Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department

RECEIVED 17 April 2014
REVISED 1 July 2014
ACCEPTED 2 July 2014
PUBLISHED ONLINE FIRST 16 July 2014

Yizhao Ni[1], Stephanie Kennebeck[2], Judith W Dexheimer[1,2], Constance M McAneney[2], Huaxiu Tang[1], Todd Lingren[1], Qi Li[1], Haijun Zhai[1], Imre Solti[1,3]

ABSTRACT

Objectives (1) To develop an automated eligibility screening (ES) approach for clinical trials in an urban tertiary care pediatric emergency department (ED); (2) to assess the effectiveness of natural language processing (NLP), information extraction (IE), and machine learning (ML) techniques on real-world clinical data and trials.

Data and methods We collected eligibility criteria for 13 randomly selected, disease-specific clinical trials actively enrolling patients between January 1, 2010 and August 31, 2012. In parallel, we retrospectively selected data fields including demographics, laboratory data, and clinical notes from the electronic health record (EHR) to represent profiles of all 202795 patients visiting the ED during the same period. Leveraging NLP, IE, and ML technologies, the automated ES algorithms identified patients whose profiles matched the trial criteria to reduce the pool of candidates for staff screening. The performance was validated on both a physician-generated gold standard of trial–patient matches and a reference standard of historical trial–patient enrollment decisions, where workload, mean average precision (MAP), and recall were assessed.

Results Compared with the case without automation, the workload with automated ES was reduced by 92% on the gold standard set, with a MAP of 62.9%. The automated ES achieved a 450% increase in trial screening efficiency. The findings on the gold standard set were confirmed by large-scale evaluation on the reference set of trial–patient matches.

Discussion and conclusion By exploiting the text of trial criteria and the content of EHRs, we demonstrated that NLP-, IE-, and ML-based automated ES could successfully identify patients for clinical trials.

Key words: Automated Clinical Trial Eligibility Screening, Natural Language Processing, Information Extraction, Machine Learning

RESEARCH AND APPLICATIONS

OBJECTIVE

This study investigates use of state-of-the-art natural language processing (NLP), information extraction (IE), and machine learning (ML) technologies for automated clinical trial eligibility screening (ES). Our specific aims are to (1) develop an automated ES approach for clinical trials enrolling in the emergency department (ED) at an urban tertiary care pediatric hospital and (2) assess the effectiveness of NLP, IE, and ML techniques on real-world clinical data and trials. The overall objective is to develop a high-sensitivity automated ES approach to identify patients who meet eligibility characteristics of a trial to reduce the pool of potential candidates for staff screening. To assist the readers, a complete list of acronyms used in the paper is presented in the online supplementary appendix table A1.

BACKGROUND AND SIGNIFICANCE

Clinical trials are critical to the progress of medical science; however, awareness and access to clinical trials pose significant challenges to patients and physicians alike. Several reports have described the initial benefit of leveraging electronic health record (EHR) information to enhance trial recruitment.[1–3] However, in most circumstances, ES is still conducted manually. Manual screening typically requires a lengthy review of

Correspondence to Dr Yizhao Ni, Cincinnati Children's Hospital Medical Center, Department of Biomedical Informatics, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA; yizhao.ni@cchmc.org

©The Author 2014. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

For numbered affiliations see end of article.

patient records, a cumbersome process that creates a significant financial burden for an institution.[4] In a busy clinical care center, the task of screening patients for clinical trials without bias is labor-intensive.[5,6] For pharmaceutical companies, the clinical trial phase is the most expensive component of drug development, and any improvement in the efficiency of the recruitment process would be highly consequential.[7] For these reasons, identifying eligible participants automatically on the basis of EHR information promises great benefits for translational science. In recent years, EHR-based eligibility screening for clinical trials has become a very active area for research and development; as such, several automated/semiautomated systems have been developed.[8–19]

These ES systems either (1) manually design specific triggers for a clinical trial (eg, age, gender, and diagnosis) to identify eligible patients[8,9,17,18] or (2) automatically match patterns between clinical trial description and EHR content to identify eligible patient cohorts.[12–14,16,19] However, trial-specific triggers normally lack generalizability to new clinical trials. A recent study also demonstrated that alert fatigue affects physicians' responsiveness, possibly because of the low accuracy of the triggers.[20]

For automated trial–patient pattern matching, several methods have been proposed to standardize trial criteria.[21–27] These methods enable the creation of computable patterns from trial description (and patient EHRs) and effectively advance the development of automated ES systems.[12–14,16,19] The annual Text REtrieval Conference (TREC) recently included a medical record track dedicated to ES, where participants attempted to rank patients for a clinical query based on the content of physician notes.[28–36]

Despite these efforts, many barriers remain.[37,38] First, although automated ES systems should ideally be evaluated on real-world data, this goal is hindered by the lack of access to production EHRs.[37] Only a handful of studies provided evaluations on real-world trial–patient matching, and most of them focused on one specific clinical trial.[12–14,16,19,39] Even the TREC medical record track had to use synthetic clinical queries because of the lack of available real-world trial–patient matches. Second, not all automated ES algorithms proposed in the literature improve performance (eg, the term expansion algorithm proposed in the TREC track reports only worsens the performance).[29,31,35] Finally, few studies explicitly report trial screening efficiency with and without automated ES; additional study is required to fill this gap in our knowledge.

To address these barriers and evaluation gaps, we customized state-of-the-art NLP, IE, and ML technologies and developed an automated ES approach. Utilizing a physician-generated gold standard of trial–patient matches and a reference standard of historical trial–patient enrollment decisions on a diverse set of clinical trials, we will contribute to the body of knowledge of automated ES by (1) evaluating a state-of-the-art automated ES approach on real-world clinical data and trials, (2) further assessing the ES algorithms proposed in the TREC literature, and (3) comparing trial screening efficiency both with and without automated ES.

DATA AND METHODS

We focused on clinical trials for pediatric patients who visited the ED at Cincinnati Children's Hospital Medical Center between January 1, 2010 and August 31, 2012. The study was approved by the institutional review board. In current practice, enrollment decisions in the ED are made on a per patient visit basis. A clinical research coordinator matches current patients with the actively enrolling trials open on the patients' date of visit based on the information collected during the visit (eg, demographics and diagnosis). Therefore, in this study, we also treated each patient visit (referred to as an 'encounter') as the unit of analysis and made an eligibility prediction for each encounter.

