JAMIA Open, 1(2), 2018, 265–274
doi: 10.1093/jamiaopen/ooy010
Advance Access Publication Date: 28 September 2018
Research and Applications

Research and Applications

Learning relevance models for patient cohort retrieval

Travis R. Goodwin and Sanda M. Harabagiu

Department of Computer Science, Human Language Technology Research Institute, University of Texas at Dallas, Richardson, Texas, USA

Corresponding Author: Travis Goodwin, PhD, Department of Computer Science, University of Texas at Dallas, PO Box 830688, MS EC31, Richardson, TX 75080-0688, USA (travis@hlt.utdallas.edu).

Received 30 December 2017; Revised 26 February 2018; Editorial Decision 20 March 2018; Accepted 5 September 2018

ABSTRACT

Objective: We explored how judgements provided by physicians can be used to learn relevance models that enhance the quality of patient cohorts retrieved from Electronic Health Records (EHRs) collections.
Methods: A very large number of features were extracted from patient cohort descriptions as well as EHR collections. The features were used to investigate retrieving (1) neurology-specific patient cohorts from the de-identified Temple University Hospital electroencephalography (EEG) Corpus as well as (2) the more general cohorts evaluated in the TREC Medical Records Track (TRECMed) from the de-identified hospital records provided by the University of Pittsburgh Medical Center. The features informed a learning relevance model (LRM) that took advantage of relevance judgements provided by physicians. The LRM implements a pairwise learning-to-rank framework, which enables our learning patient cohort retrieval (L-PCR) system to learn from physicians' feedback.
Results and Discussion: We evaluated the L-PCR system against state-of-the-art traditional patient cohort retrieval systems, and observed a 27% improvement when operating on EEGs and a 53% improvement when operating on TRECMed EHRs, showing the promise of the L-PCR system. We also performed extensive feature analyses to reveal the most effective strategies for representing cohort descriptions as queries, encoding EHRs, and measuring cohort relevance.
Conclusion: The L-PCR system has significant promise for reliably retrieving patient cohorts from EHRs in multiple settings when trained with relevance judgments. When provided with additional cohort descriptions, the L-PCR system will continue to learn, thus offering a potential solution to the performance barriers of current cohort retrieval systems.

Key words: medical informatics, information storage and retrieval, search engine, machine learning

OBJECTIVE

Electroencephalography (EEG) records the electrical activity along the scalp and measures spontaneous electrical activity of the brain, which makes it a primary tool for diagnosis of brain-related illnesses.[1,2] But, as noted in Beniczky *et al.*,[3] the EEG signal is complex, and moreover, when EEG reports are created, the interobserver agreement in EEG interpretation is known to be moderate. Both these problems can be addressed by providing clinical experts with the ability to automatically retrieve similar EEG signals and EEG reports through a patient cohort retrieval (PCR) system operating on an Electronic Health Record (EHR). A multi-modal EEG PCR system called MERCuRY was presented in Goodwin and Harabagiu,[4] which leverages the heterogeneous nature of EEG data by processing both the clinical narratives from EEG reports as well as the raw electrode potentials derived from the recorded EEG signal data. Because the patient cohort criteria are expressed in natural language, the MERCuRY system is driven by its ability to rank relevant patients based on the narratives available from the EEG reports. However, as reported in Edinger *et al.*[5] the current state-of-the-art methods are not yet satisfactory for retrieving relevant patients from clinical narratives.

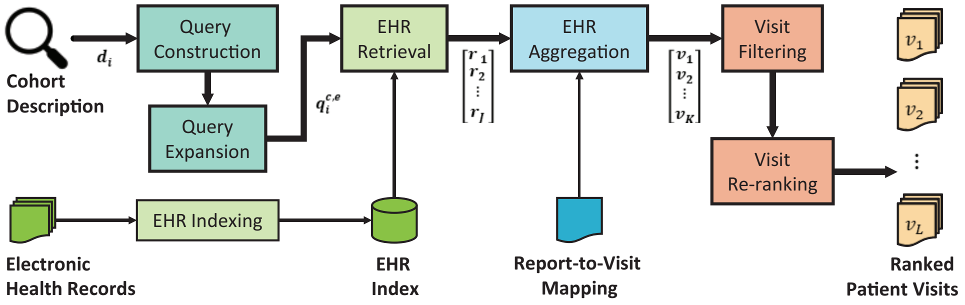


Figure 1. Architecture of a typical patient cohort retrieval system evaluated in TRECMed.

The primary objective of our study is the design, implementation and validation of a novel PCR system that *learns* how to optimally rank patients based on relevance judgements, providing improvements to current state-of-the-art methods. We demonstrate that by using a *learning-to-rank* framework informed by (1) features automatically extracted from the cohort description and from the clinical narratives and (2) physician feedback, we enhance the relevance of patients in cohorts by 27–53% above state-of-the-art. Moreover, our system for learning to rank patient cohorts is easily portable across EHR collections. Thus, our approach provides a framework for improving a PCR system when relevance judgements become available.

BACKGROUND AND SIGNIFICANCE

The automatic identification of patient cohorts satisfying a wide range of criteria—including clinical, demographic, and social information—has numerous applications,[6] e.g. (1) clinical trial recruitment; (2) outcome prediction; and (3) survival analysis. Because patient cohort identification relies on the processing of EHRs, many systems use statistical techniques or machine learning methods informed by natural language processing of the clinical narratives. However, these systems cannot *rank* the identified patients based on the *relevance* to the cohort criteria. Relevance is at the core of information retrieval (IR) systems.[7] Thus, viewing the problem of patient cohort identification as an IR problem, i.e. considering the problem of PCR, enables not only the identification of patients from a cohort, but to also the ranking of patients based on relevance to the inclusion and exclusion criteria used in the cohort description. Without considering patients' relevance, patient cohort identification systems produce only a binary decision: the patient either belongs or does not belong to the cohort. Ranking of the patients in the cohort was essential in the usability studies performed with the MERCuRY system,[4] as it enabled neurologist researchers to rapidly identify effective interventions for epilepsy accompanied by mental health comorbidities. However, not all the patients from the cohorts discovered by MERCuRY were relevant to the cohort criteria. Relevance judgements produced by neurologists indicated limitations of the system, but also provided important lessons that can be used for learning how to rank patients. Similarly, the analysis reported in Edinger *et al.*[5] indicates the limitations of PCR systems developed for the TREC Medical Records (TRECMed) track[8] in the annual Text REtrieval Conference (TREC) hosted by the National Institute for Standards and Technology (NIST). In the 2011 TRECMed evaluation, 24 PCR systems were tested against the same medical records and the same cohort descriptions, and their results were evaluated by 25 physicians.[9] The relevance judgements produced during the TRECMed evaluations could also be used for learning how to rank patients.

