*
- *invasive lobular carcinoma of the breast*
- *must meet the following in Part 2:*
    - *known invasive or in situ primary breast cancer identified on pre-surgical breast MRI to have an unsuspected 2nd lesion recommended for biopsy and biopsy has not yet been performed*

Exclusion Criteria:

- *known diabetes mellitus*

Eligibility Criteria expressed as Free-text

All men and women had 2 to 8 diopters of simple myopia and were correctable to 20/20 or better with glasses or contact lenses. All patients had the stability of their myopia documented by previous records. Patients were at least 21 years of age and lived in the metropolitan area of the study centers. Each patient agreed to have surgery on one eye and to wait 1 year for surgery on the other eye. Patients with systemic diseases that might affect corneal wound healing and patients with high corneal astigmatism were excluded from the study.

Fig. 2. Examples of eligibility criteria text (tabular-format vs. free-text).

appearing in a minimum number of documents (i.e., *R*). The list of frequent tags is then post-processed to remove noise and duplicates.

First, we discarded irrelevant tags based on the nestedness information, i.e., the frequency of a tag being a subsequence of longer tags. To this aim we used the *C*-values [35,36]. These coefficients represent the nestedness degree of a string combining its total frequency in the collection, its length, its frequency as part of longer strings, and the number of these distinct longer strings. A low *C*-value indicates that a string is highly related to its superstrings. Therefore, we discarded substrings with low *C*-values because most likely related to longer tags and, hence, without a relevant meaning as standalone terms (e.g., "coronary artery" was discarded by the presence of "coronary artery diseases").[4]

Second, we removed tags appearing too frequently because they were generally not useful to distinguish trials. To do this, we weighted the relevance of each tag in the collection through the Inverse Document Frequency (IDF) metric [37] and removed tags with an IDF value of at least 4 standard deviations lower than the mean (this removes about 5% of tags). Examples of highly frequent tags that were discarded are "primary", "doctor", "patient", and "written consent form". The remaining tags define the controlled vocabulary to index the clinical eligibility text.

2.2. Indexing eligibility text

At indexing time the eligibility criteria of each clinical trial in the repository were parsed with the method described in Section 2.1.1 and only the *n*-grams included in the controlled vocabulary were retained. The collection of all tag-based representations builds the eligibility criteria index that can be exploited to improve the computational reuse of clinical trials.

At this time, tags can also be embedded with contextual details inferred from the text (e.g., role, temporal definition, numerical validity ranges, negation) and related to their semantic type (e.g., "condition", "laboratory result", "finding") as well as to the applicative scenario. As an example, in this study we considered characterizing tag role and possible negation (e.g., "without breast cancer") in the criterion to which they refer. Therefore, a tag that unambiguously refers to an inclusion or exclusion criterion is prefixed with "ic" or "ec" respectively (e.g., when criteria are clearly listed in tabular format); conversely, a more ambiguous tag is prefixed with "nt" (i.e., "no type") to indicate that its classification cannot be deduced from text. Similarly, when a tag appears as negated in a criterion sentence, the "NOT" prefix is added to it in the tag-based representation. To do this, we applied NegEx, a widely used regular expression algorithm that implements several phrases indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases [38].

In addition, we considered tags related to age and gender differently from all the others in order to reduce the sparseness of the information (e.g., female gender could be expressed as "woman", "women", "female" in the trials). So, when age or gender details are identified in eligibility criteria, we mapped them to regular standard patterns, such as, "gender = male", "minimum age = 18", "maximum age = 70".

Last, Table 1 reports a sample of four clinical trials from ClinicalTrials.gov represented by their tags.

2.3. Evaluation

We used all 137,889 clinical trials available on the ClinicalTrials.gov repository as of December 2012. Evaluation of algorithms for eligibility criteria indexing poses difficulties due to the lack of

[4] The numeric scale of the *C*-values is related to the size of the corpus on which they are computed. Therefore for generalizing the filtering algorithm, we normalized this value according to the size of the collection, and we discarded tags appearing at least 70% of the time as substring.