Gold standard trial–patient matches

To create a gold standard set of trial–encounter matches for evaluation, we randomly sampled 5 days from the study period and collected all 1475 encounters and 13 disease-specific clinical trials (inclusion/exclusion criteria included one or more diseases) enrolling on those days. Owing to labor limitation, we further narrowed down the population by randomly selecting 600 encounters from the 1475 samples. The resulting 13 trials and the 600 encounters formed the dataset for the gold standard.

Two board-certified, pediatric emergency medicine physicians each with more than 10 years' experience independently reviewed all charts for each encounter and the criteria for each trial enrolling on the encounter date and made an eligibility decision for every trial–encounter pair. Differences between the physicians' decisions were resolved during adjudication sessions. Inter-annotator agreement between the two physicians was calculated using the F-value to define the agreement in gold standard.[40]

Historical trial–patient enrollment decisions

We collected all 239547 encounters in the ED during the study period. Of these, 36752 encounters between 00:00 and 8:00 and during holidays were excluded because of no clinical trial staffing in that time frame, providing a population of 202795 encounters. The 13 trials used in the gold standard and the 202795 encounters then formed a reference set for large-scale evaluation, in which a set of historical trial–patient enrollment decisions were leveraged as trial–patient matches. The enrollment decisions include all patients who were approached and their eligibilities confirmed in person (the patients could opt out of enrollment). The decisions do not build a gold standard because some eligible patients might not have been approached if the clinical research coordinators were busy enrolling other patients. However, the historical set includes all patients found eligible by the coordinators irrespective if they later declined enrollment. Consequently, the set forms a useful reference standard to evaluate ES algorithms in replicating eligibility decisions in a clinical practice setting.

Clinical trial description and patient EHR data

We collected the description of the 13 clinical trials as used by the research coordinators during manual screening, including

Figure 1: An example clinical trial description (trial 9 in online supplementary table A2).

Title: The utility of computerized neuropsychological testing (ImPACT) in the pediatric emergency department for predicting prolonged recovery and return to play after concussive injuries

Purpose

To determine if ImPACT test performance conducted in the PED acutely (<24 hours) after concussive injury for young athletes can predict recovery course.

Criteria

Inclusion Criteria:

- Pediatric Patients ages 11-18 years old
- Sports related blunt trauma to head or body
- GCS of 13-15
- Injury within 24 hours of ED presentation

Exclusion Criteria:

- Patient needing acute intervention or urgent admission
- Patient suffering multiple trauma needing prolonged immobilization
- Patient recently received medications that affect neurocognitive functioning

title, purpose, and inclusion/exclusion criteria. An example clinical trial description is shown in figure 1, and a description of the trials is presented in online supplementary table A2.

On the basis of the prestudy communication with the ED physicians, we extracted 15 EHR data fields that were commonly reviewed by clinical research coordinators during ES to represent the patients' profiles. The data fields were categorized into two groups: (1) structured fields, such as demographics and laboratory data; (2) unstructured text-based fields, such as diagnosis and clinical notes. A description of the data fields is presented in table 1. The structured fields were used to build logical constraint filters (LCFs), while the unstructured fields were used in NLP-based matching components. Not every encounter had all unstructured fields present, and the descriptive statistics of these fields are shown in figure 2.

Automated ES approach

We customized and implemented state-of-the-art NLP, IE, and ML algorithms to build the ES framework (figure 3). Given a clinical trial and the encounter candidates, the approach applied LCFs to exclude ineligible encounters based on structured data fields derived from the trial criteria (step 1 in figure 3). The unstructured data fields of the prefiltered encounters were then processed, from which the medical terms were extracted and stored in the encounter pattern vectors (step 2). The same process was applied to the trial criteria to construct the trial pattern vector (step 3); the vector was also extended with informative patterns extracted from EHRs of previously eligible patients to capture hyponyms relevant to the trial criteria (step 4). Finally, IE algorithms matched the trial vector with the encounter vectors and returned a ranked list of potentially eligible encounters (step 5).

Logical constraint filters

Some characteristics of a patient —for example, age and gender—have been beneficial in earlier studies.[29–31] Hence, we manually extracted the criteria of these characteristics from the trial description and applied LCFs on the structured data fields (table 1) to exclude ineligible encounters.

Text processing and medical term/assertion identification

The text processing utilized advanced NLP algorithms to extract informative textual patterns from patients' unstructured data fields. The process first combined the documents based on field types. For example, if two physician notes and two medical history entries were written during an encounter, the clinical narratives were concatenated on the basis of field types and generated two documents (one physician note and one medical history). The content was then segmented into sentences using the Stanford sentence parser, and all duplicate sentences (eg, copy-pasted and 'templated' narratives) within the same field-type-based document were removed.[41] Noninformative tokens, such as stop words, were also removed in this process. All remaining words from the documents were stored as bag-of-words in the encounter pattern vectors.

The importance of medical information hidden within clinical narratives has been increasingly recognized as a critical component in describing a patient's profile.[29,30,35] Building on our experience with Mayo Clinic's clinical Text Analysis and Knowledge Extraction System (cTAKES), we adapted it to extract text-derived, term-level medical information from the unstructured data fields.[42] cTAKES assigned concept unique identifiers (CUIs) from the Universal Medical Language System (UMLS), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes, and the clinical drug codes of RxNorm to text strings.[43–45] Our customized cTAKES implementation is described in our earlier publications.[46,47] None of the trial description or the patient data used in the present study were included in the earlier training of our customized cTAKES model.