The MERCuRY system and most of the PCR systems participating in TRECMed had architectures similar to the one illustrated in Figure 1, providing a unifying framework for applying learning-to-rank. Learning-to-rank is a framework for using machine learning techniques for generating ranking models in IR.[10] In the architecture illustrated in Figure 1, the cohort description was processed with the goal of generating a machine-readable *query* (an essential component of any IR system). Many PCR systems generated a query by relying on MetaMap[11,12] to discern concepts from the Unified Medical Language System[13] (UMLS) from the cohort description[14]; while some systems also used NegEx[15] for detecting negated concepts.[12] In addition, there were systems that (1) mapped the cohort description to ICD-9 codes[16] or (2) simply considered bags of words.[17] After the queries were constructed, systems expanded the query by introducing new terms (e.g. synonyms, related words, etc.) using different techniques, including (1) pseudo-relevance feedback (PRF)[18]; (2) applying *Personalized PageRank*[19] to the UMLS Metathesaurus[20]; (3) using semantic vectors provided by Random Indexing.[21,22] Thus, each cohort description d_i was transformed in a query $q_i^{c,\ e}$ through a query construction method *c* and a query expansion method *e*. Moreover, as illustrated in Figure 1, the EHR collection was indexed using Apache Lucene,[23] Indri,[24] or Terrier.[25] The resultant index informed a variety of relevance models that produced an ordered list of medical records, $[r_1,\ r_2,\ \cdots,\ r_j]$. Because TRECMed focused on ranking hospital visits—groups of medical records generated during a patient's stay—rather than individual medical records, teams considered methods to aggregate the relevance scores or rankings of retrieved medical records to produce a ranking of hospital visits, $[v_1, v_2,\ \cdots,\ v_K]$, informed by a report-to-visit mapping provided to TRECMed participants. The resulting visit ranking was, in some cases, filtered to account for specific cohort criteria such as age or gender,[26] and/or re-ranked[27] to produce the final ranking of visits, $[v_1,\ v_2, \cdots,\ v_L]$. The learning PCR system presented in this paper allows the various techniques used in the components of the TRECMed systems to be unified within a single architecture, enabling the exploration of the impact of each technique on the optimal ranking of patients.

MATERIALS AND METHODS

Datasets and experimental settings

In this work, we explored the design of learning PCR systems in 2 settings: (1) a neurology-specific setting focusing on cohorts identified from a large archive of EEG reports and (2) a general setting allowing the recognition of cohorts from multiple forms of hospital

Table 1. Examples of cohort descriptions used to train and evaluate the learning cohort retrieval system

EEG reports	TRECMed 2011	TRECMed 2012
Patients experiencing seizures and generalized shaking	Patients with complicated GERD who receive endoscopy	Adult patients with Alzheimer's disease admitted from nursing homes with pressure ulcers
Multiple sclerosis and seizure	Women with osteopenia	Elderly patients with subdural hematoma
Patients under 18-year-old with absence seizures	Female patient with breast cancer with mastectomies during admission	Patients admitted with Hepatitis C and IV drug use
Patients over age 18 with history of developmental delay and EEG with electrographic seizures	Adult patients who are admitted with asthma exacerbation	Patients treated for post-partum problems including depression, hypercoagulability, or cardiomyopathy
Patients evaluated for seizures vs stroke	Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix	Patients with inflammatory disorders receiving TNF-inhibitor treatment
Brain tumor and sharp waves, spike/polyspike, and wave or spikes	Children admitted with cerebral palsy who received physical therapy	Adults under age 60 undergoing alcohol withdrawal
EEG showing triphasic waves	Patients co-infected with hepatitis C and HIV	Patients with AIDS who develop pancytopenia
Patients with anoxic brain injury and EEG reports denoting brain death	Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes	Patients with hypertension on anti-hypertensive medication
EEGs without sharp waves, spikes, or spike/polyspike and wave activity in patient's diagnosed with epilepsy	Patients with dementia	Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression
EEG showing generalized periodic epileptiform discharges	Cancer patients with liver metastasis treated in the hospital who underwent a procedure	Patients who develop thrombocytopenia in pregnancy

records. Approval was obtained for both EHR collections from the Institutional Review Board (IRB) at the University of Texas at Dallas (UTD).

The Temple University electroencephalogram corpus

For the neurology-specific setting, we relied on the publicly-available collection of EEG reports from the Temple University Hospital (TUH) EEG Corpus.[28,29] It contains EEG reports collected over 25 000 sessions for 15 000 patients over 12 years. While the TUH EEG corpus contains EEG signal information as well as EEG reports, we considered only the EEG reports.

The University of Pittsburgh

PCR systems participating in the TRECMed challenges had access to a large repository of 95 702 de-identified narratives from medical reports provided by the University of Pittsburgh Medical Center. This EHR repository consisted of 1 month of reports from multiple hospitals. Each medical report was associated with exactly 1 hospital visit (an individual patient's single stay at a hospital). The data set contained 93 551 medical reports mapped into 17 264 visits.

