Journal of Biomedical Informatics 52 (2014) 112–120

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 1 February 2014

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.

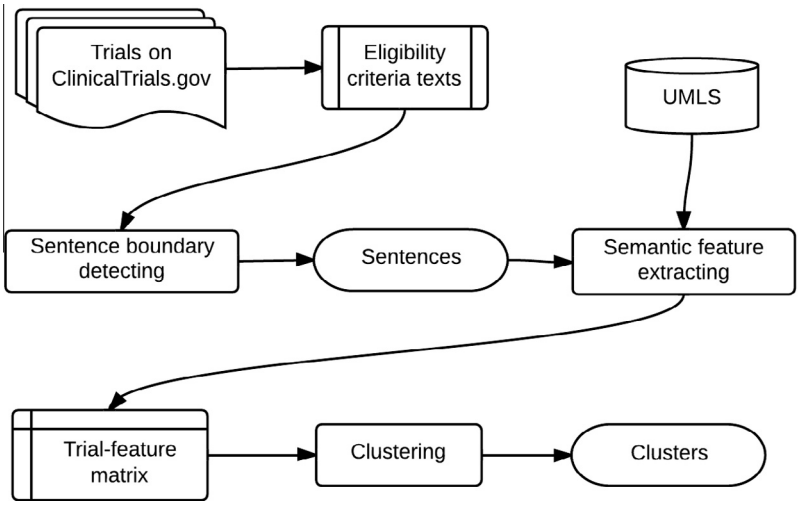


Fig. 1. The framework for automatically identifying clinical trial clusters based on eligibility criteria similarity.

points between w_i and w_j); otherwise, i was set to $i+1$ for next round of matching until i equals $length(p)-1$.

A term can be associated with multiple UMLS concepts with different semantic types. We performed concept disambiguation using a set of predefined semantic preference rules [13]. For example, a term "pregnancy test negative" was associated with two UMLS concepts, one being "pregnancy test negative" of the semantic type "Laboratory or Test Result" and the other being "reported negative home pregnancy test" of the semantic type "Finding". In the UMLS Semantic Network, "Laboratory or Test Result" is a subtype of "Finding". Hence, the more specific concept "pregnancy test negative" was assigned to this term.

We did not distinguish between inclusion and exclusion criteria for semantic feature extraction because not all trials, such as trial NCT00000114, had separate inclusion and exclusion criteria sections. The extracted unique semantic features were used for generating a trial-feature matrix and for calculating trial similarity. In the trial-feature matrix, each row corresponds to a set of semantic features from a certain trial and each column shows a certain feature existing in different trials. If a trial t_i contains semantic feature sf_m, row i and column m was recorded as 1, otherwise as 0.

2.2. Determining pairwise similarity

There are plenty of measures of semantic similarity between concepts used in Natural Language Processing [14–18]. Pedersen et al. [19] presented the adaptation of six domain-independent measures and showed that an ontology-independent measure was most effective. Particularly for text clustering, Huang [20] compared 5 widely used similarity measures on 7 datasets and showed that the Jaccard similarity coefficient achieved best score on a well-studied dataset containing scientific papers from four sources. We adopted the Jaccard similarity coefficient for calculating pairwise similarity as it can assess both similarity and diversity [21]. For a collection of trials $T=\{t_1, t_i, \ldots t_j, \ldots t_k\}$ containing k trials, the pairwise similarity $Simi$ of any two trials t_i and t_j was calculated as follows:

$$Simi(t_i,t_j)=\begin{cases}0, & |SF(t_i)\ |\text{or}|\ SF(t_j)|=0\\ \frac{|SF(t_i)\cap SF(t_j)|}{|SF(t_i)\cup SF(t_j)|}, & \text{otherwise}\end{cases}$$

$SF(t_i)$ and $SF(t_j)$ are two sets of semantic features corresponding to t_i and t_i, respectively. If either $SF(t_i)$ or $SF(t_j)$ contains no semantic features, then the similarity is recorded as 0. Otherwise, it is calculated as the number of shared features ($SF(t_i)\cap SF(t_j)$) divided by the number of features in the union ($SF(t_i)\cup SF(t_j)$).

Due to the large number of trials and the large volume of semantic features, calculating the similarity between every possible pair of trials would be computationally intensive. To improve efficiency, we first ranked all trials by their counts of semantic features. Trial pairs with a large difference in their feature counts were discarded, since the large count gap would lead to a low similarity as the shared features were too few compared to the union features. We defined two rules to select similar trial pairs: $|SF(t_i)|>2^*|SF(t_j)|$ and $|SF(t_i)|<|SF(t_j)|/2$, indicating that trial pairs with similarity below 0.5 were considered to have unsatisfactory similarity and discarded.