To convert negation expression, we implemented a negation detector based on the NegEx algorithm.[48] For example, the

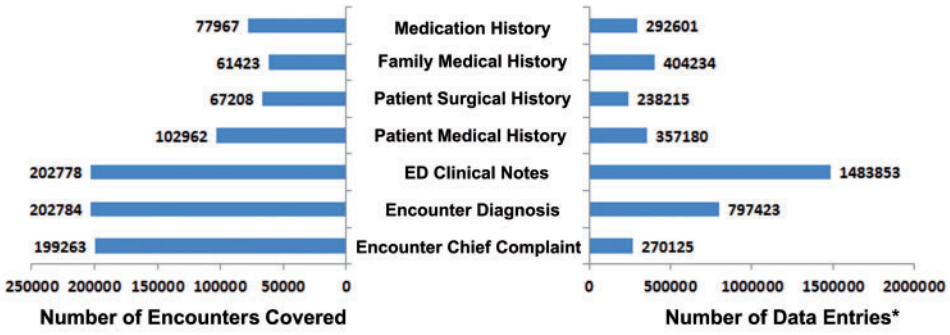
Table 1: Structured and unstructured data fields extracted from patients' electronic health records

Data field	Data field description	Data field class
Age	Patient's age	Demographics (S)
Gender	Patient's gender	Demographics (S)
Language	Patient's spoken language	Demographics (S)
Acuity	Acuity of the patient's chief complaint (from 1 to 5:1 indicates urgent complaint and 5 non-urgent complaint)	Encounter information (S)
Guardian presence	Whether the patient is escorted by his/her legal guardian	Encounter information (S)
Pregnancy, Yes/No	Whether the patient is pregnant	Encounter information (S)
Vital signs	Patient's first vital sign measurements (eg, temperature) in the ED	Encounter information (S)
GCS*	Patient's Glasgow Coma Scale	Encounter information (S)
Chief complaint	Patient's chief complaint documented during the encounter	Encounter information (US)
Diagnosis	Patient's diagnosis documented during the encounter	Encounter information (US)
ED clinical notes	Clinical notes written in the ED during the encounter	Encounter information (US)
Medical history	Patient's medical history documented during the encounter. Includes patient's historical diagnoses (eg, diagnosis name, diagnosis date, and brief comment about the diagnosis) prior to this encounter	History information (US)
Surgical history	Patient's surgical history documented during the encounter. Includes surgery (eg, surgery name and date of surgery) to the patient prior to this encounter	History information (US)
Family history	Relevant medical histories of patient's family members documented during the encounter. Include the problems of the family members provided by the patient	History information (US)
Medication history	Patient's medication history documented during the encounter. Includes all medications used by the patient prior to this encounter	History information (US)

'S' in 'Data field class' indicates a structured field and 'US' an unstructured text-based field.
*A default score of 15 was generated for GCS if the chief complaint was not related to head trauma.
ED, emergency department.

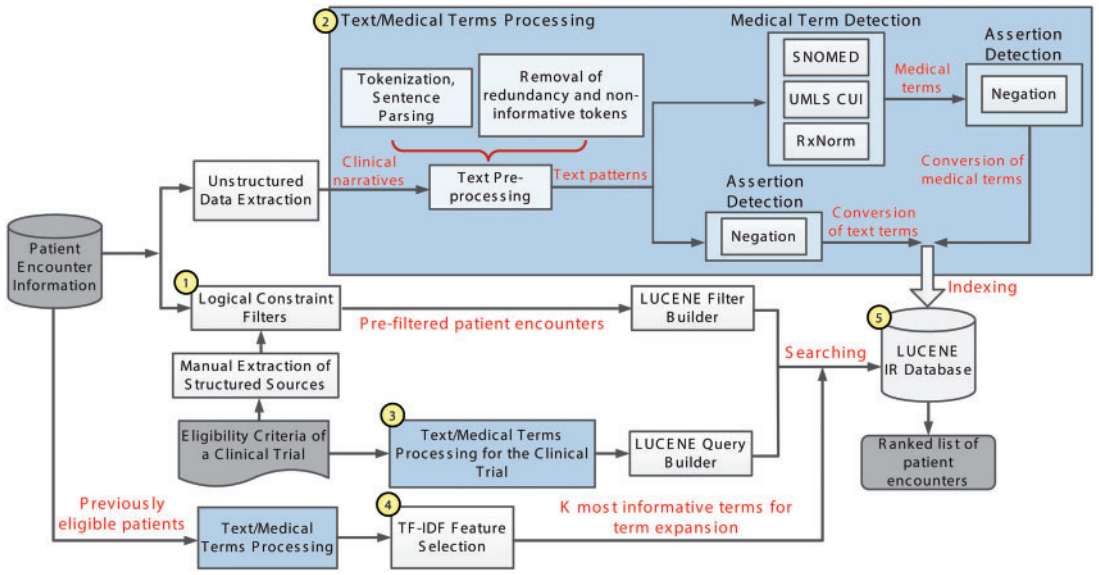
RESEARCH AND APPLICATIONS

Figure 2: Numbers of encounters covered by the unstructured data fields and numbers of entries for the fields.



*A data entry is a piece of information documented at some point during an encounter. A data field (particularly clinical notes) may have multiple data entries during an encounter.

Figure 3: Architecture of the proposed automated eligibility screening approach.



phrase 'Negative for abdominal pain' was converted into 'NEG_C0694551' in the assertion detection component. All text and medical terms were converted if necessary before being added to the encounter vectors.

For the trial eligibility description, the same text and medical term processing was applied to the inclusion and exclusion criteria to extract term-level patterns for the trial vectors. All terms extracted from the exclusion criteria were converted into negated format.

Supervised term expansion

Term expansion is the process of expanding a query with additional terms, mainly hyponyms of query words, to improve the match between the query and its candidates. Some of the top-performing approaches in the TREC medical record track have tried this technique for ES, which attempted to expand the trial vector with all possible hyponyms from the UMLS hierarchy.[29–31,35] For instance, beside using the word 'cancer' from the eligibility criteria, the algorithms looked up all words related to 'cancer' from the UMLS hierarchy (eg, 'neuroblastoma' and 'glioma') and added them to the trial vector. This unsupervised expansion was detrimental to screening performance because it introduced many irrelevant terms.[29,31,35] To address this problem, we developed a more principled term expansion component using supervised learning techniques: we used the most informative patterns retrieved from the EHRs of previously eligible patients for a trial to find the hyponyms (eg, football) relevant to the trial criteria (eg, sports-related trauma). This is the first, known to us, introduction of supervised learning to ES for enhancing trial–patient matching, which we refer to as supervised term expansion (STE). Mathematically, the text/medical terms from the encounter vectors of previously eligible patients were weighed by term frequency–inverse document frequency (TF-IDF) feature selection, where the top K terms ($K = 100$) were expanded to the trial vector.[49] The population of 'previously eligible patients' is described in the Experiments section below.

IE algorithms

The encounter vectors were used to represent patients' profiles and stored in a Lucene information retrieval database.[50] The same processes along with STE were applied to build the pattern vector for a trial. The IE algorithms then matched between the patterns of the trial and the prefiltered encounters and ranked the candidates based on TF-IDF similarity.[51] Finally, the ranked list of encounters was generated to facilitate the staff screening.