Cohort descriptions

Patient cohorts were recognized from each of the EHR collections based on descriptions provided by practicing clinicians. When using the TUH EEG corpus, 30 cohort descriptions were generated by 4 practicing neurologists. For TRECMed, we used the official cohort descriptions released by the task organizers. Thirty-four cohort descriptions were evaluated in 2011[9] and 47 additional descriptions were evaluated in 2012.[30] Examples of cohort descriptions used to train and evaluate the learning patient cohort retrieval (L-PCR) system are shown in Table 1.

Relevance judgments

To train and evaluate the retrieval performance of our system, we used visit-level *relevance judgments* produced by neurologists and clinicians for each cohort description described above. In both collections, physicians were asked to judge visits retrieved for each cohort as being RELEVANT, PARTIALLY RELEVANT, or NOT RELEVANT to the cohort. Supplementary Material Appendix H provides details on how the judgments were obtained for each collection.

The learning patient cohort retrieval system

In this study, we focused on the design of a L-PCR system. Unlike traditional PCR systems, such as MERCuRY[4] or those developed for TRECMed,[9,30] the L-PCR system uses a *learning-to-rank* approach for identifying patient cohorts that takes advantage of physician feedback. The learning-to-rank paradigm allows the L-PCR system to consider *relevance judgments* performed by clinicians to *learn* an improved patient relevance model used for retrieving and ranking patients for any given cohort descriptions.[10] The L-PCR system illustrated in Figure 2 includes 5 main components:

- a **query processing** component processes a given cohort description d_i to produce a machine-readable query, $q_i^{c,e}$;
- an **EHR processing** component produces an index of the narratives from the EHR collection;
- a **visit retrieval** component retrieves a sub-set of "candidate" visits from the EHR collection, $[v_1, \cdots, v_M]$, to be ranked by the learning relevance model (LRM);
- a **feature extraction** component extracts features vectors $[\boldsymbol{x}_1^i, \cdots, \boldsymbol{x}_M^i]$ corresponding to each candidate visit the relationship between the visit and the cohort description; and
- the **learning relevance model** uses a random forest (RF) classifier to infer the *relevance scores* $[s_1^i, \cdots, s_M^i]$ for each candidate

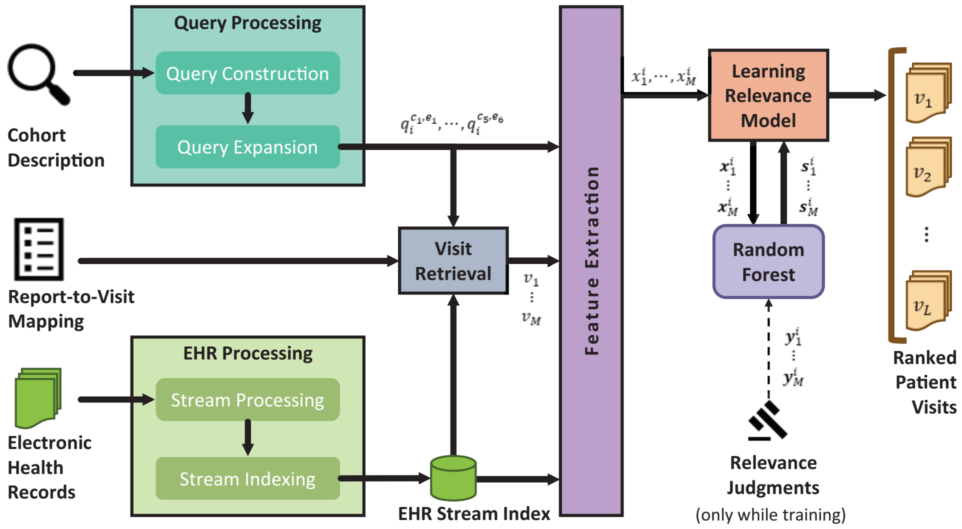
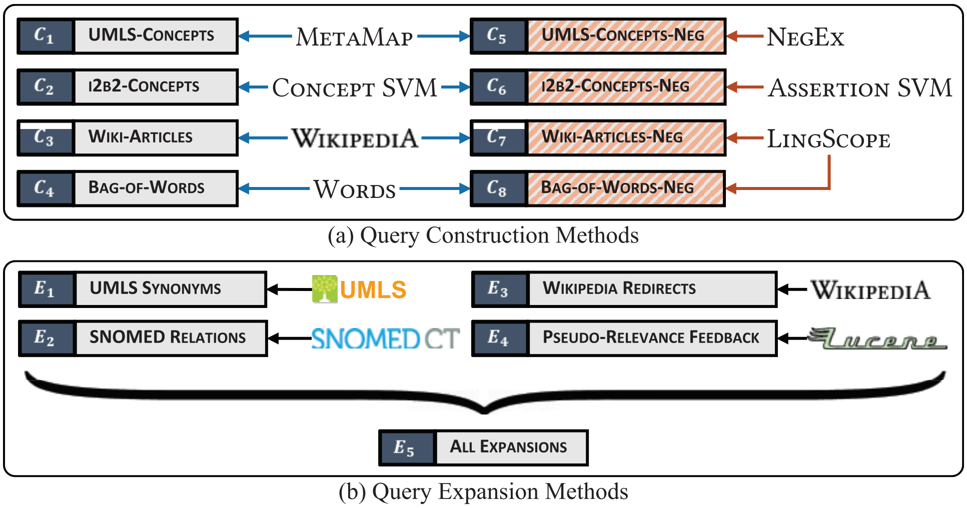


Figure 2. Architecture of the learning patient cohort retrieval system.



(a) Query Construction Methods

(b) Query Expansion Methods

Figure 3. Overview of the different approaches for (a) query construction and (b) query expansion used for feature extraction in the learning patient cohort retrieval system.

visit $[v_1^i, \cdots, v_M^i]$ based on their associated feature vectors $[x_1^i, \cdots, x_N^i]$; the RF is trained using the relevance judgments $[y_1^i, \cdots, y_M^i]$ provided by physicians.

An overview of each of these components is provided below, with additional details provided in Supplementary Material Appendices A–F.

