Intrainstitutional EHR Collections for Patient-Level Information Retrieval

Stephen Wu
Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97214, USA. E-mail: wst@ohsu.edu

Sijia Liu
Mayo Clinic, 200 1st Street SW, Rochester, MN, USA. E-mail: liu.sijia@mayo.edu

University at Buffalo, The State University at New York, 338 Davis Hall, Buffalo, NY, USA. E-mail: sijialiu@buffalo.edu

Yanshan Wang
Mayo Clinic, 200 1st Street SW, Rochester, MN, USA. E-mail: wang.yanshan@mayo.edu

Tamara Timmons
Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97214, USA. E-mail: timmonst@ohsu.edu

Harsha Uppili
Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97214, USA. E-mail: harsha.uppili@gmail.com

Steven Bedrick
Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97214, USA. E-mail: bedricks@ohsu.edu

William Hersh
Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97214, USA. E-mail: hersh@ohsu.edu

Hongfang Liu
Mayo Clinic, 200 1st Street SW, Rochester, MN, USA. E-mail: liu.hongfang@mayo.edu

Research in clinical information retrieval has long been stymied by the lack of open resources. However, both clinical information retrieval research innovation and legitimate privacy concerns can be served by the creation of intrainstitutional, fully protected resources. In this article, we provide some principles and tools for information retrieval resource-building in the unique problem setting of patient-level information retrieval, following the tradition of the Cranfield paradigm. We further include an analysis of parallel information retrieval resources at Oregon Health & Science University and Mayo Clinic that were built on these principles.

Received August 2, 2016; revised December 30, 2016; accepted March 8, 2017

 DOI: 10.1002/asi.23884

Introduction

Research in clinical information retrieval (IR) from electronic health records (EHRs) has long been stymied by the

lack of open resources. Restrictions on the use of patients' private health information (e.g., in EHRs), paint a vastly different resource landscape than the more public traditional domains for IR (e.g., web).

We believe that both clinical IR research innovation and legitimate privacy concerns can be served by the creation of intrainstitutional, fully protected resources. In this article, we provide some principles and tools for resource building in the unique problem setting of patient-level IR. These arise in part from the experience of creating such resources at two sites: Oregon Health & Science University (OHSU) and Mayo Clinic. We report them here as a guidepost for other institutions who may choose to address resource development in patient-level clinical IR.

We follow the long tradition in IR of test collections and challenge evaluations, specifically structuring our resources for Cranfield-style IR evaluations (Cleverdon & Keen, 1966; E. Voorhees, Harman, & National Institute of Standards and Technology (U.S.), 2005). Cranfield evaluations require (a) a set of documents, (b) a set of test topics, expressed as queries, and (c) judgments of whether documents are relevant, for each query.

Each of these three collection components needs to be reenvisioned for its role in patient-level IR. In the remainder of our paper, we will walk through the three components and provide principles for their design, highlighting how they differ from a "traditional" test collection. We will then describe the implementation of these principles in our own multi-institutional resource building project. We will end with some analysis of the corpora and the processes used to produce them.

Our long-term goal is to enable data-driven investigations of patient health and disease by providing resources and tools for working with clinical text. So, this article's emphasis on intrainstitutional collection construction leaves an open question: how can collections be shared across institutions? The work we present here is designed to be compatible with the promising Evaluation-as-a-service (EaaS) paradigm (Hanbury et al., 2015; Lin & Efron, 2013). In contrast to standard IR evaluation campaigns which send data to systems for evaluation, EaaS sends systems to data. However, we leave shareability issues to future work.

Related Work

Efforts to make medical language corpora available have often been developed by focused, funded projects, and have been disseminated through their use in evaluation challenges. Initially, most of the effort in producing medical text resources arose for Natural Language Processing (NLP) tasks. Notably, the i2b2 (Informatics for Integrating Biology and the Bedside) NLP challenges made use of medical records from multiple institutions for tasks like named entity recognition (Uzuner, South, Shen, & DuVall, 2011), coreference resolution (Uzuner et al., 2012), temporal relations (Sun, Rumshisky, & Uzuner, 2013), or disease-specific end-to-end processing (Stubbs, Kotfila, Xu, & Uzuner, 2015). These corpora were fully de-identified and relatively small compared to modern IR collections (the largest of the i2b2 datasets had 1,304 patient records) because their focus was on adding linguistic or domain annotations—or even on de-identification itself (Stubbs & Uzuner, 2015).

The CLEF (Conference and Labs of the Evaluation Forum) eHealth series of challenges (Goeuriot et al., 2014; Goeuriot et al., 2015; Mowery et al., 2014; Suominen et al., 2013) also used a number of clinical text resources. The MIMIC-II database (Saeed et al., 2011) contains automatically deidentified patient records, and was used in CLEF eHealth to evaluate NLP-oriented tasks like named entity recognition and template filling (Mowery et al., 2014; Suominen et al., 2013). Task 3 in 2013 applied MIMIC-II data to IR, by using discharge summaries to provide context for search topics (Goeuriot et al., 2013). However, the data in MIMIC-II was primarily gathered to be an intensive care unit (ICU) research database, with multimodal information primarily describing 25,328 ICU stays; thus, it cannot reliably support clinical information needs within EHR data if they are not ICU-related. It is important to note that MIMIC-II data is contributed by a single institution, Beth-Israel Deaconness Medical Center; it is a rare case for an Institutional Review Board at an academic medical center to approve the release of such a large scale of automatically de-identified data with as unrestrictive a license. Furthermore, expanding out from the ICU setting provides additional complexities in patient identifiability in the data. Therefore, MIMIC-II does not serve as a pattern for building multi-institutional text-related resources.