Experiments

Evaluation metrics

We adopted three evaluation metrics to measure performance. (1) To assess the overall quality of the ES output, we applied the mean average precision (MAP) commonly calculated in information retrieval:

$$\mathrm{MAP} = \frac{1}{M}\sum_{m=1}^{M}\frac{\Sigma_{n=1}^{N_m} P(n)\delta(n)}{N_m},$$

where M is the number of trials, N_m is the number of eligible encounters for trial m, n denotes the n-th encounter in the output, P(n) is the precision at cut-off n, and $\delta(n)$ is an indicator function equaling 1 if the n-th encounter is eligible for trial m and is 0 otherwise.[52] (2) To compare the screening efficiencies of different ES approaches, we used the 'workload' metrics, defined as the number of encounters required to be reviewed from the output to identify all eligible patients.[39] The workload

equals the number of predicted eligible encounters (ie, true positive + false positive) when recall = 100%. (3) To assess the recall at different algorithm cut-offs, we thresholded the ES output with 10–100% cut-offs and plotted the recall curve. These evaluations were applied to both the gold standard and the reference standard experiments.

Comparison of ES approaches

The baseline approach (denoted by BASELINE) simulated the screening process without automated ES. It was implemented by randomly shuffling the encounter list for a clinical trial. We then compared its performance with three variants of the ES approach that cumulatively integrated the proposed components: (1) LCF: the approach used the LCF component to exclude ineligible encounters and randomly shuffled the prefiltered encounters for a trial; (2) LCF + NLP: the approach specified in figure 3 without the STE component; (3) LCF + NLP + STE: the STE component was also included. To fill in the gap in the TREC literature, we additionally validated the contribution of different pattern sets on the LCF + NLP approach: we tested the four pattern sets (Text, UMLS CUI, SNOMED CT, and RxNorm) individually and in combination and assessed the MAP performance respectively. The best combination of the pattern sets was used in LCF + NLP + STE.

In all experiments no manual customizations were made to our ES algorithms (eg, adding additional rules to the negation detector) to over-fit the current datasets. The STE component was always trained on the data that were never part of the test set in each experiment.

Evaluation scenarios

We first performed twofold cross-validation on the gold standard set to evaluate the ES approaches. For each fold we used 300 encounters as candidates and evaluated the ES outputs for each trial against the gold standard eligibility decisions. The eligible patients in the other 300 encounters were regarded as 'previously eligible patients' to train the STE component. To assess the performance of STE with different sizes of training samples, we also used 1–100% of the eligible patients from the reference standard set to train the component (figure 4B). To assure the integrity of the evaluation, all patients in the gold standard were removed from the training data, providing 3864 'previously eligible patients'. In the case of 1%/2%/5% of the training data, the experiments were repeated 100/50/20 times on each fold to enable the use of all training samples. The results were then averaged over the experiments as the performance of that fold. For the rest of the portions, the experiments were repeated 10 times. For all experiments, the statistical significance of the performance difference was assessed using the paired t test. Because of the number of different tests conducted, we also applied the Bonferroni correction to the p values to account for the increased possibility of type I error.[53]

To conduct the evaluation on the reference standard set, we simulated the current practice and assessed the ES approaches on a day-by-day basis—that is, given an open trial and all encounters on day X, we ran the ES algorithms on the encounters for this trial and evaluated the outputs against the historical decisions on day X. The performance was averaged over all open days of the trial as performance for this trial. In this scenario, the patients found eligible for a trial up to day X were used to train the STE component. Hence, on day 1, the STE was not used because no previously eligible patients were available, while, on day 2, all patients found eligible in day 1 were used to train the STE, and so forth.

RESULTS

Descriptive statistics of evaluation data

For the gold standard set, the physicians reviewed 3061 trial–-encounter pairs and found 75 matches (2.45% average eligibility rate). The numbers of eligible candidates for the trials are presented in online supplementary table A2. The overall inter-annotator agreement was 96.81%, indicating good agreement on the eligibility decision.

Among the 202795 encounters, patients in 4177 encounters were found eligible for any of the 13 trials in historical enrollment decisions, providing 4210 trial–encounter matches in the reference standard set (see online supplementary table A2) (average eligibility rate 1.25%).