Table 1
A sample of clinical trials from ClinicalTrials.gov with the associated tags: (ic) inclusion criteria, (ec) exclusion criteria, (nt) no type.

Clinical trial	Tags
NCT01364194 Local Infiltration with Bupivacaine to Increase Quality of Postoperative Pain Control in Total Knee Replacement	ec:creatinine clearance, ec:hypersensitivity, ec:liver function, gender = both, ic:body mass index, maximum age = 80, minimum age = 50
NCT00001535 Twins Study of Gene Therapy for HIV Infection	ec:contraceptive methods, ec:gravidity, ec:hepatitis b, ec:hepatitis c, ec:lymphoma, ec:substance abuse problem, gender = both, ic:hiv infections
NCT00103883 A Novel Method to Determine HIV Incidence Among Youth	ec:emotionally unstable, gender = both, ic:hiv infections, maximum age = 24, minimum age = 12
NCT00004997 Leucovorin for the Treatment of 5q Minus Syndrome	gender = both, nt:NOT ecog performance status, nt:NOT finding of platelet count, nt:NOT gravidity, nt:NOT hiv infections, nt:NOT leukemia, nt:NOT pharmacologic substance, nt:NOT pharmacotherapy, nt:NOT skeletal bone marrow, nt:NOT transplantation

well-established gold standards. In addition, the use of manual reviewers introduces subjectivity and is labor-intensive and error-prone. With these considerations, we propose an evaluation framework aimed at showing (1) statistics on how the tags apply to the data repository and (2) the benefits introduced by the frequent tags in one of the potential applicative scenarios (e.g., filtering clinical trial search results).

As a baseline, we used a tool annotating independently each eligibility criterion with its UMLS concepts [13,39]; we refer to this method as "UMLS-only" tagging. The latter was based on the algorithm described in Section 2.1.1 (i.e., *n*-gram filtering according to: UMLS lexicon, 27 UMLS semantic categories, grammar properties, etc.) to process each clinical trial independently. Using *C*-value analysis, we discarded irrelevant substrings within each trial representation as well. Similarly to Section 2.1.2, we also removed highly frequent but not discriminative tags across the collection based on their IDF value. Last, UMLS-only annotations were embedded with contextual details as in Section 2.2.

2.3.1. Tag distribution

We evaluated the size of the controlled vocabulary and the distribution of tags in the eligibility criteria index according to different values of R (i.e., minimum percentage of trials in the collection in which a tag must appear to be considered frequent). We also analyzed the differences among vocabularies obtained from different sized corpuses of trials. This evaluates the risk the tags becoming out-of-date, a common problem of controlled vocabulary-based approaches.[5] To do this, we measured the percentage of tags in common between vocabularies mined using subsets composed of N random trials and the tags mined using the entire collection. To generalize the results, for each value N, we performed three analyses based on different randomly sampled corpuses of trials and reported their average results.[6]

2.3.2. Tags to filter clinical trial search results

As one of the potential applicative scenarios, we propose to apply the controlled vocabulary to filter the results of a clinical trial search engine. The idea is that tags can be suggested to the users to reduce the results achieved by initial general searches not processing the trial eligibility text information. To do this, we used 50 different conditions (available in Appendix A) linked to more than 1,000 trials from ClinicalTrials.gov as queries and conducted simple searches (i.e., each search returned more than 1,000 trials). We used a controlled vocabulary of 115 tags mined from the entire ClinicalTrials.gov repository with $R = 3\%$ ("cvocab-3.0"), i.e., a tag had to appear in at least 4,137 trials to be considered frequent (see Appendix B).

First, for each query, we considered the five most frequent tags in the resulting documents and used them for query expansion (i.e., adding one of these tags to the initial query). We measured the percentage of resulting documents that were discarded when expanding the query with an additional tag ("document reduction rate"). Second, we performed an experiment showing that it is feasible to reduce the initial search resulting documents to a manually reviewable list by selecting a limited number of tags. We envisioned an applicative domain where the list of the ten most frequent tags in the resulting set is shown beside the rank list for further filtering based on eligibility criteria. These tags are updated dynamically at every user selection according to the remaining trials. For each of the 50 query conditions, we performed 500 distinct simulations (i.e., a total of 25,000 experiments) based on random tag selections to filter the result sets until only one trial remained. We evaluated the number of tags required to reach a preset threshold for the resulting documents (i.e., at most 3, 5, 10, 20, or 50 trials).