It should also be said that although the IR-oriented CLEF eHealth challenge tracks (Goeuriot et al., 2013, 2014; Palotti et al., 2015) were about finding medical information, the problem setting was still traditional web search. This differs from our setting in that it does not focus on EHR data or a health provider's needs, but rather on consumer health.

Annotation efforts such as MiPACQ (Albright et al., 2013), SHaRe (Suominen et al., 2013), SHARPn (Chute et al., 2011), and THYME (Bethard et al., 2016) provided layers of NLP-oriented annotation; many of these projects were collaborative efforts with the aforementioned challenges and corpora. These corpora and task settings differ from ours in that they evaluate on specific linguistic or medical structures within text, rather than on retrieval at a patient level, and they are typically on much smaller amounts of data.

Other shared tasks utilized multimodal and image search resources in medical records, with interesting methods of dealing with confidentiality concerns. CLEF eHealth 2015 Task 1a used synthetic data from the NICTA Synthetic Nursing Handover Data, rather than dealing with protected health information; Task 1b used the QAERO French Corpus, which chose publicly available drug information, patents, and biomedical publications rather than dealing with confidentiality concerns in an EHR setting (Goeuriot et al., 2015). The VISCERAL project (www.visceral.eu), a multimodal resource which annotated radiology images,

DOI: 10.1002/asi

somewhat sidestepped confidentiality concerns by preprocessing the identified text and giving term lists, essentially providing an immutable upstream NLP process. Another strategy, used by the Khresmoi project (Hanbury, Boyer, Gschwandtner, & Müller, 2011), was to choose an IR information-seeking activity that did not require protected health information. This approach was exhibited in both their multilingual search queries and summaries, and the web documents for the CLEF eHealth IR challenges through 2015.

Most like the resources developed in this paper is the Pittsburgh NLP Repository, in that it is a collection of EHR records built for Cranfield-style IR evaluations. It was distributed exclusively for the 2011-2012 TREC Medical Records Tracks (Voorhees & Hersh, 2012; Voorhees & Tong, 2011), and sought to address patient confidentiality concerns by providing a de-identified clinical IR collection usable by the research community (albeit with limited availability) for a patient cohort retrieval task. By design, searching this collection bore the marks of patient confidentiality concerns: retrieve hospital visits—as a stand-in for patients—in response to a query. Although this was an important step forward in that it moved beyond document retrieval, it did not (and could not) fully embrace patient-level IR. Furthermore, it is no longer available for research activities.

None of these previously constructed shared corpora can support the connectedness, granularity, and scope of information that is possible in the intrainstitutional corpora that we propose here.

Design Principles

Here, we outline the structure and assumptions of the EHR-based IR collections. These principles touch on each of the three components of Cranfield evaluations. Although some principles were implied by the desired use cases of such collections, others arose from our experience implementing such collections across the two institutions.

Principle 1 - Significance: Finding Patients or Cohorts in a Patient Population

Epidemiological investigations typically examine a patient population to determine the distribution and determinants of diseases. This population-level setting is implicit in an EHR-based IR collection; namely, patients whose records are part of the collection constitute a population that may be studied. Because of this, one must be mindful of which patients' EHR data are included. A catch-all "all patients that have ever been documented at tertiary care hospital X" is likely to constitute a highly irregular patient population with significant missing data for a large number of patients. At the opposite extreme, a convenient "only patients who have an ICU visit" may be biased in its representation of comorbidities.

Subgroups of a population may be sought for various reasons. Clinical research frequently uses cohorts of patients to study a disease or disorder—these are subgroups with additional inclusion and exclusion criteria. Alternatively, a physician may seek to pull up the EHR record for a recently treated patient for subsequent care—this is a diminutive subgroup of one individual.

Practically, our principle of considering an IR collection to be a patient population translates into sets of primary care patients with sufficient data at a given medical institution; primary care patients are more likely to constitute a geographically cohesive, demographically representative, and phenotypically demonstrative population. In this setting, a boolean (or thresholded) search for patients matching some criteria may then be considered a cohort.

Of course, it is possible to build a cohort-level (rather than population-level) IR collection; searches would then help characterize that cohort. However, the institutional investment in building an IR collection is better justified when multiple cohorts of patients can be found from a population; furthermore, the population-level collection can still characterize cohorts by means of simple drill-down searches.