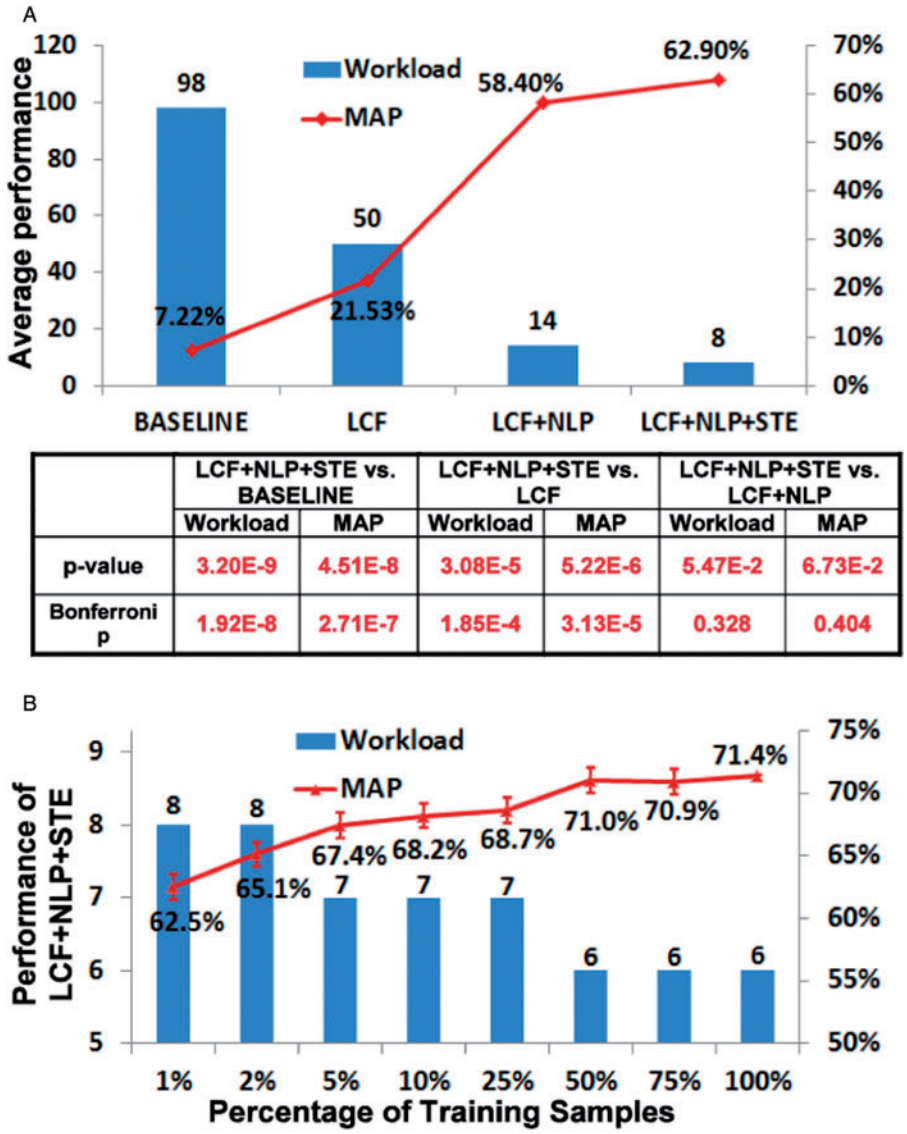
Gold standard experiments

Figure 4A presents the average workload and MAP performance of the ES approaches over all trials in the twofold cross-validation experiment. Without automated prescreening (BASELINE), a clinical research coordinator would have to screen on average 98 encounters per trial to identify all eligible patients in the gold standard set. With the automated approach LCF + NLP + STE, the workload was reduced dramatically by more than 90% to eight screened encounters per trial. A similar trend was observed when MAP between different approaches was compared, where the improvements of LCF + NLP + STE over BASELINE and LCF were statistically significant. In the cross-validation experiment, LCF + NLP + STE did not significantly outperform LCF + NLP because of insufficient training data (figure 4A). However, we observed consistent improvement in its performance when more training data from the reference standard were used (figure 4B). In the case of 10% training data (387 samples), it outperformed LCF + NLP statistically significantly on both evaluations (p = 1.00E-9 on workload and p = 4.86E-2 on MAP).

Figure 5 presents the recall curves at different algorithm cut-offs. LCF + NLP + STE (trained on eligible patients in the alternative fold of the gold standard) achieved 90% recall when thresholding the top 22% of its output as eligible candidates, suggesting that the screening efficiency was improved by about 450% while missing only 10% of eligible patients.

Finally, we assessed the contribution of the four pattern sets in table 2. The LCF + NLP approach with all patterns achieved the best performance (combination 15), followed closely by LCF + NLP using Text, SNOMED CT, and UMLS CUI (combination 14). The improvements of the best pattern combination were statistically significant over the variants using Text, SNOMED CT, and RxNorm individually (combination 1/2/4), or

Figure 4: Average workload and mean average precision (MAP) performance of the eligibility screening (ES) approaches on the gold standard set (A) and the performance of LCF + NLP + STE with different sizes of training samples (B). Statistical significance tests (paired t test) of the performance difference between LCF + NLP + STE and the other ES approaches are also presented. LCF, logical constraint filter; NLP, natural language processing; STE, supervised term expansion.



	LCF+NLP+STE vs. BASELINE		LCF+NLP+STE vs. LCF		LCF+NLP+STE vs. LCF+NLP	
	Workload	MAP	Workload	MAP	Workload	MAP
p-value	3.20E-9	4.51E-8	3.08E-5	5.22E-6	5.47E-2	6.73E-2
Bonferroni p	1.92E-8	2.71E-7	1.85E-4	3.13E-5	0.328	0.404

any combination of Text and SNOMED CT with RxNorm (combination 6/8). It is worth noting that the LCF + NLP approach with UMLS CUI (combination 2), or any combination of Text, SNOMED CT, and UMLS CUI (eg, combination 9/10/13) also achieved high performances, which were close to that of the best combination.

Reference standard experiments

Figure 6 illustrates the evaluation on the reference standard set, where we observed identical trends of performance for the four ES approaches. Again, LCF + NLP + STE achieved the best performance, and its improvements over the other approaches were statistically significant.

DISCUSSION

In the gold standard experiments, the LCF approach showed good capability in excluding ineligible patients (workload reduction 49%, 50 vs 98 screened encounters). However, without the information from clinical narratives, it was unable to match descriptive criteria (eg, diagnosis) with patients' profiles. By applying the NLP and IE algorithms, LCF + NLP further improved the performance (workload reduction 86%, 14 vs 98 screened

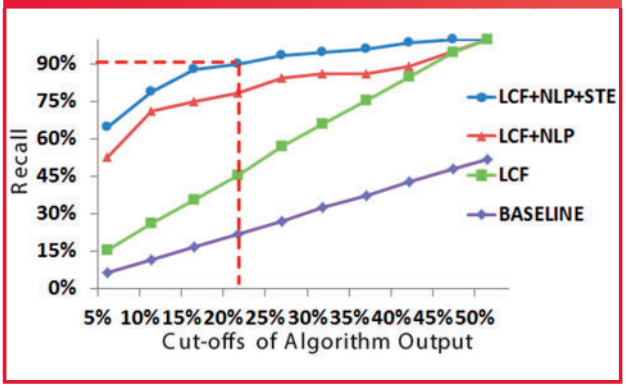


Figure 5: Recall performance of the eligibility screening approaches at different cut-offs of algorithm outputs. LCF, logical constraint filter; NLP, natural language processing; STE, supervised term expansion.