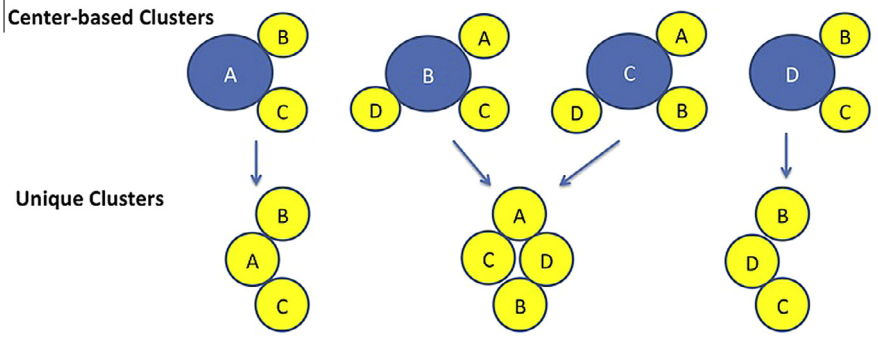
2.3. Clustering trials

There are many clustering models based on connectivity, centroid, distribution, and so on [22–26]. Methods such as K-means and hierarchical clustering were also assessed. Inspired by the known algorithm Nearest Neighbor Search (NNS) [27], for each unique trial, we constructed a cluster by using this trial as the center and by identifying all its near neighbors. To measure nearness, we calculated the distance between each neighbor and the central trial using the following formula: distance = 1-similarity, where the similarity was the previously calculated pairwise similarity between the trial pair. For any central trial x, only trials whose similarities to the trial were greater than or equal to a predefined similarity threshold δ were included in the cluster centered on x. Therefore, we refer to these clusters as center-based clusters. Connected center-based clusters were merged to form similarity-based clinical trial network using the DBScan algorithm [28]. In order to facilitate visualization and statistical analyses of clusters, we removed structural information (i.e., center vs. neighbor) and identified trial sets with distinctive membership compositions from all center-based clusters and named these sets as unique clusters. Fig. 2 illustrates the center-based and unique clusters for example trials.

2.4. Evaluation design

The Amazon Mechanical Turk (MTurk[9]) is an online crowdsourcing platform that enables human workers to perform human intelligence task (HIT) [29]. It has been shown to be effective for similarity evaluation. For example, Snow et al. [30] reported high agreement between MTurk non-expert annotations and existing gold standards

[9] https://www.mturk.com/.



Similarity Matrix of four trials A,B,C,D; 1=identical; 0= completely different; our threshold for forming clusters is 0.7

	A	B	C	D
A	1	.8	.9	.6
B	.8	1	.8	.9
C	.9	.8	1	.8
D	.6	.9	.8	1

Fig. 2. Center-based clusters and unique clusters constructed from four example trials.

provided by experts on five natural language tasks including a word-similarity evaluation task. MTurk has also been used for evaluating biomedical informatics research. For instance, Maclean and Heer [31] presented crowdsourcing patient-authored text medical word identification tasks to MTurk non-experts and achieved results that were comparable in quality to those achieved by medical experts. Therefore, we used MTurk for evaluating the similarity within our clusters.

Regarding the limitations of using MTurk, Mason and Suri [32] pointed out some of these in their systematic review of MTurk with respect to efficiency improvement, quality assurance, security, ethics and privacy. They also suggested possible techniques to control submission quality, which were partially applied in our evaluation. Lee [33] summarized the 9 benefits and 4 limitations of using MTurk, including "instructions needed clarification over several tests". Following these suggestions, we designed a two-phase evaluation and defined answer rejection rules to ensure data quality from MTurk. In the first phase, we determined the optimal similarity threshold. In the second phase, we evaluated the clusters generated using the optimal threshold.