Principle 2 - Scope: Retrieve Patient-Level EHRs

The prototypical ad-hoc search task is in libraries or for the web, where the desired unit of retrieval is a single document or website. However, patient-level IR is about finding patients, not just documents. A natural consequence to looking at whole patients is that diverse types of data records must be considered. Of the many documents and types of documents associated with a patient, any combination of the evidence might carry the information that is most salient to a given query. Examining only single documents of a patient's EHR may thus be insufficient. Therefore, the task of scoring and retrieving whole (multidocument) patients is a necessary part of the scope of clinical IR. This may be done directly, or scores may be aggregated from traditional document-level scores into patient-level scores.

Of course, other scopes of analysis are important for healthcare. At one end of the spectrum, a clinician-user may be interested in passage retrieval within a single patient's record. This is, of course, still situated within the context of a patient-level subset of the EHR, and is thus served by the creation of a patient-level collection as we have been proposing. At the other end of the spectrum, digital health information about a single patient may include patient portals with private messaging, online patient forums, and social media. Although these sources of information may certainly augment EHR data, our work focuses on overcoming the confidentiality barriers inherent to nonpublic EHR data, so search across such modalities is outside the scope of this work.

Principle 3 - Sharing: Keep Data Intrainstitutional

The third principle for IR test collections is that they should be built within institutions. Shareability should be achieved at a different point in the research life cycle. This is

DOI: 10.1002/asi

in response to the legal and ethical context of health information. For example, in the United States, health data is protected by laws of the Health Insurance Portability and Accountability Act (HIPAA), and therefore cannot be easily shared outside the secure servers of a healthcare institution. The Related Work section discusses several strategies that are used to solve this problem in other medical language corpora: deidentification, access through an interface, generating synthetic data, and selecting nonconfidential domains or data.

Each strategy yields unacceptable consequences for true patient-level search. Healthcare institutions are typically hesitant to share even de-identified data; known counterexamples are almost exhaustively listed in the Related Work. In all known cases, vital links between pieces of medical information are necessarily destroyed or obfuscated during any de-identification process, or in any interface that would present nonconfidential data. For example, MIMIC broke links beyond ICU data, and Pittsburgh's NLP Repository broke links beyond the level of a hospital visit; neither is a full picture of a patient's EHR. Synthetic data is difficult to generate at a scale viable for IR, and rarely comes from the same distribution as true patient data. Sidestepping the problem by choosing data from another domain still ignores the potential benefits of "unlocking" EHR data.

The principle, then, implies that data need not change hands or be improperly exposed; rather, research collaboration and sharing needs are left to be accomplished by a (later) step, such as by Evaluation-as-a-Service (EaaS) approaches.

Principle 4 - Topic Sources: Diversity and Practicality

In selecting and developing topics on which to assess relevance, one must consider how EHR-based cohorts retrieved by a system might be used. Real-world information needs display broad diversity: cohorts may be sought for research study recruitment, preliminary screening for a later manual review, evidence-based clinical care, or characterization of population health in epidemiological studies. Thus, topics in a patient-level IR collection should reflect this diversity of real-world use cases.

This principle can be served by selecting topics from information seekers (e.g., researchers and clinicians), or from information providers (e.g., data warehouses and management services). The common technique of analyzing and utilizing query log information is an example of using data from information providers, and could robustly augment an EHR system that already provides a search interface and stores query logs, though the authors have no knowledge of EHR systems that provide a search interface or store query logs.

Principle 5 - Topic Format: Diverse Topic Representations

Information needs from varied topic sources may come in a variety of different representations: text terms, structured data queries, etc. Inspired by past work on evaluating the stability of evaluation measures (Buckley & Voorhees, 2000), topics in clinical IR collections should have each topic (information need) expressed with a similar diversity of format. Without flexibility in topic representations, it is impossible to convey the intent of all original source topics in a lossless manner; for example, a structured query with tens of inclusion criteria cannot easily be cast as a simple phrase that is shorter than a sentence. It is possible to unintentionally misrepresent a topic when condensing or expanding, and this risk is compounded if there is only one acceptable granularity for topic representation. In practice, this means that a chosen topic must be cast in (a chosen set of) topic representations. This allows for faithful representation of the underlying information need, but it also enables us to calculate stable metrics and isolate the contribution of different topic representations.

Principle 6 - Assessment Task: Relevance Assessment as Chart Review

The assessment of relevance to a (cohort) query is tantamount to manual medical record (chart) review. Unlike traditional IR assessment (on a single document), many pieces of information (possibly thousands of text and structured data documents corresponding to a single patient) may be considered in making a judgment on relevance. Oftentimes during chart review, ad hoc rules or guidelines are developed; a reviewer weighs evidence for and evidence against an overall decision according to a medical-knowledge-informed logic. This dynamic will be present as long as it is medical records that are being assessed.

A corollary of this principle is that relevance assessments cannot be crowd-sourced, as is possible in some other domains. Chart reviews require specialized medical knowledge to be meaningful, and patient privacy concerns prevent the sharing of documents from a patient's EHR. Instead, patient-level IR collections should employ intrainstitutional medical experts to take on the costly step of patient-level relevance assessments. By keeping both the data collection and the relevance assessments intrainstitutional, patient confidentiality concerns are fully observed. This does not stipulate what incentives might compensate the medical experts for their relevance assessments.

Principle 7 - Assessment Tools: Visualizing Chart Review

An upshot of the principle above is that any improvements made to the process of relevance assessing are also process improvements for chart review. Manual chart reviews are a bottleneck in traditional clinical research settings because each member of a population needs to be evaluated for some number of diseases or symptoms. Chart review in medical research is often done with a clinician's full EHR interface, elaborate spreadsheets, and manual record keeping.

Instead, relevance assessments should be done with a user interface that streamlines the process of chart reviews.

DOI: 10.1002/asi

TABLE 1. Clinical notes document structure at OHSU.

XML field name	Description	Data type
OHSU_MRN	The MRN (medical record number) of the patient	String
SOURCE_SYSTEM_PAT_ID	The patient ID from the Epic database	String
SOURCE_SYSTEM_ENC_ID	The Epic database's unique identifier for an Encounter (visit)	String
SOURCE_SYSTEM_NOTE_CSN_ID	The unique identifier for the note, from the Epic database	String
NOTE_TYPE	The type of note (operative note, consults, op report)	String
NOTE_DATE	The date that the current version of the note was created	Date
NOTE_CREATED_DATE	The date that the original version of the note was created	Date
NOTE_FILING_ DATE	The date that the current version of the note was filed	Date
AUTHOR_NAME	The full name of the note author (usually null)	String
AUTHOR_SPECIALTY	The first specialty listed for the note author (usually null)	String
COSIGNER_NAME	The full name of the note cosigner (usually null)	String
COSIGNER_SPECIALTY1	The first specialty listed for the note cosigner (usually null)	String
NOTE_TEXT	The actual text of the note	String

Key to this is the visual display and navigation of complex data. It is also important that intermediate evidence and results can be "staged," and later, logically combined into a final patient-level assessment.

Principle 8 - Assessment Pools: Simulated Competitions

Cranfield-style evaluations such as those in TREC and CLEF typically utilize a set of runs (here, ranked lists of patients) from a shared task to determine a pool of documents (patients) to judge. In clinical IR, the lack of shareability and access means that there are no participant-submitted runs that we can use to select which patients will be assessed. Therefore, assessment pools in clinical IR should be drawn from simulated competitions with multiple baseline runs.

A few notes about this choice. First, when the included baseline runs are more diverse but relevant, the relevance assessments are more representative. Without diversity, new systems being evaluated will be disfavored. Without relevance, a topic may have insufficient examples of relevant patients to be useful in IR evaluation. Second, and perhaps more importantly, using simulated competitions decouples relevance assessing from competition organizing and its associated confidentiality concerns for medical data. This makes it feasible for complete clinical IR collections to be built without solving an a priori shareability problem.

Implementation

We describe the three Cranfield components (collection, topics, judgments) when implemented with the principles above. It should be noted that principles about the collection are implemented in parallel at two institutions: OHSU and Mayo Clinic. It may be possible to identify principles based on the experience of building a single IR collection; however, in our experience, many unknowns accompany the EHR data of any new institution or repository. Thus, parallel collections allow us to compare and contrast the practical effects of our resource design considerations.

Parallel Collections

OHSU collection. At OHSU, patients were included in the population if they had inpatient or outpatient encounters with primary care departments (Internal Medicine, Family Medicine, or Pediatrics), with three or more encounters and five or more text entries, between 1/1/2009 and 12/31/2013. This resulted in a population of 99,965 unique patients and 6,273,137 unique encounters. Documents are clinical text or other structured data generated by medical professionals documenting patient encounters. Document types from OHSU include both text and structured data: clinical notes, order result comments, demographics, ambulatory encounters, hospital encounters, encounter diagnoses, problem list, medications (ordered, current, recorded administrations), lab results, surgeries, vital signs, microbiology results, procedures, and imaging.

The OHSU data was initially collected from OHSU's Epic EHR and stored in the corresponding Clarity database. Patient-level structure is provided by linking all documents with a patient ID; thus, for each patient there are typically many documents of many data types (see statistics in the section on Collection Statistics). Within each document, there are multiple fields such as medical record number (patient ID), note text, lab results, or diagnosis. For example, Table 1 shows the document structure for the Notes; fields are either in string format or in date format.

Mayo Clinic collection. In the Mayo Clinic collection, patients were included in the population if they were under institutional-provided insurance at 01/01/2013 (i.e., employees and their family members) with research authorization. All clinical notes from 1998 to 2013 were gathered. This resulted in a population of 138,228 patients and 15,486,886 clinical notes. Clinical notes from Mayo Clinic include various textual documentations for all clinical encounters and are in CDA1.0 format where the sections are standardized across the institution. Each document in Mayo Clinic collection has an event type, which is partially comparable to document types in OHSU collection. Mayo Clinic collection has a total of 37 event types. Examples of these types are limited evaluation, miscellaneous, test-oriented

DOI: 10.1002/asi

miscellaneous, consult, multisystem evaluation and hospital admission note. Within each document, there are multiple sections. The Mayo Clinic collection includes 78 different sections, such as family history, diagnosis, immunizations, lab tests, vital signs, and current medication. The set of sections corresponding to each event type varies.

Interinstitutional relational mapping. The EHRs at OHSU and Mayo Clinic are significantly different in structure, with EHRs from different vendors (Epic vs. GE Centricity/Cerner) and institution-specific customizations. Structured data fields at OHSU such as vital signs or lab results contain a full record of entries with multiple subfields, whereas Mayo "fields" are sections of clinical notes, which often summarize findings. Thus, no fields can have a truly direct 1:1 relationship. Given this, consistency was crucial to identify useful and meaningful relationships between collections, which was accomplished through the categorization of fields; these relational mappings help determine how queries can be executed in different environments, and how results can be interpreted. Because the clinical relevance of different document types had to be considered, our lead relevance assessor at OHSU (coauthor and medical expert TT) did this mapping, with input from developers at both sites.

We established relationships between the parallel collections at the document type and field levels. For example, 15 fields at Mayo correspond to surgery information whereas at OHSU there are 11 fields in 3 document types. These share the concept of surgery information and report, but although operative reports are recorded in "notes:NOTE_TEXT", this field contains all notes so there is limited overlap. Similarly, surgery codes are well-documented in OHSU fields, but may not be included in each surgery report at Mayo.

All fields were assigned to one or more relevant criteria categories and subcategories. Criteria categories include demographics, diagnosis, encounter, labs, medications, and so on, whereas subcategories are similar to an individual criterion such as age, as a subcategory of demographics. The fields, sorted by categories and subcategories, were compared and matched into 3 relationship types: directly mapped, indirectly mapped, or conceptually related. When appropriate, the relationship may be with an OHSU document type, and each field or document type may have multiple relationships. A directly mapped relationship is between 1 Mayo field and 1 OHSU field or document, indicating that they are essentially equivalent. Indirectly mapped relationships may be many-many or 1-many, and share a category and subcategory. Last, conceptual relationships are also many-many or 1-many. These share a concept, but may be in different categories and overlap less than in an indirect relationship.

The resulting mappings can be seen in a supplement to this article. Future work on inter-institutional data mapping may require each institution to standardize its data to a common data model, rather than directly reconciling between two institutions.

Topics

A total of 56 test topics were developed based on defined patient cohorts drawn from five sources, illustrating a variety of use cases. Cohort descriptions from these sources, generally composed of eligibility criteria, serve as models for test topics.

Sources. Clinical study data requests, as submitted by researchers to the Oregon Clinical and Translational Research Institute (OCTRI), OHSU's Research Data Warehouse (RDW), provided the basis for 29 topics. One data request, out of the 30 provided by OCTRI, was excluded from development because it specified retrieval of clinic notes rather than individual patients. Additional topics were modeled after cohorts from the Phenotype KnowledgeBase (PheKB) (seven topics), Rochester Epidemiology Project (REP) (nine topics), and National Quality Forum (NQF) (12 topics). Finally, Mayo Clinic provided cohort descriptions from its own RDW to create two topics. Cohorts with similar characteristics were merged during topic development to avoid redundancy (one OHSU/REP topic, and two OHSU/PheKB topics), resulting in the total of 56 topics.

There is a significant level of variation in length, format, level of detail, and complexity among the cohort descriptions from these sources. Adapting these into a common framework allows for consistency among topics; maintaining the general eligibility criteria and objectives from the source description results in cohorts differing in subject matter, complexity, and precision.

The test topics produced through this process are therefore representative of diverse use cases and real-world information needs, with varied subjects and complexity presented in a consistent manner.

Formats. We provide the 56 topics with titles and three different formats for possible queries: (a) summary statement; (b) brief summary and clinical—a shorter summary statement plus a mock clinical case incorporating a patient and scenario that typify the topic criteria; (c) brief summary plus structured data—a summary statement plus criteria listed as defined or structured data field values.

For example, a topic concerning adults with rheumatoid arthritis is formatted in Figure 1.

All 56 topics, each in three topic representations, are included as supplemental material in a single XML file.

Semistructured queries. In IR collections, topics are the "official" representation of the information need; queries submitted to an IR system, on the other hand, may differ greatly and are typically left up to each system to determine. However, we include here further implementation details for site-specific, semistructured queries corresponding to representation C.

A full set of OHSU semistructured queries is included as supplementary material to this article. We believe this to be necessary because of our medical setting and its common

DOI: 10.1002/asi

A. Adults 18-64 years old with rheumatoid arthritis who have had a lab test for cyclic citrullinated peptide IgG antibody with a result greater than 40 units.

B. Adults 18-64 years old with rheumatoid arthritis and lab result for positive anti-CCP IgG > 40 units.

I. A 58-year-old female presents with morning stiffness and joint pain in her hands, especially her fingers, which improves after about 30 minutes, but doesn't remit fully. On examination she is found to have ulnar deviation, decreased grip strength, and joint tenderness over the MCP and PIP joints. She has a positive rheumatoid factor and is positive for anti-CCP Ab at 45 units.