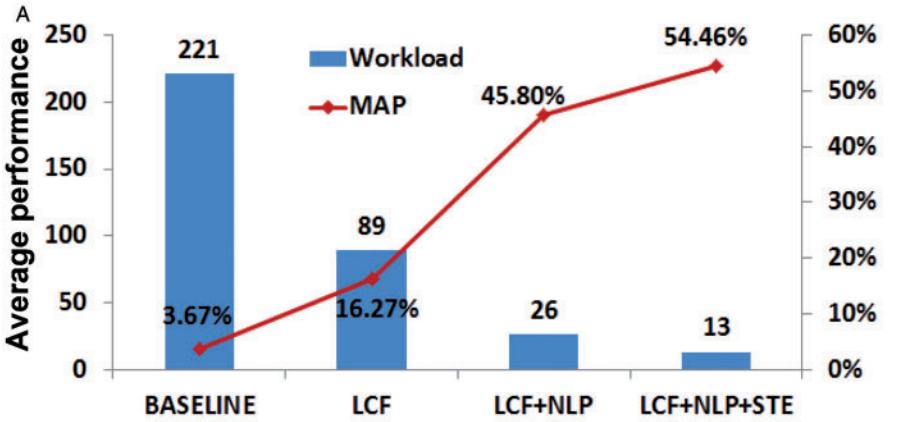
encounters). This result verifies the effectiveness of the NLP and IE techniques and confirms the findings of some reports of the TREC medical record track on real-world data.[29–31,33,35] For LCF + NLP + STE, we observed consistent improvement in performance when the STE training data increased (figure 4B). When a training size similar to the test data was used, the approach achieved better performance than LCF + NLP (figure 4A, workload reduction 43%, 8 vs 14 encounters). Given sufficient training samples (figure 4B), LCF + NLP + STE outperformed LCF + NLP statistically significantly. This promising result showed the great potential of STE in boosting the performance of automated ES. One representative example was observed on trial 9, where the inclusion criterion 'sports related blunt trauma' was ambiguous and found hardly any matches in patients' clinical notes. By exploring the EHRs of previously eligible patients, STE additionally picked up sport-related terms (eg, football and soccer) for the trial vector and greatly

Table 2: Average MAP of the LCF + NLP approach using different combinations of pattern sets; statistical significance tests (paired t test) of the performance difference between the best pattern combination and the others are also presented

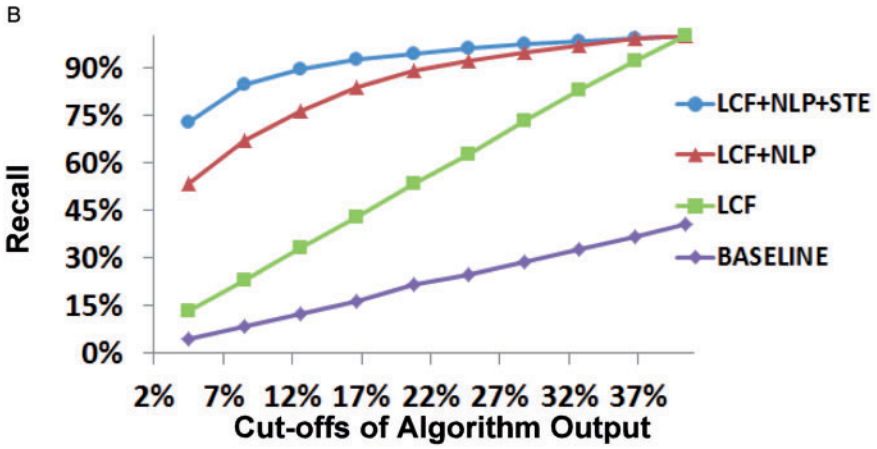
Pattern set					MAP	p Value
Combination	Text	SNOMED	CUI	RxNorm		
1	×	×	×	√	0.296	1.61E-4*
2	×	×	√	×	0.559	0.354
3	×	√	×	×	0.502	7.30E-3*
4	√	×	×	×	0.527	4.10E-2*
5	×	×	√	√	0.553	0.322
6	×	√	×	√	0.503	1.48E-2*
7	×	√	√	×	0.554	9.10E-2
8	√	×	×	√	0.527	3.45E-2*
9	√	×	√	×	0.562	0.160
10	√	√	×	×	0.565	0.260
11	×	√	√	√	0.548	6.38E-2
12	√	×	√	√	0.562	0.161
13	√	√	×	√	0.565	0.285
14	√	√	√	×	0.583	8.91E-2
15	√	√	√	√	**0.584**	N/A

√, pattern set used; ×, otherwise.
Bold number indicates the best result.
N/A indicates that the performances between the two ES approaches are identical and no p value is returned.
*The performance difference between the two ES approaches is statistically significant at the 0.05 level.
CUI, concept unique identifier; ES, eligibility screening; LCF, logical constraint filter; MAP, mean average precision; NLP, natural language processing.

Figure 6: Average workload and mean average precision (MAP) performance of the eligibility screening approaches on the reference standard set (A) and the recall performance of the approaches at different algorithm cut-offs (B). Statistical significance tests (paired t test) of the performance difference between LCF + NLP + STE and the other approaches are also presented. LCF, logical constraint filter; NLP, natural language processing; STE, supervised term expansion.



	LCF+NLP+STE vs. BASELINE		LCF+NLP+STE vs. LCF		LCF+NLP+STE vs. LCF+NLP	
	Workload	MAP	Workload	MAP	Workload	MAP
p-value	0.00	0.00	0.00	0.00	8.14E-66	2.03E-27
Bonferroni p	0.00	0.00	0.00	0.00	4.88E-65	1.22E-26



improved the trial–patient matching (evidenced by a 66.7% workload reduction over LCF + NLP on this trial, 2 vs 6 screened encounters). The identical trends for the four approaches observed in the reference standard experiments confirmed the above findings and validated the scalability of our ES algorithms.

By investigating the contribution of different pattern sets (table 2), we found that no single pattern covered a complete list of information, and the UMLS CUI was shown to be more informative than the others. The Text set was less informative, and combining it with UMLS CUI slightly improved the performance (combination 9 vs 2). A similar trend was observed on the SNOMED CT set, which did not contribute much additional information when UMLS CUI was used (combination 7 vs 2). Since drug-related information in the trial criteria was sparse (see online supplementary table A2), the RxNorm set contributed little information for trial–patient matching. Consequently, combining RxNorm with the other patterns barely influenced the results in our case. These observations suggest that, when designing the ES algorithms, one should customize pattern sets on the basis of trial requirements (eg, whether it contains drug information). Adding more patterns will not always increase the screening performance.