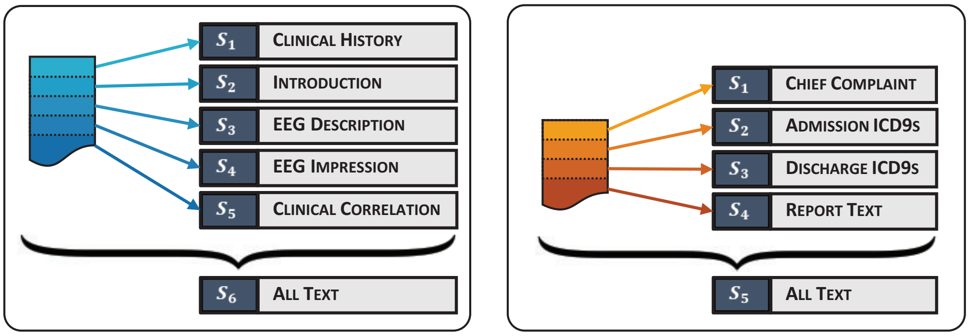
Query processing

As with the typical PCR system illustrated in Figure 1, each cohort description is first processed by a query construction step followed by query expansion.

Query construction. The L-PCR system incorporates 8 query construction methods, illustrated in Figure 3(a). Methods C_1, C_2, and C_3 represent the cohort description as a set of medical concepts detected by MetaMap,[11] classified by a support vector machine (SVM),[31,32] or corresponding to the titles of Wikipedia articles,[26,27] respectively. By contrast, method C_4 represents the cohort description as a set (i.e. "bag") of words. To account for the possibility of exclusion criteria in cohort descriptions (e.g. "without a diagnosis [of] schizophrenia"), we introduced a second version of C_1-C_4 in which the negation of any query component was detected using NegEx,[15] an SVM,[27,33] or LingScope,[34] respectively. Further details, rationale, and examples are provided in Supplementary Material Appendix A.

Query expansion. Figure 3(b) lists the 5 query expansion methods implemented within the L-PCR system. The first 4 query expansion methods incorporate synonyms from UMLS,[13] related concepts from SNOMED CT,[35] synonyms and misspellings from Wikipedia,[26] and individual words from related patient visits using PRF[7]. The fifth query expansion method is the combination of E_1-E_4. Further details and examples are provided in Supplementary Material Appendix B.

Note: in addition to query construction and expansion, we extracted any age or gender criteria from the cohort description using a grammar and lexicon previously described in Goodwin et al.[26] and described in Supplementary Material Appendix C.



(a) EEG Reports (b) Hospital Records

Figure 4. Indexed Streams from EEG reports (left) and hospital records (right). EEG: electroencephalography.

Electronic health record processing

Stream processing. We unified indexing, searching, and feature extraction across both EHR collections, by representing each EHR as a set of multiple, abstract *streams*[36] of unstructured information. Each stream corresponds to one or more sections in the EHR collection. Conceptually, each stream acts as a "lens" that determines which sections of the EHR are considered during feature extraction and retrieval. The stream representation allows the L-PCR system to automatically account for the semantics of each stream, without the semantics being explicitly encoded. Figure 4 illustrates the streams used for each EHR collection, while Supplementary Material Appendix D provides additional information about the content of each stream.

Stream indexing. To expedite feature extraction from the EHRs associated with each hospital visit, we separately indexed the content of each EHR collection using Apache Lucene.[23] We used a *tiered* indexing approach in which each stream was indexed independently, allowing individual streams of each EHR to be retrieved during feature extraction and retrieval. No pre-processing was applied beyond tokenization with Lucene's English Analyzer.[23]

Visit retrieval

To reduce complexity and improve scalability of the L-PCR system, rather than extracting features from every EHR in the collection, we rely on a basic retrieval step to identify a high-recall set of "candidate" visits likely to be relevant to the cohort description. These candidate visits are obtained by constructing a query with Bag-of-Words (C_1), expanding by all expansions (E_5), and identifying the top M ranked EHRs by the All Text stream (S_4/S_5) with the BM25 ranking function (in our experiments we used $M = 2,\ 000$). This allowed the set of "candidate" visits to be obtained by mapping the retrieved EHRs to their corresponding patient visits.

Feature extraction

Determining whether a "candidate" patient visit v_j is relevant to (i.e. satisfies the criteria from) a given cohort description d_i requires access to a rich set of features derived from (1) the cohort description d_i, (2) the patient visit v_j, and (3) the interactions between d_i and v_j. To account for the variation between cohort descriptions, we considered multiple strategies for transforming d_i into queries. Let $q_i^{c,e}$ represent the query obtained when using query construction method c and query expansion method e. Likewise, we considered multiple strategies for representing the information encoded in each visit v_j. Hence, we considered r_k^s the textual content provided by stream s of the electronic health record r_k, and define $v_j^s = \left\{r_1^s,\ r_2^s,\ \cdots,\ r_{N_j}^s\right\}$ as the content of stream s from each report associated with visit v_j. We produced a single feature vector $\boldsymbol{x}_j^i$ encoding information about d_i and v_j by extracting the 14 high-level multivalued features listed in Table 2.

As shown, 10 of the 14 features illustrated in Table 2 are multivalued, i.e. consist of distinct values for each possible query representation $q_i^{c,e}$ of d_i and each stream s of v_j^s (where applicable). Each of these values corresponds to a single entry in the resultant feature vector, i.e. F_1 corresponds to 5 entries in the generated feature vector. Moreover, features F_3, F_6 and F_{10}-F_{14} capture the distribution of feature values extracted for each component of the query (F_1) or for each report associated with the hospital visit (F_6, F_{10}-F_{14}) using 5 *aggregation methods* (described below). Of note are features F_{10}-F_{14}, which incorporate standard relevance models from IR to measure the relevance between the criteria in $q_i^{c,e}$ and each stream of visit v_j^s.

Aggregation methods. To capture the distribution of feature values obtained using different streams or for each report associated with a candidate visit, we considered 5 aggregating statics $A = \{mean,\ minimum,\ maximum,\ variance,\ sum\}$.

The learning relevance model

The role of the LRM is to infer a *relevance score* s_j^i between every candidate visit v_j and the cohort description d_i using the feature vector $\boldsymbol{x}_j^i$ extracted above. This is accomplished by using the *pairwise* strategy of learning-to-rank.[10] Given (1) feature vectors $\left[\boldsymbol{x}_1^i,\ \cdots,\ \boldsymbol{x}_N^i\right]$ associated with candidate visits $[v_1, \cdots,\ v_N]$ and (2) "gold-standard" relevance judgments $\left[y_1^i,\ \cdots,\ y_N^i\right]$ indicating the relevance of each candidate visit to d_i, the RF is trained to infer the scores $\left[s_1^i,\ \cdots,\ s_N^i\right]$, which result in the optimal ordering of hospital visits as indicated by $\left[y_1^i,\ \cdots,\ y_N^i\right]$. Additional information about the pairwise learning-to-rank strategy, the RF, and model parameters is provided in Supplementary Material Appendix F. After training, the LRM uses the RF to produce the final ranked list of hospital visits by (1) inferring the relevance score s_j^i for each candidate visit v_j retrieved for d_i and (2) returning the L highest scoring visits (in our experiments we used $L = 1\,000$).

RESULTS

We evaluated the performance of the L-PCR system when automatically identifying patient cohorts in 2 settings: (1) a neurology-specific setting operating exclusively on EEG reports and (2) a

Table 2. Features extracted for a cohort description d_i and hospital visit v_j

	Feature description	Domain of values
(Features encoding information about the cohort description $\boldsymbol{d_i}$)		
F_1	Number of **criteria** detected in cohort description d_i with each construction method c	$\mathbb{N}^{\lvert C\rvert}$
F_2	Number of **terms** in $q_i^{c,e}$ for each $c \in C$, and each expansion method $e \in E$	$\mathbb{N}^{(\lvert C\rvert\times\lvert E\rvert)}$
F_3	*Statistics* of the **normalized inverse document frequency** (**IDF**) of $q_i^{c,e}$ in each stream $s \in S$ for each $c,\ e$.	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$
(Features encoding information about the candidate visit $\boldsymbol{v_j}$)		
F_4	Number of **reports** associated with v_j	$\mathbb{N}$
F_5	Distribution of **report types** associated with v_j	$\mathbb{R}^{\lvert T\rvert}$
F_6	*Statistics* of the **number of words** in each $r^s \in v_j^s$ for every s	$\mathbb{N}^{(\lvert A\rvert\times\lvert S\rvert)}$
(Features encoding the relationship between the cohort description $\boldsymbol{d_i}$ and candidate visit $\boldsymbol{v_j}$)		
F_7	Whether the **age** (if any) specified in cohort description i matches the age in any stream of any report $r^s \in v_j$	$\{0,\ 1\}$
F_8	Whether the **gender** (if any) specified in cohort description i matches the most frequently-mentioned gender in any stream of any report $r^s \in v_j$	$\{0,\ 1\}$
F_9	Whether the **hospital status** in cohort description i matches the hospital status in any stream of any report $r^s \in v_j$	$\{0,\ 1\}$
F_{10}	*Statistics* of the **Dirichlet**-smoothed **language model similarity**[37] (**LM: Dir**) between $q_i^{c,e}$ and each $r^s \in v_j$ for every $c,\ e,\ s$	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$
F_{11}	*Statistics* of the **Jelinek-Mercer**-smoothed **language model similarity**[37] (**LM: JM**) between $q_i^{c,e}$ and each $r^s \in v_j$ for every $c,\ e,\ s$	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$
F_{12}	*Statistics* of the **BM25 similarity**[38] between $q_i^{c,e}$ and each $r^s \in v_j$ for every $c,\ e,\ s$	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$
F_{13}	*Statistics* of the **TF-IDF similarity**[7] between $q_i^{c,e}$ and each $r^s \in v_j$ for every $c,\ e,\ s$	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$
F_{14}	*Statistics* of the **term frequency** (**TF**) between $q_i^{c,e}$ and each $r^s \in v_j$ for every $c,\ e,\ s$	$\mathbb{R}^{(\lvert A\rvert\times\lvert C\rvert\times\lvert E\rvert\times\lvert S\rvert)}$

Additional details for each feature are provided in Supplementary Material Appendix E. $\mathbb{N}$ represents the natural numbers, $\mathbb{R}$ represents the real numbers, and the exponent (if provided) indicates the *dimensionality*, or number of values produced by that feature in the resultant feature vector).

Table 3. Patient cohort retrieval performance on (a) EEG reports and (b) TRECMed

Setting	MAP	NDCG	BPref	rPrec	P@10
(a) Retrieval performance when retrieving patient cohorts from EEG reports					
BM25 baseline: 10-fold CV	0.4996	0.6144	0.4064	0.5213	0.6
L-PCR: 10-fold CV	0.6634	0.7171	0.5900	0.6088	0.6
MERCuRY (text-only): 10-fold CV	0.5220	0.5441	0.4483	0.5081	0.5
(b) Retrieval performance when retrieving the patient cohorts using in TRECMed from Hospital Records.					
BM25 baseline: evaluated on 2011	0.4052	0.5202	0.5082	0.4112	0.600
BM25 baseline: evaluated on 2012	0.2930	0.3462	0.3462	0.3135	0.464
L-PCR: 10-fold CV on 2011	0.6316	0.8816	0.5788	0.5859	0.706
L-PCR: 10-Fold CV on 2012	0.5100	0.8194	0.4703	0.5028	0.589
L-PCR: trained on 2012 and evaluated on 2011	0.6127	0.8675	0.5638	0.5763	0.674
L-PCR: trained on 2011 and evaluated on 2012	0.5145	0.8167	0.4735	0.5072	0.596
Best submitted to TRECMed 2011	—	—	0.5502	0.4400	0.656
Best submitted to TRECMed 2012	0.2860	0.5780	—	—	0.592