C. Adults 18-64 years old with rheumatoid arthritis and lab result for positive anti-CCP IgG > 40 units.

I. Demographics inclusion
  a. Age: 18-64 years
II. Diagnosis inclusion
  a. Rheumatoid arthritis (ICD-9): 714.0
III. Lab inclusion
  a. Cyclic citrullinated peptide IgG antibody (anti-CCP IgG): > 40 units

FIG. 1. Three representations for topic 15.

use of structured queries. Cohort identification is commonly performed relying on only the information in structured data fields—using database queries which select records where the set of specified fields contain the given values. The third topic representation, C, includes eligibility criteria in a list form resembling this approach. Although representations A and B can be used as plain-text queries, significant work (involving significant knowledge of an underlying institution-specific data model) is involved in producing even a baseline query for representation C.

We intend that, by supplying a semi-structured query implementation of topic representation C, we enable other similar IR collections to make use of our topic representations. It should be noted that because these semistructured queries are site-specific, a mapping (such as the OHSU-Mayo mapping presented in the section on Parallel Collections) of the SSQ would also be necessary.

For example, Topic 21 has a topic representation C with criteria includes a laboratory procedure: "ALANINE TRANSAMINASE." The corresponding OHSU and Mayo fields in which to search have an indirectly mapped relationship, but they are both in the category "labs: lab names and codes" (see supplement). For OHSU, these are found in structured fields `lab_results:COMPONENT_NAME`, `procedures_ordered:PROC_NAME`, and `result_comments:PROC_NAME`. The related fields to specify at Mayo are sections `20146:Labs` and `20119:Admission_Findings_Test_Results`.

Relevance Assessments

Pooling by run sampling. Implementing the principle of simulated competitions of Principle 8, we varied four parameters to automatically create a diverse set of 48 runs per topic.

- **Topic representation**: {*A*, *B*, *C*} As the base query text, we utilized the entire contents of three different topic representations.
- **Text subset**: {*all*, *text*} The documents to match against (and thus retrieval statistics) could be filtered to include only text Notes (OHSU) or text-type events (Mayo Clinic), or all documents could be matched against.
- **Aggregation method**: {*sum*, *max*} After documents for a patient were scored, the function that combines scores for all of the documents belonging to that patient could be either a summation of scores (log probabilities) or a maximum.
- **Retrieval model**: {*BM*25, *DFR*, *LMDir*, *Lucene*} Despite some incongruences from the literature, we used Lucene's default implementations of BM25 (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1995), divergence from randomness (Amati & Van Rijsbergen, 2002), language modeling with Dirichlet smoothing (Zhai & Lafferty, 2001), and Lucene Default scoring.

Searches were carried out on each topic with these 48 different parameter settings, producing 48 ranked lists per topic. A pool for each topic was chosen from these 48 ranked lists by selecting, for each ranked list:

1. The top 15 patients (+15 to the pool)
2. 25% of the next 85 (+21 to the pool)
3. 1% of the next 900 (+9 to the pool)

As typical in TREC, ranked lists are limited to the 1000 top patients, so each of the 48 parameter settings adds at most 45 new patients to the pool of patients to be judged. Because of duplicates, and because ranked lists may contain fewer than 1,000 patients, the practical pool sizes are below the theoretical maximum of 2,160. Several alternative sampling methods were tested and refined, eventually producing these proportions; for instance, we found that evenly sampling from logarithmic depths in the runs (i.e., 100% of 10, 10% of 100, 1% of 1000, etc.) overemphasized irrelevant patients.

PRAI interface and usage. To support the process of patient-level relevance assessment, we have designed the EHR Patient Relevance Assessment Interface (PRAI). PRAI is a web application written in Rails; it is connected to a