Table 3: False positive errors made by the LCF + NLP + STE approach

Cause of false positive errors identified by the chart review	Error (%)
1. The ES approach matched similar signs and symptoms (eg, RLQ and RUQ abdominal pain) but omitted the other criteria	30.68
2. The ES approach matched the correct diagnosis but could not identify ineligible patients because the exclusions did not exist in the collected EHR data fields (eg, less than 32 weeks' gestational age)	17.04
3. The ES approach omitted the negation expression of the signs and symptoms (eg, Mom denied patient had diarrhea) and hence caused wrong patient recommendation	14.78
4. The ES approach matched the correct diagnosis but omitted some inclusions/exclusions implied in the clinical narratives (eg, symptoms >72 h)	13.63
5. The ES approach matched the terms expanded by the STE component (eg, football, soccer and skating) but omitted the primary criteria (eg, diagnosis)	2.27
6. Wrong diagnosis, other reasons	21.59

EHR, electronic health record; ES, eligibility screening; LCF, logical constraint filter; NLP, natural language processing; RLQ, right lower quadrant; RUQ, right upper quadrant; STE, supervised term expansion.

Error analysis, limitations, and future work

We performed error analysis for LCF + NLP + STE by reviewing the charts for all false positives made in the workload evaluation on the twofold cross-validation experiment. The LCF + NLP + STE approach made 88 errors, which were grouped into six categories in table 3. About 44% of the errors were ascribed to the confusion between similar signs and symptoms (cause 1, eg, recommending a patient with 'RUQ abdominal pain' to a trial for 'RLQ abdominal pain') and the omission of exclusions implied in the clinical narratives such as time-related criteria (cause 4, eg, omitting the clue 'pain started four days ago' indicating that the symptom had lasted for more than 72 h). This is because our ES approach uses 'bag-of-words' patterns, which limits its ability in finding semantic relations between consecutive words. To alleviate this problem, we will extend the pattern set to 'bag-of-phrases' in our future work and apply advanced NLP algorithms to analyze the semantic and temporal relations within the context.

Another set of errors were caused by missing inclusion/exclusion criteria in the EHR data fields (cause 2, eg, we did not collect the field of gestational age, a criterion used in trial 1 and 13 in this study). The approach will be more powerful if we integrate more EHR fields into the LCF component (eg, additional demographics and laboratory data). Since we did not manually customize the ES algorithms to over-fit the current data, the mistakes made by the components (eg, negation detector and STE) were propagated and caused errors in patient recommendation (causes 3 and 5). We will tune these components on our current data (eg, introducing additional rules in the negation detector) to improve their accuracies in future study.

One limitation of the study is that its evaluation is restricted to retrospective data. In the future, we will evaluate the practicality of automated ES in a randomized controlled prospective test environment. To verify the generalizability of the ES algorithms, we plan to test our approach on a more diversified patient population (eg, adult patients), multiple institutions, and clinical data under different formats (eg, clinical record formats used in different vendors' EHR product).

CONCLUSION

By leveraging NLP, IE, and ML technologies on both the eligibility criteria and the patient EHRs, we demonstrated that NLP-, IE-, and ML-based automated ES could successfully identify patients for disease-specific clinical trials. Using a physician-generated, gold-standard-based evaluation of real-world clinical data and trials, the approach achieved more than 90% workload reduction potential in patient cohort identification and showed the potential of a 450% increase in trial screening efficiency. This work also verified the effectiveness of the NLP, IE, and ML algorithms and UMLS components in a real-world dataset. Large-scale evaluation on the historical trial–patient enrollment decisions confirmed the findings and validated the scalability of the proposed algorithms. Consequently, we hypothesize that the automated ES approach, when rolled out for production, will have potential for significant impact in reduction of time and effort for executing clinical research, particularly as important new initiatives greatly expand the number of, and access to, potential clinical trials for patients.

ACKNOWLEDGEMENTS

Particular thanks go to Jordan Wright for his detailed reading and comments on the manuscript. The authors also thank Melanie Hounchell and Andrea Kachelmeyer for their support in providing the data.

CONTRIBUTORS

YN coordinated the development of the gold standard, extracted the patient EHR data, ran all the experiments, analyzed the results, created the tables and figures, and wrote the manuscript. SK and CMM developed the gold standard set of trial–patient matches, provided suggestions in developing the ES approach, analyzed the errors, and contributed to the manuscript. JWD extracted the data of historical trial–patient enrollment decisions, consulted on data quality and cleaning, coordinated the development of the gold standard, and contributed to the manuscript. IS coordinated the work, supervised

the experiments, data cleaning, analysis of the results, and contributed to the manuscript. HT, TL, QL and HZ coordinated the experiments and contributed to the manuscript.

FUNDING

This work was partially supported by NIH grants 5R0 0LM010227-05, 1R21HD072883-01 and 1U01HG006828-01.

COMPETING INTERESTS

YN, HT, TL, QL and HZ were also supported by internal funds from Cincinnati Children's Hospital Medical Center.

ETHICS APPROVAL

Cincinnati Children's Hospital Medical Center Institutional Review Board.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. *BMC Med Inform Decis Mak* 2008;8:13.
2. Thadani SR, Weng C, Bigger JT, *et al*. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009;16:869–73.
3. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc* 2009;16:316–27.
4. Penberthy LT, Dahman BA, Petkov VI, *et al*. Effort required in eligibility screening for clinical trials. *J Oncol Pract* 2012;8:365–70.
5. Ibrahim GM, Chung C, Bernstein M. Competing for patients: an ethical framework for recruiting patients with brain tumors into clinical trials. *J Neurooncol* 2011;104:623–7.
6. Wynn L, Miller S, Faughnan L, *et al*. Recruitment of infants with sickle cell anemia to a Phase III trial: data from the BABY HUG study. *Contemp Clin Trials* 2010;31:558–63.
7. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004;5:417–29.
8. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using real time recuiting. *Proc AMIA Symp* 2000;2000:111–15.
9. Embi PJ, Jain A, Clark J, *et al*. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA Annu Symp Proc* 2005;2005:231–5.
10. Grundmeier RW, Swietlik M, Bell LM. Research subject enrollment by primary care pediatricians using an electronic health record. *AMIA Annu Symp Proc* 2007;2007:289–93.
11. Nkoy FL, Wolfe D, Hales JW, *et al*. Enhancing an existing clinical information system to improve study recruitment and census gathering efficiency. *AMIA Annu Symp Proc* 2009;2009:476–80.
12. Treweek S, Pearson E, Smith N, *et al*. Desktop software to identify patients eligible for recruitment into a clinical trial: using SARMA to recruit to the ROAD feasibility trial. *Inform Prim Care* 2010;18:51–8.
13. Heinemann S, Thüring S, Wedeken S, *et al*. A clinical trial alert tool to recruit large patient samples and assess selection bias in general practice research. *BMC Med Res Methodol* 2011;11:16.
14. Pressler TR, Yen PY, Ding J, *et al*. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Med Inform Decis Mak* 2012;12:47.
15. Ding J, Erdal S, Borlawsky T, *et al*. The design of a pre-encounter clinical trial screening tool: ASAP. *AMIA Annu Symp Proc* 2008;931.
16. Penberthy L, Brown R, Puma F, *et al*. Automated matching software for clinical trials eligibility: measuring efficiency and flexibility. *Contemp Clin Trials* 2010;31:207–17.
17. Embi PJ, Jain A, Clark J, *et al*. Effect of a clinical trial alert system on physician participation in trial recruitment. *Arch Intern Med* 2005;165:2272–7.
18. Beauharnais CC, Larkin ME, Zai AH, *et al*. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. *Clin Trials.* 2012;9:198–203.
19. Petkov VI, Penberthy LT, Dahman BA, *et al*. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. *Exp Biol Med.* 2013;238:1370–8.
20. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. *J Am Med Inform Assoc* 2012;19(e1):145–8.
21. Stanfill MH, Williams M, Fenton SH, *et al*. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010;17:645–51.
22. Weng C, Tu SW, Sim I, *et al*. Formal representation of eligibility criteria: a literature review. *J Biomed Inform* 2010;43:451–67.
23. Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *J Biomed Inform* 2011;44:927–35.
24. Weng C, Wu X, Luo Z, *et al*. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011;18(Suppl 1):116–24.
25. Tu SW, Peleg M, Carini S, *et al*. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform* 2011;44:239–50.
26. 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.

27. Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. *J Biomed Inform.* 2013;46:805–13.
28. Text REtrieval Conference [Website]. 2013. http://trec.nist.gov/ (accessed 17 April 2014).
29. Gurulingappa H, Müller B, Hofmann-Apitius M, *et al*. A semantic platform for informational retrieval from E-Health records. The twentith Text REtrieval Conference Proceedings (TREC); 2011. http://trec.nist.gov/pubs/trec20/papers/Fraunhofer-SCAI.med.update.pdf (accessed 17 April 2014).
30. King B, Wang L, Provalov I, *et al*. Cengage Learning at TREC 2011 medical track. The twentith Text REtrieval Conference Proceedings (TREC) 2011. http://trec.nist.gov/pubs/trec20/papers/Cengage.medical.update.pdf (accessed 17 April 2014).
31. Limsopatham N, Macdonald C, Ounis I, *et al*. University of Glasgow at medical records track 2011: Experiments with Terrier. The twentith Text REtrieval Conference Proceedings (TREC) 2011. http://trec.nist.gov/pubs/trec20/papers/uogTr.crowd.microblog.web.update4-20.pdf (accessed 17 April 2014).
32. Demner-Fushman D, Abhyankar S, Jimeno-Yepes A, *et al*. A knowledge-based approach to medical records retrieval. The twentith Text REtrieval Conference Proceedings (TREC) 2011. http://trec.nist.gov/pubs/trec20/papers/NLM.medical.pdf (accessed 17 April 2014).
33. Callejas PMA, Wang Y, Fang H. Exploiting domain thesaurus for medical record retrieval. The twentith firstText REtrieval Conference Proceedings (TREC) 2012. http://trec.nist.gov/pubs/trec21/papers/udel_fang.medical.nb.pdf (accessed 17 April 2014).
34. Limsophatham N, McCreadie R, Albakour MD, *et al*. University of Glasgow at TREC 2012: Experiments with Terrier in Medical Records, Microblog, and Web Tracks. The twentith first Text REtrieval Conference Proceedings (TREC); 2012. http://trec.nist.gov/pubs/trec21/papers/uogTr.medical.microblog.web.final.pdf (accessed 17 April 2014).
35. Qi Y, Laquerre PF. Retrieving medical records with "sennamed": NEC labs America at TREC 2012 medical records track. The twenty first Text REtrieval Conference Proceedings (TREC); 2012. http://trec.nist.gov/pubs/trec21/papers/sennamed.medical.final.pdf (accessed 17 Apr 2014).
36. Zhu D, Carterette B. Exploring evidence aggregation methods and external expansion sources for medical record search. The twenty first Text REtrieval Conference Proceedings (TREC); 2012. http://trec.nist.gov/pubs/trec21/papers/udel.medical.final.pdf (accessed 17 April 2014).
37. Chapman WW, Nadkarni PM, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;18:540–3.
38. Edinger T, Cohen AM, Bedrick S, *et al*. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. *AMIA Annu Symp Proc* 2012;2012:180–8.
39. Schmickl CN, Li M, Li G, *et al*. The accuracy and efficiency of electronic screening for recruitment into a clinical trial on COPD. *Respir Med* 2011;105:1501–6.
40. Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC);2008.
41. De Marneffe M, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In Proc. of the International Conference on Language Resources and Evaluation (LREC); 2006;449–454.
42. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
43. Universal Medical Language System [Website]. 2013. http://www.nlm.nih.gov/research/umls/ (accessed 17 April 2014).
44. Systematized Nomenclature of Medicine [Website]. 2013. http://www.ihtsdo.org/snomed-ct/ (accessed 17 April 2014).
45. RxNorm: normalized naming system for clinical drugs [Website]. 2013. https://www.nlm.nih.gov/research/umls/rxnorm/ (accessed 17 April 2014).
46. Li Q, Zhai H, Deleger L, *et al*. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *J Am Med Inform Assoc.* 2013;20:915–21.
47. Deleger L, Li Q, Lingren T, *et al*. Building gold standard corpora for medical natural language processing tasks. *AMIA Annu Symp Proc* 2012;2012:144–153.
48. Champman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34:301–10.
49. Jing L, Huang H, Shi H. Improve feature selection approach TFIDF in text mining. In Proc. of the first International Conference of machine learning and cybernetics; 2002; 944–6.
50. Apache Lucene [Website]. 2013. http://lucene.apache.org/ (accessed 17 April 2014).
51. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 1972; 28:11–21.
52. Baeza-Yates RA, Ribeiro-Neto BA. *Modern information retrieval*. Addison Wesley, 1999.
53. Bland J, Altman D. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.

AUTHOR AFFILIATIONS

[1]Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[2]Division of Pediatric Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA