Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Clustering clinical trials with similar eligibility criteria features

Tianyong Hao [a], Alexander Rusanov [b], Mary Regina Boland [a], Chunhua Weng [a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, United States
[b] Department of Anesthesiology, Columbia University, New York, NY, United States

ARTICLE INFO

Article history:
Received 13 July 2013
Accepted 24 January 2014
Available online xxxx

Keywords:
Medical informatics
Clinical trial
Cluster analysis

ABSTRACT

Objectives: To automatically identify and cluster clinical trials with similar eligibility features.
Methods: Using the public repository ClinicalTrials.gov as the data source, we extracted semantic features from the eligibility criteria text of all clinical trials and constructed a trial-feature matrix. We calculated the pairwise similarities for all clinical trials based on their eligibility features. For all trials, by selecting one trial as the center each time, we identified trials whose similarities to the central trial were greater than or equal to a predefined threshold and constructed center-based clusters. Then we identified unique trial sets with distinctive trial membership compositions from center-based clusters by disregarding their structural information.
Results: From the 145,745 clinical trials on ClinicalTrials.gov, we extracted 5,508,491 semantic features. Of these, 459,936 were unique and 160,951 were shared by at least one pair of trials. Crowdsourcing the cluster evaluation using Amazon Mechanical Turk (MTurk), we identified the optimal similarity threshold, 0.9. Using this threshold, we generated 8806 center-based clusters. Evaluation of a sample of the clusters by MTurk resulted in a mean score 4.331 ± 0.796 on a scale of 1–5 (5 indicating "strongly agree that the trials in the cluster are similar").
Conclusions: We contribute an automated approach to clustering clinical trials with similar eligibility features. This approach can be potentially useful for investigating knowledge reuse patterns in clinical trial eligibility criteria designs and for improving clinical trial recruitment. We also contribute an effective crowdsourcing method for evaluating informatics interventions.

1. Introduction

The past few decades have witnessed heightened expectations for transparency in scientific research. Vast troves of clinical and research data have been digitized and made publicly available by governmental agencies, corporations, and private organizations. The availability of these data has generated a great need for innovative methods that leverage such Big Data to improve healthcare delivery and to accelerate clinical research [1]. However, gaining meaningful insights from this Big Data is fraught with challenges.

For example, in one of the largest clinical trial repositories, ClinicalTrials.gov[1], there are more than 145,745 clinical trials as of May 2013. Information overload is an unsolved problem when searching for relevant clinical trials in this repository. Methods have been developed to address this problem [2–8], such as web-based EmergingMed[2], SearchClinicalTrials.org[3], and the UK Clinical Trials Gateway[4], and mobile device-based NCITrials@NIH[5], ClinicalTrials Mobile[6], and ClinicalTrials.app[7]. Although these methods are helpful in narrowing the search for trials, they require users to come up with effective queries, which can be a difficult task given the complexity of eligibility criteria [9] and of medical terminologies.

One alternative to clinical trial search based on a user query is case-based search, which identifies trials similar to an example trial. Such an approach can remove the burden for query formulation from the user and is deemed to be useful in multiple usage scenarios. For clinical trial volunteers, a trial for which they qualify but cannot join due to closed enrollment, geographic distance from the recruitment site, or other practical reasons, can serve as a

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, United States. Fax: +1 2123053302.
E-mail address: cw2384@columbia.edu (C. Weng).

[1] http://clinicaltrials.gov/.
[2] http://www.emergingmed.com.
[3] http://searchclinicaltrials.org/.
[4] http://www.ukctg.nihr.ac.uk.
[5] http://bethesdatrials.cancer.gov/app/.
[6] http://www.clinicaltrials.com/industry/clinicaltrials_mobile.htm.
[7] http://www.iphoneclinicaltrials.com/.

http://dx.doi.org/10.1016/j.jbi.2014.01.009
1532-0464/

starting point in the search for trials recruiting similar patients. For clinical trial investigators, case-based search might help identify colleagues recruiting similar patients for related diseases and inform the eligibility criteria design of a new trial. For meta-analysis researchers, this method can identify studies with similar eligibility features and help uncover knowledge reuse patterns among related studies or improve the efficiency of systematic reviews.

To support the aforementioned use cases, in this paper we present an automated approach to identifying clinical trials with similar eligibility criteria, across and within diseases, based on the similarity in semantic eligibility features. In the context here, a semantic feature is a clinically meaningful patient characteristic, such as a demographic characteristic, a symptom, a medication, or a diagnostic procedure, used to determine a volunteer's eligibility for a trial. It contains either one word, (e.g., "cardiomyopathy") or multiple words (e.g., "biopsy-proven invasive breast carcinoma") [8]. We focused on similarity measures at the concept level because as noted by Korkontzelos et al. [10], decreasing the length of lexical units, from sentences to phrases or tokens, can solve the sparsity problem in identifying eligibility criteria that are important for a particular study, though a potential tradeoff of this method is that unimportant functional words and phrases are more frequent than meaningful ones in the biomedical domain.

An important premise of our proposed approach is that numerical values in eligibility criteria, such as constants in expressions for age and laboratory results, are not necessary considerations for determining eligibility criteria similarity at the concept level. For example, our method does not differentiate "Age: 50–65" from "Ages: 10–17", or differentiate "HbA1C > 6.5" from "HbA1C < 6.5". For clinical trials with a small number of eligibility criteria features, this limitation might result in incorrect clustering of trials with semantically different eligibility criteria. However, eligibility criteria are rich in features, with an average of 38.5 features per trial on ClinicalTrials.gov. When two trials are deemed similar using our method, a majority of eligibility features must match; therefore, the differences in the attributes associated with any feature have minimal influence on overall trial similarity. In other words, it is unlikely that a trial recruiting patients aged 50–65 would match a trial recruiting patients aged 10–17 in all other eligibility features. The presence of many features helps our method distinguish trials recruiting different target populations despite the disregard for numerical values in any given feature.

The rest of this paper is organized as follows. We first describe our processes for semantic feature extraction and trial clustering based on feature similarities. Then we introduce a crowdsourcing method for evaluating the similarities of the resulting clusters using Amazon's Mechanical Turk. On this basis, we present the performance metrics for this method.

2. Materials and methods

Fig. 1 illustrates the methodology framework. We obtained the free-text eligibility criteria for all registered trials (N = 145,745 as of September 2013) listed on ClinicalTrials.gov. We then used the Unified Medical Language System (UMLS) Metathesaurus to recognize all biomedical concepts, which serve as the semantic features, and assigned a suitable UMLS semantic type for each of them. On this basis, we constructed a trial-feature matrix to cluster trials using pairwise similarity. Our design rationale and implementation details are further provided below.

2.1. Extracting semantic features

Although UMLS's parser, MetaMap, is the mostly widely used parser for biomedical concept recognition, we chose to develop our own concept recognition algorithm to avoid the limitations in MetaMap output as identified by Luo et al. [11]. For example, the criterion "Patients with complications such as serious cardiac, renal and hepatic disorders" was parsed by MetaMap Transfer (MMTx) as {Patients |Patient or Disabled Group} {with complications |Pathologic Function} {such as serious cardiac, renal |Idea or Concept} {and|} {hepatic disorders |Disease or Syndrome}. These results were not granular enough. Additionally, MMTx returned the phrase "such as serious cardiac, renal" as a single constituent, which was problematic.

Excluding trials with no or non-informative text, such as "please contact site for information" (e.g., NCT00000221), for each remaining trial listed on ClinicalTrials.gov, we extracted its eligibility criteria text and preprocessed it by removing white spaces. We then performed sentence boundary detection for feature extraction. We first tried commonly used sentence boundary detectors such as the NLTK *sent_tokenize* function [12] but they alone were ineffective due to the variability in the formatting of the criteria text, e.g., some sentences lacked boundary identifiers or used different bullet symbols as separators. Therefore, we first applied bullet symbols or numbers as splitting identifiers and then applied NLTK on the remaining text chunks. For example, the eligibility criteria text of trial NCT00401219 contained both bullet symbols and a sentence boundary identifier. Therefore, the text was first split using the bullet symbols and then chunked using the identifiers. We improved the NLTK function to handle words like "e.g." and "etc.", which were incorrectly separated by the period symbol.

We identified terms using a syntactic-tree analysis after part-of-speech (POS) tagging. This method was better than an n-gram-based method for pair-wise similarity calculation because the latter generated overlapping terms, which could lead to overestimation of similarity, or omitted candidate features that were not sufficiently frequent, which could cause underestimation of similarity. After testing several parsers, we utilized an open library[8] to generate syntactic trees based on POS tags labeled by NLTK. Using predefined parsing rules, we traversed the syntactic trees and extracted phrases using NLTK WordNet lemmatizer and stemming modules. For example, from the sentence "a multi-center study of the validity" the algorithm would generate the following syntactic tree: {(S a/DT (NP (NBAR multi-/NN center/NN study/NN)) of/IN the/DT (NP (NBAR validity/NN)))}. From the tree, two noun phrases were extracted using NBAR tag (one predefined rule): "multi-center study" and "validity".

Being candidate semantic features, all terms were looked up in the UMLS using normalized substring matching rather than exact string matching. The advantage of this fuzzy term mapping strategy is that partial or complete term could be mapped to a UMLS concept. For example, we can extract a semantic feature "serious hypertensive disease", where "hypertensive disease" is a UMLS concept, from term "serious systemic arterial hypertension" even if the latter as a whole does not exist in UMLS. For a term p, each word w was assigned as a start point for substring generation after checking with a list of English stop words, a list of non-preferred POS tags, and a list of non-preferred semantic types. For a start point w_i, substring from w_i to an end point word w_j ($i < j < length(p)$, $w_j \in p$) was generated as s_{ij} with j from reverse direction (largest substring first). s_{ij} was then processed through UTF decoding, word normalization (by NLTK WordNet Lemmatizer and word case modifier), word checking (on punctuations, numeric, English stop words, and medical related stop words), and acronym checking to match with UMLS concepts. If there was no match, it moved to substring $s_{i(j-1)}$ for next matching until $j = i + 1$. Once there was a match, the start point w_i was set to point w_j (skip the start

[8] https://gist.github.com/alexbowe/879414.