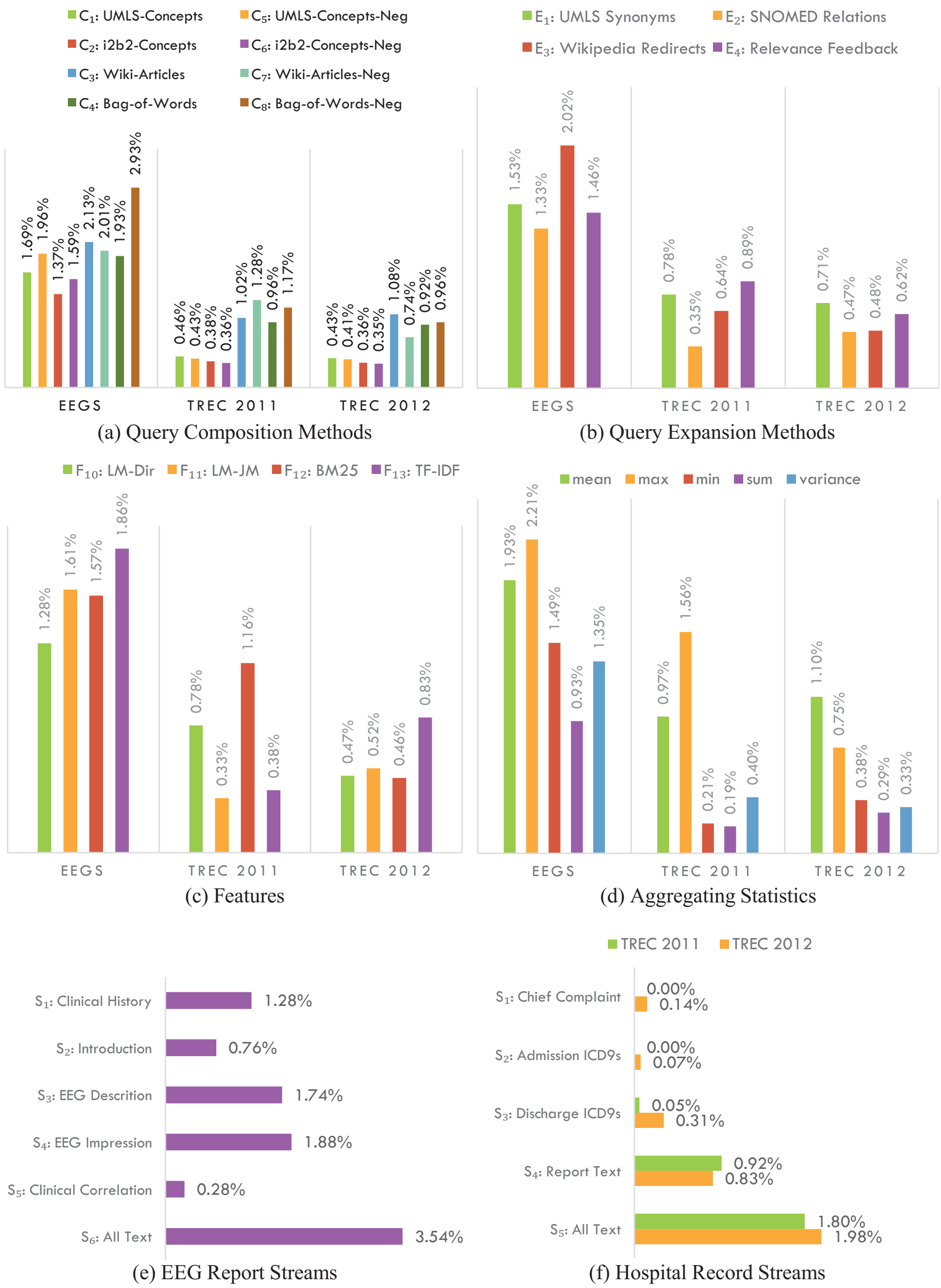
general hospital setting associated with a variety of EHR types. In both settings, we measured the performance of our approach using 5 measures commonly used to evaluate IR systems[39]: (1) the Mean Average Precision[7] (MAP), (2) the Normalized Discounted Cumulative Gain[7] (NDCG), (3) the Binary Preference[40] (BPref), (4) the *R*-Precision[7] (rPrec), and (5) the Precision within the first 10 returned visits[9] (P@10); details and discussion of these metrics are provided in Supplementary Material Appendix G.

Patient cohort retrieval from EEG reports

We measured the performance of L-PCR system for neurology-focused cohorts by considering EEG reports from the TUH EEG Corpus[28] using the relevance judgments described in *Datasets and Experimental Settings*. Table 3(a) presents the performance of the L-PCR system compared to (1) a BM25 baseline and (2) a text-only variant of MERCuRY[4] (a multi-modal retrieval system incorporating polarity information and the BM25 ranking function) using cross validation at the cohort description level (e.g. we evaluated on each set of 3 cohort descriptions when training on the remaining 27). Note: the cohort descriptions used to evaluate MERCuRY in this study are different and more complex than those used in Goodwin and Harabagiu.[4]

Patient cohort retrieval from hospital EHRs

We evaluated the performance of our approach in a general hospital setting using the patient cohort descriptions produced for TRECMed during 2011 and 2012 (described in *Datasets and Experimental Settings)*. We evaluated the performance of the L-PCR system using the cohort descriptions produced during both years of the evaluation with 10-fold cross validation (using the same strategy described above), as well as when training on the cohort descriptions



(a) Query Composition Methods

(b) Query Expansion Methods

(c) Features

(d) Aggregating Statistics

(e) EEG Report Streams

(f) Hospital Record Streams

Figure 5. Average feature importance as measured using EEG reports and TRECMed hospital records. EEG: electroencephalography.

produced during 2011 and testing on the cohort descriptions produced in 2012 (and vice versa). Table 3(b) presents these results, as well as performance of (1) a BM25 baseline system and (2) the best submitted systems reported by NIST for each year, which to the best of our knowledge are still state-of-the-art. An overview of these systems is provided in Supplementary Material Appendix J.