3. Results

In the tables presented in the following sections, the symbol (∗) after a numeric value denotes that the difference between the proposed approach and UMLS-only is statistically significant at the 5% level.

3.1. Tag distribution statistics

Table 2 reports number of tags and document coverage (i.e., percentage of trials represented by at least one tag over the entire collection) for controlled vocabularies mined from the ClinicalTrials.gov repository using different values of R. Values of $R < 1\%$ lead to a higher number of tags and likely noisier vocabulary, while values between 1% and 5% achieve an acceptable number of tags (i.e., between 100 and 500), defining a more compact index. However, R does not affect the document coverage obtained by the controlled vocabulary, which approaches the result achieved by UMLS-only with about 30,000 tags.

Fig. 3 shows the distribution of tags across clinical trials (i.e., how many trials have been assigned with a certain number of tags). Again, the value of R affects the number of tags assigned to the trials, with the controlled vocabulary generally obtaining more compact indexes than UMLS-only. This is manifested by the longer

[5] A controlled vocabulary becomes out-of-date when it is not able to represent the new documents added to a repository. In this case, tags fit the collection used for mining, but data are too various to define significant steady patterns and the new documents cannot be represented by the current vocabulary.

[6] Random corpuses for each value of N were sampled without replacement whenever possible through all the experiments. However, given the limited number of trials available, random corpuses started to have a minimum overlap for values of $N \geqslant 45{,}000$.

Table 2
Distribution of tags over the ClinicalTrials.gov repository. We compared controlled vocabularies mined using different values of *R* (i.e., 0.5%, 1%, 3%, 5%, referred as "cvocab-*R*") as well as tags from the UMLS-only approach. "Tag Number" is the number of tags in the vocabulary; "Document Coverage" is the percentage of documents in the collection represented by at least one tag.

	Tag vocabulary				
	UMLS-only	cvocab-0.5	cvocab-1.0	cvocab-3.0	cvocab-5.0
Tag Number	31,153	1,054	486	115	64
Document Coverage (%)	99.3	99.0	98.9	98.8	97.7

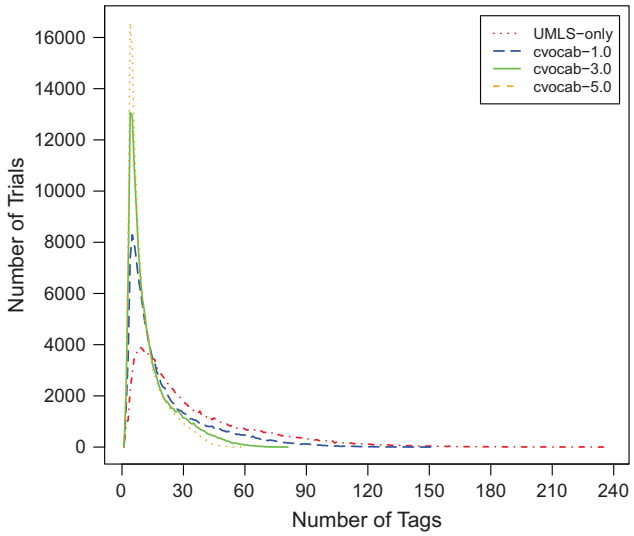

Fig. 3. Distribution of controlled vocabularies over the ClinicalTrials.gov repository obtained by different values of *R* (i.e., 1%, 3%, 5%, referred as "cvocab-*R*"). The curve obtained by UMLS-only is shown as well for comparison.

tail present in the UMLS-only curve, which shows that several trials were indexed with a much larger number of tags (i.e., more than 150 tags) than the others, leading to a more dispersive representation.

Last, Fig. 4 shows the percentage of tags shared between vocabularies mined from sub-samples of *N* trials and those mined using the entire collection of trials for different values of *R*. As can be seen, when $N \geqslant 40{,}000$, the number of tags in common approaches 95% for most of the *R* values (i.e., *R* = 3–5%). This means that the controlled vocabulary is likely to remain unchanged despite the collection of trials used to mine its tags, and consequently it is unlikely to become out-of-date.

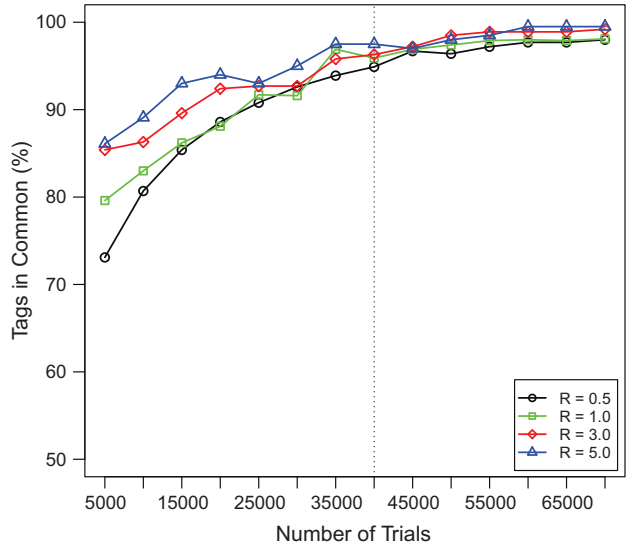

Fig. 4. Percentage of tags in common between vocabularies mined with different sub-collections of size *N* and the vocabulary mined using the entire collection of trials for different values of *R* (i.e., 0.5%, 1%, 3%, 5%). For each *N*, results are averaged over experiments performed on three different random corpuses of trials. The dashed line highlights when the number of tags in common exceeds 95% for most of the *R* values.

3.2. Effectiveness of tags to filter clinical trial search results

Table 3 reports the percentage of search results discarded when expanding the query with an additional tag. Results also show the average reduction achieved expanding the query with any 2-tag combination of the top five tags considered. As can be seen, the index based on the controlled vocabulary leads to significantly improved search result reduction compared to UMLS-only approach. As an example of the applicative scenario, Table 4 compares the tags suggested from both approaches to three queries.

Tags were recommended including their classification as inclusion or exclusion criteria; we believe this scenario would be more useful for a user in refining the resulting set of trials. Moreover, it should be noted that the reduced set also includes the trials indexed with the chosen tag prefixed by "nt" (i.e., no type) in order to avoid losing relevant information. For example, if the tag recommended were "ec:gravidity", the filtered resulting sets would include the trials indexed with that tag as well as those indexed with "nt:gravidity", in order to include trials having "gravidity" used in ambiguous exclusion criteria. Similarly, refined results for the recommended tag "nt:gravidity" would include trials indexed with "ec:gravidity" and "ic:gravidity" as well.

Last, Table 5 reports the number of tags required to reduce the number of trials to a specified upper limit. For each query condition, we measured the minimum, maximum, and mean number of tags over the 500 random simulations, and we report the results averaged over all the queries. Again, cvocab-3.0 generally achieves significant better results than UMLS-only, enabling users to reduce the resulting document space with a minor number of operations.

4. Discussion

Experiments on ClinicalTrials.gov confirmed a few of our hypotheses. First, frequent tags are unlikely to change radically as the repository grows and hence do not need to be continuously updated. In fact, given the high overlap among tags obtained using sub-samples of the collection (see Fig. 4) and the continuous but not excessive release of clinical trials,[7] existing frequent tags are likely to represent most of the new coming trials as well. Second, frequent tags can index almost the same number of trials as uncontrolled UMLS concept-based annotations, yet using significantly fewer tags. Third, in addition to efficiency, the appropriate size of frequent tags permit a manual review of the tag list (e.g., to improve readability), which is impossible for the annotation method based on the much larger number of UMLS jargons.