Instruction:

- For each text cluster you will be shown 2- 3 texts. The first one will be labeled "Query" and the others will be labeled "Candidate".
- You will compare each of "Candidate" text with "Query" text then select answer choice based on overall similarity. If all the candidate texts are very similar to Query, then select "Strongly Agree"; if candidate texts are similar to Query, select "Agree"; if somewhat similar but not all candidate are similar to Query, then select "Neither agree nor disagree", and so on.
- The judgment is based on semantic (overall meaning). Please disregard text formats and specific numbers, such as age and specific laboratory values.
- As an example: "Patients older than 55 with a history of heart disease" and "Patients older than 65 with a history of heart disease" are very similar because the only difference is the age (should ignore specific numbers). By contrast: "Patients older than 55 with a history of heart disease" is different from "Patients older than 55 with pancreatic cancer" though they shared "Patients" (should select "Disagree" option).
- **IMPORTANT:** One text you read will contain the phrase "waldo is hiding here" for quality control. When you see it, chose the option "Found waldo". Failed detection of the option will result in rejection of your submission. Missing of answers or submission from same worker can also lead to rejection.

This is a good cluster containing similar texts. Do you agree?

Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo
Click to view cluster	○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree ○ Found waldo

Fig. 3. The user interface of HIT designed for MTurks for cluster evaluation.

2.4.1. Phase I – threshold determination

Based on empirical results we selected three candidate thresholds for optimization: i.e., 0.7, 0.8 and 0.9. We generated trial pairs having a similarity equal to each threshold. To obtain a representative sample, we plotted the distribution of average word counts per pair of texts as a box-plot. We then selected a total of 20 pairs of texts, 5 from each quartile, from this distribution curve for each of the three thresholds. In total we generated 60 pairs of texts for evaluation.

The evaluation set was published as a HIT on the MTurk website with a reward of $1.20 offered for completion of the entire HIT. For each pair of texts, workers were asked whether they agreed with the statement "The texts in the pair are similar." Workers were instructed to avoid differences in actual numbers (e.g., age, cutoffs for laboratory values, etc.) and to focus only on broad criteria concepts when calculating similarity. Available answer choices were "Strongly agree", "Agree", "Neither agree nor disagree", "Disagree", and "Strongly disagree". To quantify the mean and standard deviation (SD) of the answers for each candidate threshold, we mapped each of the above choices to numbers 5, 4, 3, 2, and 1, respectively. The optimal similarity threshold was selected by comparing the means and SDs of the three candidate thresholds.

To assure that workers were paying attention and not just randomly selecting answers, and to filter out "spammers" or "bots" [32], we inserted the hidden phrase "Waldo is hiding here" into one of the comparison text pairs. In the instructions, workers were informed that this phrase would appear in some of their eligibility criteria texts. They were instructed to select the answer choice "Found Waldo" upon seeing the phrase in the text rather than any of the choices pertaining to similarity. The instructions also explained that workers who failed to discriminate between pairs with or without the hidden phrase would result in rejection, without payment, of their work and a negative review of their performance posted to their profile. The entire HIT for any worker who failed the hidden phrase identification was deemed invalid and excluded. We continued recruiting workers until we got a total of 10 valid, completed HITs. This resulted in 10 evaluations by unique workers of the entire 60-pair set.

2.4.2. Phase II – cluster evaluation

Using the optimal threshold from Phase I, we generated unique clusters. To ensure a fair sampling of cluster sizes, the distribution of cluster sizes, measured in the number of nodes, was presented as a box plot. An evaluation set of 40 clusters, 10 selected from each quartile, was then generated. The evaluation set was published as a HIT on the MTurk website with a reward of $1.20 offered for completion of the entire HIT. For each cluster, workers were asked to rate if every cluster contained similar texts using a 5-Likert scale. As in Phase I, workers were instructed to ignore the differences in attribute values, e.g., age, cutoffs for laboratory values, and to focus only on inclusion eligibility concepts when determining similarity for each trial pair.

The user interface design for the HIT is shown in Fig. 3. Clicking "click to view cluster" opens a page (Fig. 4) containing eligibility criteria texts from ClinicalTrials.gov for comparison.

We employed the same hidden phrase identification method for quality control as described in Phase I. Additionally, by performing the HIT ourselves, we estimated that it would be difficult to perform an accurate assessment of 40 clusters in less than 20 min. Accordingly we excluded the entire HIT if it was completed in less than 20 min. We continued recruiting workers until we got a total of 20 (double the number in Phase I) valid and complete HITs. This resulted in 20 evaluations by unique workers for the entire 40 cluster set.

3. Results

3.1. Semantic features and clusters

We extracted all 145,745 clinical trials present on ClinicalTrials.gov as of 05/17/2013. After excluding the trials whose eligibility criteria text section were missing or contained only the phrase

This page contains a total of 3 texts for comparison.

Query:
Inclusion Criteria:
- Smokers smoking at least 10 cigarettes/day as a mean within the 2 months preceding the screening visit
- Motivated to quit with a score greater than or equal to 6 on the ten-point Motivation Scale

Exclusion Criteria:
- Non tobacco cigarettes consumption
- Chronic use of marijuana
- Pregnancy, breastfeeding
- Any clinically significant disease that might interfere with the efficacy or safety evaluation of the study drug
- Concomitant use of drugs as an aid to smoking cessation or that might induce weight change

Candidate1:
Inclusion Criteria:
- Smokers smoking at least 10 cigarettes/day as a mean within the 2 months preceding the screening visit
- Motivated to quit with a score greater than or equal to 6 on the ten-point Motivation Scale

Exclusion Criteria:
- Non tobacco cigarettes consumption
- Chronic use of marijuana
- Pregnancy, breastfeeding
- Any clinically significant disease that might interfere with the efficacy or safety evaluation of the study drug
- Concomitant use of drugs as an aid to smoking cessation or that might induce weight change

Candidate2:
Inclusion Criteria:
- Smokers smoking at least 15 cigarettes/day as a mean within the 2 months preceding the screening visit
- Motivated to quit with a score greater than or equal to 6 on the ten-point Motivation Scale

Exclusion Criteria:
- non tobacco cigarettes consumption
- chronic use of marijuana
- pregnancy
- breastfeeding
- any clinically significant disease that might interfere with the efficacy or safety evaluation of the study drug
- Concomitant use of drugs as an aid to smoking cessation or that might induce weight change

Fig. 4. The eligibility criteria of trials in an example cluster for comparison by workers.

"Please contact site for information", 142,948 remained and were used as our dataset. We identified 2,770,746 sentences, from which 5,508,491 semantic features (459,936 unique) were extracted, with 38.5 features per trial on average. Of the unique features, 160,951 (34.99%) were shared by at least two trials. For instance, Table 1 aligns part of the semantic features extracted from trials "NCT00822978" and "NCT01034774". The two trials shared 30 features (marked as "Y" in "Shared" column) and had 3 different features: "iodine", "excessive alcohol consumption", and "illegal substance". As a result, their pairwise similarity is 0.91.

The percentage of unique features that are shared by at least two trials varies by the total number of trials used for feature extraction, as shown in Fig 5. With a sample of 40,000 trials, the percentage of shared features is 31.34%. Beyond this point, increasing the number of trials only slightly affects this percentage. In our dataset, 34.99% of unique features are shared across all trials. Therefore, these shared features could be used to aid standardization efforts for clinical trials eligibility criteria. The remaining 65.01% of features are unique to a given trial. These features could be useful for distinguishing trials but not useful for clustering similar trials.

After discarding trial pairs with similarity less than 0.5, a total of 386,992 pairs remained and were used for clustering. Using the optimal similarity threshold ($\delta = 0.9$), as determined by the MTurk evaluation, 8806 center-based and 3,614 unique clusters were generated. Cluster sizes ranged from 2 to 734 members. Table 2 shows the statistics for the first 9 sizes (2–10). Of the center-based clusters, most contained 2 (5680 clusters, 64.5%) or 3 (969 clusters, 11.0%) trials. Overall, the average center-based cluster size and unique cluster size are 65.34 and 2.85, respectively. There were 2157 (24.5%) center-based clusters and 314 unique clusters (8.69%) containing more than 3 trials, with the largest cluster containing 734 trials.

3.2. Evaluation results using MTurk

In Phase I of the MTurk based evaluation, 20 pairs of eligibility criteria texts were selected from each of three similarity thresholds: 0.7, 0.8, and 0.9. Distributions of text length (measured as average number words per pair) for each threshold (Table 3) were used to select a representative sample of text lengths. Five pairs of texts were randomly selected from each quartile range. For example, for threshold of 0.7, the minimum, 1st quartile, median, 3rd quartile, and maximum are 26, 69, 108, 294.5, and 845, respectively, with the corresponding ranges being 26–69, 69–108, 108–294.5, 294.5–845. As a result, a total of 20 pairs of texts for the threshold were selected. Of note, even though it appears that there is a correlation between threshold and text length based on the three thresholds presented here, such a correlation was not observed on a bigger dataset of ten thresholds and their corresponding mean text lengths.

Table 1
Parts of semantic features extracted from trials NCT00822978 and NCT01034774.

NCT00822978	NCT01034774	Shared
Hypertensive disease	Hypertensive disease	Y
Deafness	Deafness	Y
SGCG gene	SGCG gene	Y
Kidney problem	Kidney problem	Y
Hearing	Hearing	Y
Aminoglycosides	Aminoglycosides	Y
Family history	Family history	Y
Seizures	Seizures	Y
Asthma	Asthma	Y
Recent blood donor	Recent blood donor	Y
Ear structure	Ear structure	Y
Previous injury	Previous injury	Y
Diabetes	Diabetes	Y
Continue medical condition	Continue medical condition	Y
Smoker	Smoker	Y
Heart diseases	Heart diseases	Y
Iodine	Excessive alcohol consumption	N
	Illegal substance	N

Of the 13 submissions received from 13 unique workers, three were rejected because the workers failed to identify the text containing "waldo". We accepted the remaining 10 submissions for a total expense of $13.20 ($1.20/submission x 10 submissions + $1.20 commission fee). The mean and standard deviation (SD) of the scores for each threshold are shown in Table 4. The average of mean and standard deviation for all the thresholds are 3.72 and 1.08, respectively. We selected 0.9 as the optimum similarity threshold because this was the only threshold where the mean score was >4, which corresponds to the "Agree" choice.

In the Phase II of the evaluation, 20 unique workers evaluated 40 unique clusters. Ten clusters were randomly selected from each quartile range of the distribution of cluster sizes (measured as the number of trials in the cluster). The minimum, 1st quartile, median, 3rd quartile, and maximum cluster sizes were 2, 2, 2, 2, and 734, respectively. A total of 23 submissions were received, 20 of which were accepted. Three submissions were rejected either for failing to identify the text containing "waldo" or for completing the evaluation too quickly, a sign of lack of careful consideration. The average time spent by workers on this task was 29 min.

One of the accepted submissions was originally rejected due to failure to correctly identify the text containing "waldo" but, after review by the authors, was later accepted after this worker explained that he tried but failed to find the "waldo" text. As for the evaluation, the total cost for Phase II was $24 (=$1.20/submission x 20 submissions + $2.40 commission). The inter-rater reliability was calculated as 0.92, using Cronbach's Alpha[10]. The mean cluster quality score was 4.331 ± 0.796; i.e., the overall response to the statement "This is a good cluster" was between "Agree" (4) and "Strongly agree" (5).

3.3. Disease-specific network visualization

Our method can visualize results as trial networks. For example, by limiting our dataset to "Breast Cancer", we extracted 5309 trials present on ClinicalTrials.gov as of 05/20/2013. A total of 4844 trials were obtained after initial eligibility criteria filtering and 117,388 sentences were acquired after sentence boundary identification. From these, we extracted 255,614 semantic features and 98 center-based clusters (2.55 trials per cluster on average) using a similarity threshold of 0.9. A network diagram containing the clusters and unique features was automatically generated for visualization of the relatedness of the clusters (Fig. 6). The large orange nodes represent individual clinical trials, while small blue nodes represent unique semantic features. The links between one feature node and multiple trial nodes denote that the semantic feature is shared by the trials. The more semantic features shared by trials, the closer are the trials. For a semantic feature, the more trials it is connected to, the closer it is located in to the center of the network (i.e., the high density area).

3.4. A cluster-based clinical trial search interface

We used the similarity-based search to enhance the ClinicalTrials.gov search interface for clinical trials. A demonstration system for this design is available online[11]. After searching for trials using

[10] http://en.wikipedia.org/wiki/Cronbach's_alpha.
[11] http://columbiaelixr.appspot.com/cluster.

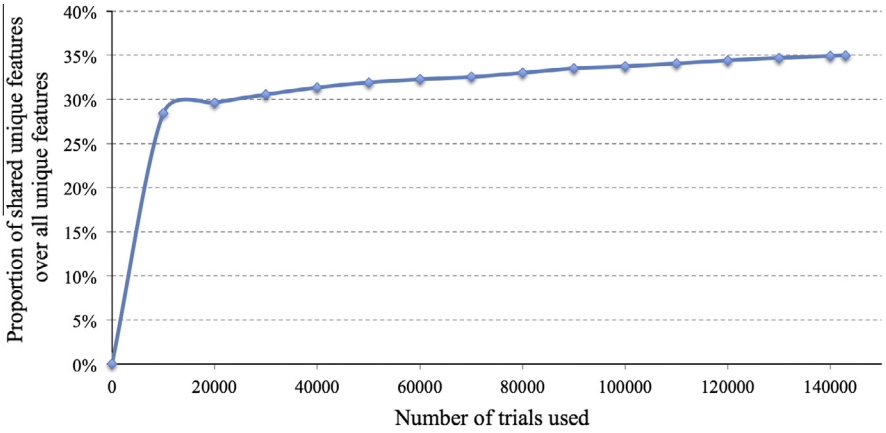


Fig. 5. Percentage of unique features shared by at least two trials as a function of total number of trials.