DOI: 10.1002/asi

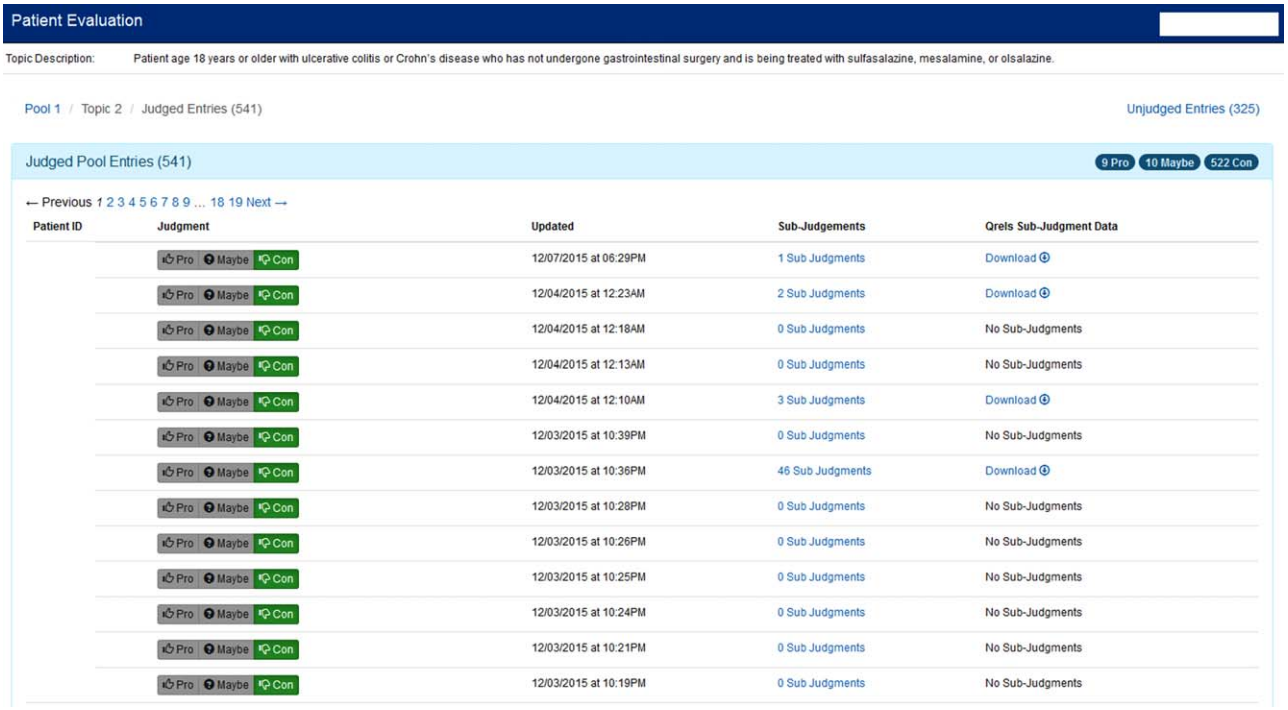

FIG. 2. Patient-level judgments (IDS removed). [Color figure can be viewed at wileyonlinelibrary.com]

PostgreSQL database for tracking judgments, and to Elasticsearch for retrieving patient data.

Patients selected for relevance judgment constitute a topic's patient pool in PRAI. The PRAI interface enables users to browse patient data much like they would in an EHR system, navigating within and between document types with the ability to search, filter, and sort.

PRAI allows users to record patient-level relevance judgments for a given topic and patient (see Figure 2). It also introduces the ability to perform "Sub Judgments" (document-level judgments, see Figure 3), whereby a single piece of data is marked as providing evidence in the overall judgment for the patient. A given sub judgment may concern criteria for patient inclusion or exclusion, and may support or contraindicate the patient's inclusion in the topic's cohort.

Patient- and document-level judgments are easily recorded by clicking on the relevant icon, and can be modified through the same process. Patient-level judgments can be recorded at multiple points enabling the medical expert to quickly make patient-level judgments when the criteria have been met.

Results and Analysis

Collection Statistics

Tables 2 and 3 show the document types gathered for the OHSU and Mayo Clinic collections, respectively. The OHSU collection in Table 2 lists patient percentages (column 2) with respect to total patients (column 1, row 1), total (ambulatory) encounter associated with each document type (column 3), and the number of records (column 4) per patient (column 5, equal to column 4 divided by column 1). with a population of just under 100,000 patients in the OHSU collection, we can see in Table 2 that >99% of patients have received an encounter diagnosis, have clinical notes written about them, and have recorded vitals. Conversely, only 18.6% of the population (18,640 patients) have had surgeries, only 27.5% have had microbiology results, and only 47.2% have had administered medications

It is important to note in Table 2 that a small number of Encounters per patient are not representative of the whole: although there were about one-fifth as many Lab Results as Encounters, 83.5% of patients had at least one Lab Result amongst their (average) 37.62 Encounters. This is a very coarse-granularity measure of missing data, but we can see that snapshot health data like a single Encounter does not fully describe patients' overall health.

Table 3 shows the statistics of the Mayo Clinic event types appearing in more than 25% of the total 138,228 patients. The event types are shown in descending order by the number of patients with at least one document of each type. Limited Evaluation and Miscellaneous are the most frequent included event types, which appear in 97.4% of the patient population. The event types of Vasectomy, Aspiration, and Nail Trimming are mentioned least frequently, included in only 1.1% of the patient records, which are around 1,500 unique patients.

Because the Mayo Clinic collection represents Mayo Clinic employees and their family members, the time spans of their EHR records are generally longer than those of the average Mayo Clinic patients. This can be reflected in the