[7] About 5,000 trials were added to ClinicalTrials.gov between August and December 2012.

Table 3
Query expansion to filter the results of a first search using a controlled vocabulary ($R = 3\%$, i.e., "cvocab-3.0") and UMLS-only tags. Results are the percentage of trials discarded when tags are added to the query (i.e., "Document Reduction Rate"). Experiments were performed adding either one tag taken from the 5 most frequent tags in the original result set or any two of these tags.

Tag Number	Tag Rank	Document Reduction Rate (%)	
		cvocab-3.0	UMLS-only
1 tag	*first*	**63.1***	59.0
	second	**68.2***	63.5
	third	**73.9***	68.6
	fourth	**77.1***	71.5
	fifth	**79.3***	74.3
	mean	**72.3***	67.4
2 tags	*mean*	**88.5***	84.2

* Indicates that the difference between the results is statistically significant.

To demonstration the value of this approach with an example, we proposed to apply the controlled vocabulary-based index to filter clinical trial search results. Our choice of this particular application is motivated by the fact that state-of-the-art clinical trial search engines, such as ClinicalTrials.gov, do not process eligibility criteria for retrieval and often return overwhelming results when users issue insufficiently specific queries [15,40]. As mentioned above, the lack of gold standards with which to compare tags precluded an information retrieval evaluation (e.g., precision and recall [31]). This application-oriented evaluation framework establishes the benefits of using tags to filter search results which are known to be relevant for the user (because they were obtained through a search performed on a well-established clinical trial search engine, i.e., ClinicalTrials.gov). We focused on simple queries that might be issued by a non-expert user to create the "worst case" scenario. Consequently, the more specific the initial query, the fewer tags required by the filtering mechanism to reach a number of trials suitable for manual review. We believe that this evaluation framework might also benefit other researchers interested in tag-based index of clinical eligibility text.

Most alternative evaluation strategies require a manual rating process. A possible approach is the expert-based systematic review, which however could be subjective and thus error-prone and time-consuming. Similarly, other applications, such as clinical trial clustering based on eligibility similarity, could be considered for evaluation, but these too require a gold standard to measure the result quality (i.e., relevant judgments on the trial similarity).

Table 4
Example of the five most frequent tags ($R = 3\%$, i.e., "cvocab-3.0") and UMLS-only annotations for 3 query-conditions.

cvocab-3.0	UMLS-only
genetic diseases – *returned* 5,149 *clinical trials*	
1. ec:gravidity	1. ec:gravidity
2. ec:hypersensitivity	2. ic:female
3. ec:malignant neoplasms	3. ic:male gender
4. ic:contraceptive methods	4. ec:hypersensitivity
5. ec:heart failure	5. ec:kidney
cerebrovascular disorders – *returned* 2,055 *clinical trials*	
1. ic:cerebrovascular accident	1. ic:cerebrovascular accident
2. ec:cerebrovascular accident	2. ec:allergy
3. ic:NOT coronary artery diseases	3. ec:cerebrovascular accident
4. ic:cognitive therapy	4. ec:brain
5. ec:NOT hemorrhage	5. ic:ischemic stroke
heart diseases – *returned* 9,163 *clinical trials*	
1. ec:myocardial infarction	1. ic:heart
2. ic:coronary artery diseases	2. ec:heart
3. ec:heart failure	3. ec:myocardial infarction
4. ec:gravidity	4. ic:coronary artery diseases
5. ic:angina pectoris	5. ec:heart ventricle – reduction

Table 5
Search result filtering in terms of number of tags a user must click (from a set of 10 tags) to reduce the number of trials to a specified upper limit, averaged over 50 query conditions. For each query condition, we ran 500 distinct simulations based on random tag selections. The pool of tags to be chosen dynamically changed at every user selection with respect to the remaining trials. The controlled vocabulary was mined using $R = 3$, i.e., "cvocab-3.0".

Filtered Trial Limit	Algorithm	Tag click number		
		Min	Max	Mean
3	UMLS-only	7.46	22.78	14.15
	cvocab-3.0	**5.26***	**17.26***	**11.18***
5	UMLS-only	7.04	21.00	12.63
	cvocab-3.0	**5.86***	**15.90***	**10.08**
10	UMLS-only	5.98	17.48	10.25
	cvocab-3.0	**4.94***	**13.48***	**8.32***
20	UMLS-only	4.98	13.78	8.11
	cvocab-3.0	**4.16***	**10.88***	**6.71***
50	UMLS-only	3.74	9.56	5.87
	cvocab-3.0	**3.20***	**7.94***	**4.86***

* Indicates that the difference between the results is statistically significant.

A tag-based index of eligibility text can benefit different clinical trial technologies, in particular those oriented to search and browsing. As shown empirically, tags can be used to reduce the results of a free-text search. In this applicative scenario, tags could be provided to the users as either tag lists beside the search results (as envisioned in Section 2.3.2), tag clouds, or progressive questions. Regardless of the presentation format, tags summarizing eligibility criteria could help users to overcome the difficulties of querying conventional clinical trial search engines. Another potential application is efficient clustering of clinical trials to assist similarity searches (i.e., find trials with eligibility criteria similar to a query trial). This could benefit investigators designing new trials and seeking guidance from similar trials or patients looking for additional trials related to those for which they might be eligible.

Last, the controlled vocabulary could be paired with techniques that process eligibility criteria independently (e.g., [11]). For example, in a search scenario, the frequent tags could be exploited for initial information filtering, whereas a more detailed representation adding a second degree of granularity could be used to suggest details more specific to the result context.

This study has a few limitations. The proposed methodology aims at establishing a general framework to index clinical eligibility text with frequently used tags. Yet, the tag-mining step can probably be improved in several ways. For example, given the generality of the tags, unsupervised approaches for topic distribution could be exploited to leverage and integrate the controlled vocabularies (e.g., [26,41]). Moreover, tags are mined without considering their interrelationships. However, having a hierarchy of the tags would provide the user with information that might be more or less precise (e.g., "breast carcinoma" vs. "carcinoma") depending on the application. Given the limited size of the tag set, a manually created hierarchy could lead to the best results. Existing algorithms for automatic learning concept hierarchies from text (e.g., [42,43]) might be useful as well and should be considered in the future. Besides exploring the directions mentioned above, future research may also consider the application of the proposed methodology to mine tags from other types of medical texts, such as scholarly journals and clinical notes.

5. Conclusion

This paper contributes a method for detecting frequent tags from clinical trial eligibility text and a novel framework for

evaluating it. We conclude that frequent tags mined from eligibility criteria in our unsupervised fashion serve as an effective and efficient index for clinical trials, which can potentially benefit different tasks, such as search and clustering. Of note, we successfully applied the index for reducing the results returned by a state-of-the-art clinical trial search engine.

Funding

The authors are supported by Grant R01LM009886 (PI: Weng) from the National Library of Medicine and Grant UL1 TR000040 (PI: Ginsberg) funded through the National Center for Advancing Translational Sciences.

Acknowledgments

The authors would like to thank the editor and reviewers for their constructive and insightful comments, as well as Vojtech Huser, James J. Cimino, Ida Sim, and Simona Carini for helpful discussions and support.

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.2013.08.012.

References

[1] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(5):760–72.
[2] Friedman C, Rindflesch TC, Corn M. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform; 2013.
[3] De Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I. Automated information extraction of key trial design elements from clinical trial publications. AMIA Annu Symp Proc; 2008. p. 141–5.
[4] Hernandez ME, Carini S, Storey MA, Sim I. An interactive tool for visualizing design heterogeneity in clinical trials. AMIA Annu Symp Proc; 2008. p. 298–302.
[5] Kiritchenko S, De Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak 2010;10:56.
[6] Chung GY. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inform Decis Mak 2009;9:10.
[7] Paek H, Kogan Y, Thomas P, Codish S, Krauthammer M. Shallow semantic parsing of randomized controlled trial reports. AMIA Annu Symp Proc, 2006: p. 604–8.
[8] Xu R, Supekar K, Huang Y, Das A, Garber A. Combining text classification and Hidden Markov Modeling techniques for categorizing sentences in randomized clinical trial abstracts. AMIA Annu Symp Proc; 2006. p. 824–8.
[9] Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J Biomed Inform 2013;46(1):33–9.
[10] Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011;18(Suppl 1):i116–24.
[11] 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 2011;44(2):239–50.
[12] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993;32(4):281–91.
[13] Korkontzelos I, Mu T, Ananiadou S. ASCOT: a text mining-based Web-service for efficient search and assisted creation of clinical trials. BMC Med Inform Decis Mak 2012;12(Suppl. 1):S3.
[14] *ClinicalTrials.gov*. April, 2013; Available from: http://www.clinicaltrials.gov/.
[15] Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.
[16] Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010;43(3):451–67.
[17] Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. AMIA Summits Transl Sci Proc 2010;2010:46–50.
[18] Chi EH and Mytkowicz, T., Understanding the efficiency of social tagging systems using information theory. ACM Hypertext, 2008: p. 81–88.
[19] Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human system communication. Commun ACM 1987;30(11):964–71.
[20] Furnas GW, Fake C, Von Ahn L, Schachter J, Golder S, Fox K, et al. Why do tagging systems work? ACM CHI; 2006. p. 36–9.
[21] Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging. Proc Natl Acad Sci USA 2007;104(5):1461–4.
[22] Lee-Smeltzer KH. Finding the needle: controlled vocabularies, resource discovery, and Dublin Core. Libr Collect Acquis 2000;24(2):205–15.
[23] Lancaster FW. Vocabulary control for information retrieval. Washington: Information Resources Press; 1972.
[24] Julien CA, Tirilly P, Leide JE, Guastavino C. Constructing a true LCSH tree of a science and engineering collection. J Am Soc Inf Sci Tec 2012;63(12):2405–18.
[25] Rindflesch TC, Aronson AR. Semantic processing in information retrieval. Annu Symp Comput Appl, Med Care; 1993. p. 611–5.
[26] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3(4–5):993–1022.
[27] Shen J, Wang M, Yan S, Hua XS. Multimedia tagging: past, present and future. ACM Multimedia 2011:639–40.
[28] Levy M, Sandler M. Music information retrieval using social tags and audio. IEEE Trans Multimedia 2009;11(3):383–95.
[29] French JC, Powell AL, Gey F, Perelman N. Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. ACM CIKM; 2001. p. 199–206.
[30] Ruthven I. Interactive information retrieval. Annu Rev Inf Sci Tech 2008;42:43–91.
[31] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press; 2008. p. 482.
[32] Bird S. NLTK: the natural language toolkit. COLING/ACL on interactive presentation sessions; 2006. p. 69–72.
[33] Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, Mcnaught J, Ananiadou S, et al. Developing a robust part-of-speech tagger for biomedical text. Adv Inform 2005;3746:382–92.
[34] Luo Z, Johnson SB, Weng C. Semi-automatically inducing semantic classes of clinical research eligibility criteria using UMLS and hierarchical clustering. AMIA Annu Symp Proc 2010;2010:487–91.
[35] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. Int J Digital Libraries 2000;3(2):115.
[36] Korkontzelos I, Klapaftis LP, Manandhar S. Reviewing and evaluating Automatic Term Recognition techniques. Advances NLP; 2008. p. 248–59.
[37] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Proc Man 1988;24(5):513–24.
[38] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
[39] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
[40] Patel CO, Garg V, Khan SA. What do patients search for when seeking clinical trial information online? AMIA Ann Symp Proc 2010;2010:597–601.
[41] Perotte A, Bartlett N, Elhadad N, Wood F. Hierarchically supervised latent Dirichlet allocation. NIPS 2011:2609–17.
[42] Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform 2011;44(1):163–79.
[43] Shamsfard M, Barforoush AA. Learning ontologies from natural language texts. Int J Hum–Comput Stud 2004;60(1):17–63.