Table 2
The relationship between cluster size and number of clusters.

Cluster size	Number of clusters	
	Center-based	Unique
2	5680 (64.5%)	2910 (80.5%)
3	969 (11%)	390 (10.8%)
4	464 (5.3%)	146 (4%)
5	222 (2.5%)	61 (1.7%)
6	78 (0.9%)	22 (0.6%)
7	79 (0.9%)	16 (0.4%)
8	20 (0.2%)	6 (0.2%)
9	53 (0.6%)	13 (0.4%)
10	50 (0.6%)	11 (0.3%)

Table 3
The quartile distribution of eligibility criteria text length measured by the average number of words per trial pair.

δ	Min	1st Quart.	Median	3rd Quart.	Max	Mean
0.7	26	69.00	108.00	294.50	845	220.50
0.8	15	54.50	96.75	272.60	959	205.20
0.9	34	43.00	43.00	65.25	909	80.43

Table 4
The mean and standard deviation of MTurk similarity ratings at different thresholds.

Threshold	Mean	Standard deviation
0.7	3.35	1.20
0.8	3.81	0.97
0.9	4.00	1.07

standard search functions on ClinicalTrials.gov, users can select any trial to perform a cluster-based search and view trials with similar eligibility criteria features. Adjacent to each trial in the cluster was an indicator showing whether this trial belonged to another cluster. A user could choose to view that cluster. In this way, users would be able to explore the network of trials, whose eligibility criteria were similar to the initial trial of interest.

4. Discussion

We presented a method for clustering trials with similar eligibility criteria features to facilitate case-based clinical trial search. Nearly 35% of the features we extracted are shared by at least two trials. Our approach produced clusters of manageable sizes. Though more than 85% of the center-based clusters contained between 2 and 7 trials, the mean cluster size was 65.34. This effect is mostly due to the presence of one large 734-member cluster. A closer look at the individual trials comprising this cluster revealed that all trials had identical eligibility criteria texts comprised of only one sentence: "No eligibility criteria". Further analysis revealed that the eligibility criteria for these trials tended to be contained in other sections of the trial description. For some of these trials, diseases listed in the "Conditions" section (NCT00005154) were the only identifiable eligibility criteria. Since search by condition (a list of standardized elements) is already a search function of ClinicalTrials.gov, we did not include the condition section in our analysis of trial similarities. For a few trials, eligibility criteria were contained in the "Detailed Description" section (NCT00005444). This, combined with the presence of vast amounts of text pertaining to aspects of the study unrelated to eligibility contained within the "Detailed description" section, prompted us to exclude this section in our analysis. Finally, some trials contained no specific eligibility criteria anywhere in the trial description but did contain some general requirements, such as age and gender (NCT00005721). Since specific criteria are required to distinguish trials, we chose to exclude these trials for analysis.

Most clusters were either 2-member (64.5%) or 3-member (11%) clusters due to the use of a high similarity threshold (0.9). The MTurk workers lacked biomedical domain knowledge, which could potentially explain their preference for literal similarity with such a high threshold, though the advantage is that there was a high inter-rater reliability.

4.1. Limitations and future work

A recent publication showed that the trial summaries on ClinicalTrials.gov were more condensed than full-text protocols for 32 studies [34]. Therefore, similarity results based on the eligibility criteria text on ClinicalTrials.gov may not hold when full-text clinical protocols are used for all studies. However, our approach can be applied to assess protocol similarity. Since full-text clinical-trial protocols are usually not freely available (especially for ongoing studies), the current use of ClinicalTrials.gov as the data source is the best solution available at present.

The clusters we generated were meaningful and contained highly similar trials, which is the counterpart to high precision in information retrieval, but we may not have identified all such clusters, which is the counterpart to low recall in information retrieval. Recall may be improved by lowering the similarity threshold, which could decrease the proportion of 2- and 3-member clusters and increase the proportion of larger clusters, decreasing the degree of positive skew of cluster sizes. However, this improvement