Measuring feature importance

To analyze the impact of the different techniques and features used by the L-PCR system, we measured the Gini importance[41] of each entry in all feature vectors extracted from the training data. Figure 5 illustrates the normalized average Gini importance of different (1) query construction methods, (2) query expansion methods, (3) features, (4) aggregating statistics, or (5/6) streams for each collection of EHRs. Supplementary Material Appendix I provides the most important features for each experiment.

DISCUSSION

In this article, we presented a learning PCR system which incorporates machine learning operating on features encoding thousands of different strategies for representing queries, EHRs, and their interactions to learn how to rank patient cohorts based on clinicians' feedback. As shown by Table 3, the L-PCR surpassed previously published state-of-the-art NDCG score by 27.1% on EEGs and by

52.5% on TRECMed 2011. It is interesting to note that for both years of TRECMed, there was no statistically significant change in performance when trained using cross validation, or using the cohort descriptions from the previous/following year ($P < .001$).

It is clear from Figure 5(a) that bag-of-words obtained the best performance compared to other query construction methods when processing cohorts from EEGs. By contrast, for general EHRs, Wikipedia titles provided higher performance. This reinforces our observations that many neurological phenomena (e.g. "spike and wave") are not associated with entities in structured knowledge bases. Moreover, the fact that, for EEGs, Wikipedia redirects of individual words in the cohort description provided the most informative expansions suggests that when structured information is available, it greatly improves performance. This is clearly demonstrated by the TRECMed cohorts, where the most impactful query expansion method was based on UMLS, followed closely by PRF. Interestingly, the impact of each individual features, as shown in Figure 5(c), varied between all 3 experiments, suggesting that the choice of relevance model is less important than the choice of query composition and expansion strategies. Overall, we found that query expansion resulted in a 6.0% (relative) increase in performance compared to no query expansion for TRECMed, and an 8.7% increase for EEG reports.

One key difference between automatic PCR systems and traditional IR systems is that each patient may be associated with multiple medical records. When analyzing different methods of aggregating features from report- to visit-level, the importance of the *variance* statistic, especially when contrasted with the *sum* statistic, suggests that the reports associated with a visit often have unequal impacts on the relevance of the visit. Moreover, as the *sum* measure closely resembles the effect of merging all reports associated with a visit into a single document (a common strategy employed by TRECMed participants), our results show that treating reports separately can be advantageous. In terms of streams, as shown in Figure 5(e–f), the ALL TEXT stream provided the most information to the model, suggesting that differentiating between streams added little value to the L-PCR system.

Error analysis

We analyzed errors made by the L-PCR system for both EHRs collection. In the TUH EEG collection, we observed several common phenomena. The first was that many of the cohort criteria were not found in existing ontologies and were often incorrectly parsed. For example, "PLED" indicates "periodic lateralized epileptiform discharges", but was not present in UMLS, SNOMED, Wikipedia, or even the Epilepsy and Seizure Ontology (EpSO).[42] A more significant source of errors was accounting for the fact that EEG cohorts are typically characterized by the *attributes* of waveform activity, rather than their presence or absence. In the previous example the "lateral" attribute highlights a major phenomenon in EEG reports—the role of spatiotemporal information. In EEG reports, the term "lateral" may not be mentioned; instead, activity is often described as occurring in specific lobes of the brain or at specific channels/nodes in the EEG. Moreover, consider that for a visit to be relevant to the criteria "generalized periodic epileptiform discharges", it is not sufficient for "epileptiform discharges" to be described. The discharges must be generalized and periodic. Each of these attributes can be described in multiple ways, for example, both "non-localized", "diffuse" indicate generalized activity.

We also observed in both collections that the performance of the L-PCR (and each baseline system) varied significantly between cohort descriptions. We found that cohort descriptions which qualified their criteria with anatomical (e.g. "lower extremity") or temporal ("history of") attributes were harder to retrieve. Moreover, we observed that the most difficult cohort descriptions were those that described relations between concepts: i.e. "inflammatory disorders receiving TNF-inhibitor treatment" or "cancer patients with liver metastasis treated in the hospital who underwent a procedure". For an in-depth analysis of errors encountered by participants of the TRECMed evaluation, we refer the reader to Edinger *et al*.[5]

Limitations

This study has several limitations. First, only 111 cohort descriptions were evaluated. While every effort was made to produce and evaluate realistic patient cohort descriptions, the time and cost associated with producing relevance judgments limited the number of cohorts that could be evaluated. Consequently, the performance results reported in Table 3 represent results of a pilot study only on 2 sets of EHR collections and may not be reflective of performance on other EHR collections. Second, many of the features extracted in our study rely on individual *streams* in the EEG/hospital reports which vary between hospitals and EHR collections. Fortunately, in our experiments, the top-performing features all relied on the ALL TEXT stream, which can be easily generalized across EHR collections. Finally, because the TRECMed collection is no longer available, we were unable to consider some promising features, such as word embeddings[43] or convolutional neural network features. In future work, we shall investigate the impact of concept embeddings on the performance of the L-PCR system.

CONCLUSION

Learning-to-rank can be successfully applied for retrieving patient cohorts from EHR when (1) judgments of relevance are available; and (2) a rich set of features is considered. In this paper, we present L-PCR system, and our experimental results on 2 EHR collections demonstrate that the L-PCR system obtains results 27–53% above state-of-the-art PCR systems when retrieving the same cohorts from the same EHR collections. Moreover, by analyzing the performance of the L-PCR system, we were able to measure the impact of a variety of PCR techniques. Overall, our results indicate the promise of the L-PCR system, but also reveal potent avenues for further improvement.

Conflict of interest statement. The authors have no conflicts of interest to declare.

CONTRIBUTORS

T.G. and S.H. originated the study. T.G. conducted the experiments. S.H. reviewed and help analyze the findings. T.G. and S.H. wrote the first and revised drafts of the manuscript.