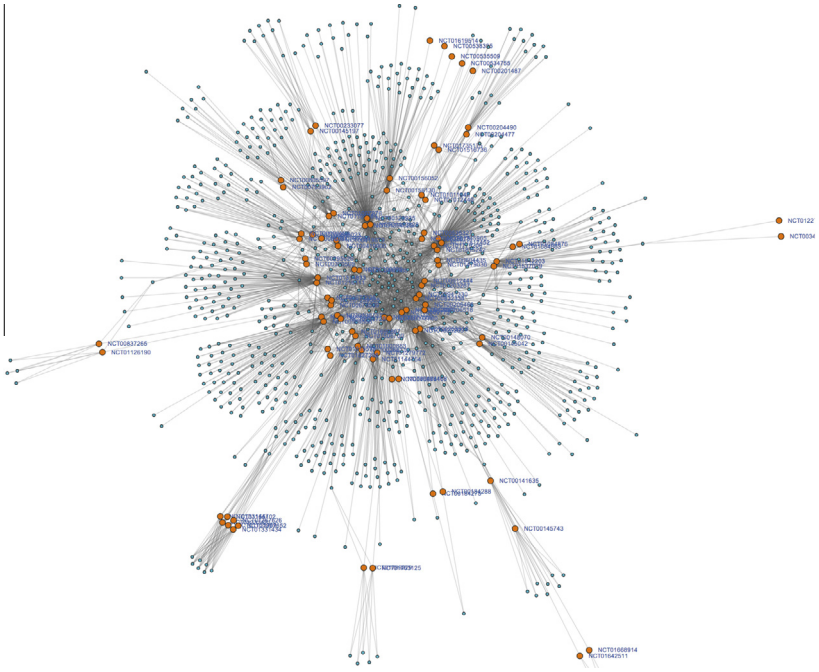


Fig. 6. The dynamically generated network diagram of all "Breast Cancer" related trials (http://columbiaelixr.appspot.com/static/cluster_breast_cancer.html).

would be at the expense of a decrease in precision (i.e., some of the clusters at the lower thresholds would not be as relevant with member trials not as similar to each other). Therefore, one must balance the tradeoff between precision and recall by optimizing similarity threshold determination. Our use of MTurk for threshold optimization proved to be effective, though our starting range included only three thresholds derived through empiric testing and we used a small subset of eligibility criteria texts. More candidate thresholds and larger amounts of eligibility criteria text could be used for threshold optimization.

Our methodology does not take into account the attributes of eligibility features, such as the numerical value ranges for quantitative features, e.g., Age > 75 years old. As discussed in the Methods section, the effect of this disregard for numerical data is minimized by the presence of a large number of features (i.e., an average of 38.5) per trial, which, when considered together, help distinguish trials with different target populations. Yet our method may still mistakenly cluster trials recruiting different populations together due to this limitation. Thus, eligibility for one trial in a cluster does not necessarily imply qualification for all trials in the cluster. We are working on a method for extracting and utilizing information from numerical values ranges to further improve trial similarity measures in the future.

As the increase of the similarity threshold tends to decrease the proportion of larger clusters, high thresholds may be too restrictive for some users by displaying too few highly trials that are similar to the center one. One solution to this problem is to optimize threshold determination. Another approach is to rank trials by their similarity scores and to pick relative higher ones in each cluster. We intend to assess the utility of these methods in the future.

Network visualization enables a global view of a similarity-based clinical trial network that can show how trials are connected by common medical concepts and how these concepts are shared by different trials, across or within diseases. Another future work idea is to combine similarity-based trial networks with other networks such as those based on geographic locations for helping elucidate relationships among clinical trials from multiple perspectives.

5. Conclusions

We developed an automated approach for clustering trials of similar eligibility criteria. Our evaluation confirmed the similarities within clinical trial clusters, which can be valuable for researchers and patients alike. Our experience with the Amazon Mechanical Turk confirmed that with careful data quality control, crowdsourcing was an effective approach to engage the public to participate in evaluations of biomedical informatics interventions. We hope our clinical trial search method can be integrated into clinical trial search engines to make clinical trial search easier for end users.

Author contributions

T.H. designed and implemented the method, performed evaluations, data analysis, and results interpretation, and led the writing of the paper. A.R. participated in the evaluation design, performed data analysis and interpretation, wrote the paper with other authors. M.R.B. participated in the evaluation design, results analysis, and paper drafting. C.W. conceptualized the idea, supervised its design, implementation, and evaluation, and wrote the paper.

Acknowledgments

This work was supported by National Library of Medicine Grants R01LM009886 (PI: Weng) and by National Center for Advancing Translational Sciences Grant UL1TR000040 (PI: Ginsberg).

References

[1] Bollier D. The promise and peril of big data. The Aspen Institute. ISBN: 0-89843-516-1; 2010.
[2] Miotto R, Jiang S, Weng C. ETACTS: a method for dynamically filtering clinical trial search results. J Biomed Inform December 2013;46(6):1060–7.
[3] Patel C, Gomadam K, Khan S, Garg V. TrialX: using semantic technologies to match patients to relevant clinical trials based on their personal health records. Web Semantics Sci Serv Agents World Wide Web 2010;8(4):342–7.

[4] Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. *Health Technol Assess* 2007;11(48). iii, ix-105.
[5] 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.
[6] 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.
[7] Milian K, Bucur A, Teije AT. Formalization of clinical trial eligibility criteria: evaluation of a pattern-based approach. In: Proc. of the IEEE international conference on bioinformatics and biomedicine; 2012. p. 1–4.
[8] Boland MR, Miotto R, Gao J, Weng C. Feasibility of feature-based indexing, clustering, and search of clinical trials. a case study of breast cancer trials from ClinicalTrials.gov. Methods Inf Med 2013;52(4).
[9] Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. AMIA Summits Transl Sci Proc 2010:46–50.
[10] 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.
[11] Luo Z, Duffy R, Johnson SB, Weng C. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. AMIA Summit Clin Res Inform 2010:26–31.
[12] Perkins J. Python text processing with NLTK 2.0 cookbook. Packt Publishing; 2010.
[13] Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. J Biomed Inform 2011;44(6):927–35.
[14] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. In: Anthony Cohn, editor. Proc. of the 21st national conference on Artificial intelligence (AAAI'06), vol 1; 2006. p. 775–80.
[15] Metzler D, Dumais S, Meek C. Similarity measures for short segments of text. Adv Inform Retrieval Lect Notes Comput Sci 2007;4425:16–27.
[16] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the int'l. conf. on research in computational linguistics; 1997. p. 19–33.
[17] Corley C, Mihalcea R. Measuring the semantic similarity of texts. In: Proc. of the ACL workshop on empirical modeling of semantic equivalence and entailment (EMSEE '05), Stroudsburg, PA, USA; 2005. p. 13–18.
[18] Hao T, Lu Z, Wang S, Zou T, Gu S, Liu W. Categorizing and ranking search engine's results by semantic similarity. In: Proc. of the 2nd international conference on ubiquitous information management and communication; 2008. p. 284–288.
[19] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(3):288–99.
[20] Huang A. Similarity measures for text document clustering. In: Proc. of the New Zealand computer science research student conference (NZCSRSC'08), Christchurch, New Zealand; 2007. p. 49–56.
[21] Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard coefficient for keywords similarity. In: Proc. of the international multi conference of engineers and computer scientists, vol I; 2013. p. 380–4.
[22] Hirano S, Sun X, Tsumoto S. Comparison of clustering methods for clinical databases. Inform Sci 2004;159(3–4):155–65.
[23] Fushman DD, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proc. of the 21st ACL; 2006. p. 841–8.
[24] Shash SF, Mollá D. Clustering of medical publications for evidence based medicine summarization. Artif Intell Med Lect Notes Comput Sci 2013;7885:305–9.
[25] Li PH, Wong WHS, Lee TL, et al. Relationship between autoantibody clustering and clinical subsets in SLE: cluster and association analyses in Hong Kong Chinese. Rheumatology 2013;52(2):337–45.
[26] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans Neural Netw May 2005;16(3):645–78.
[27] Beis JS, Lowe DG. Shape indexing using approximate nearest-neighbor search in high-dimensional spaces. In: Proc. of the 1997 conference on computer vision and pattern recognition (CVPR '97); 1997. p. 1000–6.
[28] Ester M, Kriegel H, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of the 2nd int. conf. knowledge discovery and data mining (KDD',96); 1996. p. 226–31.
[29] Alonso O, Baeza-Yates R. Design and implementation of relevance assessments using crowdsourcing. Adv Inform Retrieval Lect Notes Comput Sci 2011;6611:153–64.
[30] Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proc. of the conference on empirical methods in natural language processing (EMNLP '08). ACL; 2008. p. 254–63.
[31] Maclean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013:1. http://dx.doi.org/10.1136/amiajnl-2012-001110.
[32] Mason W, Suri S. Conducting behavioral research on amazon's mechanical Turk. Behav Res Methods 2011;44(1):1–23.
[33] Lee JH. Crowdsourcing music similarity judgments using mechanical Turk. In: Proc. of the 11th international society for music information retrieval conference (ISMIR 2010), Utrecht, Netherlands; 2010. p. 183–8.
[34] Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. J Biomed Inform 2013;46(5):805–13.