FUNDING

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGEMENTS

The authors would like to thank the Neural Engineering Data Consortium at Temple University and the TExt Retrieval Conference (TREC) Medical Records Track (TRECMed) organizers for providing the data used in our experiments.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online. Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/ dryad.pq0cs6h.

REFERENCES

1. Tatum WO. *Handbook of EEG Interpretation*. 2nd Ed. New York, NY: Demos Medical Publishing; 2014.
2. Yamada T, Meng E. *Practical Guide for Clinical Neurophysiologic Testing: EEG*. Philadelphia, PA: Lippincott Williams & Wilkins; 2012.
3. Beniczky S, Hirsch LJ, Kaplan PW, *et al*. Unified EEG terminology and criteria for nonconvulsive status epilepticus. *Epilepsia* 2013; 54: 28–9.
4. Goodwin TR, Harabagiu SM. Multi-modal patient cohort identification from EEG report and signal data. *AMIA Annu Symp Proc* 2016; 2016: 1794–803.
5. 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.
6. Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014; 21 (2): 221–30.
7. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press; 2008.
8. Voorhees EM. The trec medical records track. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM; 2013: 239.
9. Voorhees EM, Tong RM. Overview of the TREC 2011 Medical Records Track. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
10. Liu T-Y. *Learning to Rank for Information Retrieval*. 1st ed. Springer-Verlag Berlin Heidelberg; 2011. http://link.springer.com/10.1007/978-3-642-14267-3 Accessed Aug 5, 2015.
11. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the American Medical Informatics Association Annual Symposium. 2001: 17–21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/ Accessed Mar 9, 2015.
12. Cohen KB, Christiansen T, Hunter LE. MetaMap is a Superior Baseline to a Standard Document Retrieval Engine for the Task of Finding Patient Cohorts in Clinical Free Text. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
13. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993; 32 (4): 281–91.
14. Karimi S, Martinez D, Ghodke S, *et al*. Search for Medical Records: NICTA at TREC 2011 Medical Track. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
15. Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001; 34 (5): 301–10.
16. Bedrick S, Ambert KH, Cohen AM, *et al*. Identifying Patients for Clinical Studies from Electronic Health Records: TREC Medical Records Track at OHSU. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
17. Córdoba JM, Maña López MJ, Mata López J, *et al*. Medical-miner at TREC 2011 Medical Records Track. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
18. Zhang J, Lin X, Zou Y, *et al*. PRIS at TREC 2011 Medical Record Track. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
19. Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab; 1999.
20. Martinez D, Otegi A, Soroa A, *et al*. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *J Biomed Inform* 2014; 51: 100–6.
21. Kanerva P, Kristoferson J, Holst A. Random indexing of text samples for latent semantic analysis. In: Proceedings of the Annual Meeting of the Cognitive Science Society. 2000.
22. Wu S, Masanz J, Ravikumar KE, *et al*. Three questions about clinical information retrieval. In: Proceedings of the Twenty-first Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2012.
23. Jakarta A. Apache Lucene-a high-performance, full-featured text search engine library. *Apache Lucene* 2004.
24. Strohman T, Metzler D, Turtle H, *et al*. Indri: a language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis. Amherst, MA, USA; 2005: 2–6.
25. Ounis I, Amati G, Plachouras V, *et al*. Terrier: a high performance and scalable information retrieval platform. In: Proceedings of the OSIR Workshop. 2006: 18–25.
26. Goodwin T, Rink B, Roberts K, *et al*. Cohort Shepherd: discovering cohort traits from hospital visits. In: Proceedings of the Twentieth Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2011.
27. Goodwin T, Roberts K, Rink B, *et al*. Cohort Shepherd II verifying cohort constraints from hospital visits. In: Proceedings of the Twenty-first Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2012.
28. Harati A, Choi S-M, Tabrizi M, *et al*. The Temple University Hospital EEG Corpus. In: Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE. IEEE; 2013: 29–32.
29. Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. *Front Neurosci* 2016; 10; doi: 10.3389/fnins.2016.00196
30. Voorhees EM, Hersh WR. Overview of the TREC 2012 Medical Records Track. In: Proceedings of the Twenty-first Text REtrieval Conference. National Institute of Standards and Technology (NIST); 2012. http://skynet.ohsu.edu/~hersh/trec-12-med.pdf Accessed Mar 10, 2016.
31. Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. *J Am Med Inform Assoc* 2011; 18 (5): 568–73.
32. Goodwin T, Harabagiu SM. Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records. In: 2013 IEEE Seventh International Conference on Semantic Computing (ICSC). IEEE; 2013: 363–370.
33. Goodwin T, Harabagiu SM. The impact of belief values on the identification of patient cohorts. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer; 2013: 155–166.
34. Vincze V, Szarvas G, Móra G, *et al*. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *J Biomed Semantics* 2011; 2 (Suppl 5): S8.
35. Stearns MQ, Price C, Spackman KA, *et al*. SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001: 662.
36. Qin T, Liu T-Y, Xu J, *et al*. LETOR: a benchmark collection for research on learning to rank for information retrieval. *Inf Retrieval* 2010; 13 (4): 346–74.
37. Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2001: 334–342.
38. Robertson SE, Walker S, Jones S, *et al*. Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference. National Institute of Standards and Technology (NIST); 1995: 109–109.

39. Voorhees EM, Harman DK, others. *TREC: Experiment and Evaluation in Information Retrieval*. MIT press Cambridge; 2005. http://www.aclweb.org/anthology/J/J06/J06-4008.pdf Accessed Mar 10, 2016.
40. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2004: 25–32.
41. Breiman L. Random forests. *Mach Learn* 2001; 45 (1): 5–32.
42. Sahoo SS, Lhatoo SD, Gupta DK, *et al*. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. *J Am Med Inform Assoc* 2014; 21 (1): 82–9.
43. Glicksberg BS, Miotto R, Johnson KW, *et al*. Automated disease cohort selection using word embeddings from Electronic Health Records. *Pac Symp Biocomput* 2018; 23:145–56.