DOI: 10.1002/asi

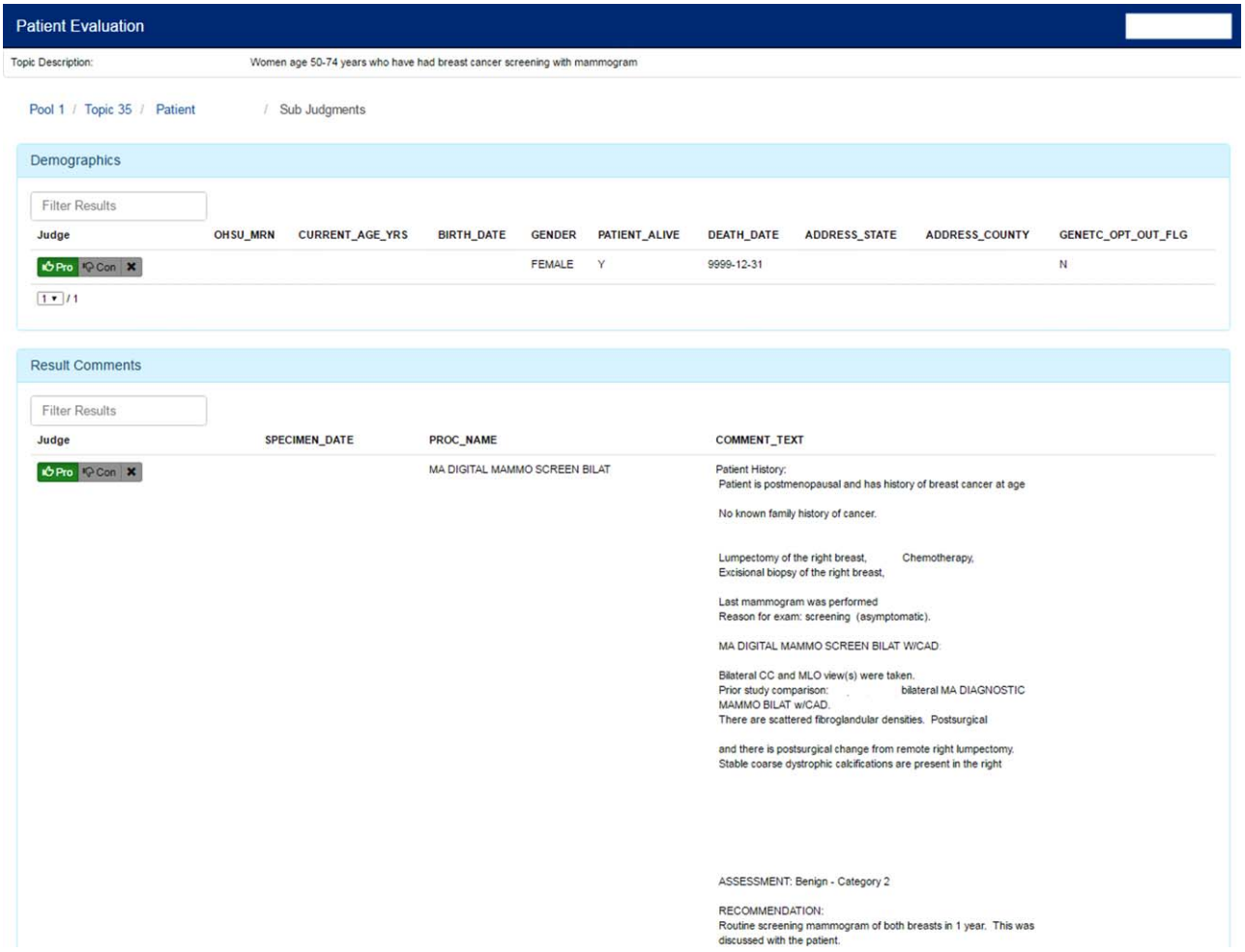

FIG. 3. Document-level "sub judgments" showing the demographics and problem list sections. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. OHSU collection data types and counts.

Type	Patients	%	Encounters	Records	Rec/Patient
Ambulatory Encounters	*99,965*	100.0	*3,760,205*	*3,760,205*	*37.62*
Demographics	99,965	100.0	-		-
Encounter Attributes	99,965	100.0	6,273,137	6,273,137	62.75
Encounter Diagnoses	99,938	100.0	3,725,603	18,170,896	181.77
Notes	99,868	99.9	3,491,659	10,111,930	101.15
Vitals	99,098	99.1	1,362,431	6,647,115	66.49
Procedures Ordered	98,514	98.5	1,880,309	7,229,854	72.32
Medications Ordered	94,089	94.1	1,388,086	5,336,506	53.38
Current Meds	92,783	92.8	-	31,997,402	320.09
Problem List	90,722	90.8	-	761,260	7.62
Lab Results	83,435	83.5	733,461	20,186,748	201.94
Hospital Encounters	73,303	73.3	466,252	466,252	4.66
Result Comments	72,716	72.7	468,814	916,554	9.17
Administered Meds	47,208	47.2	125,831	6,497,157	64.99
Microbiology Results	27,515	27.5	65,373	296,548	2.97
Surgeries	18,640	18.6	29,895	31,889	0.32

Patient IDs found in the (italicized) ambulatory encounters field were used to query the rest of the data from the epic-derived database. Thus, ambulatory encounter IDs were present in most data fields, with counts shown in the encounters column

large total number of documents for each patient. Therefore, the health records of these patients are more complete than a typical patient population.

From Table 3, we can see documents of the Multi-system Evaluation and Hospital Summary of Care types contain more sections than other types. This reflects the

DOI: 10.1002/asi

TABLE 3. Mayo clinic collection counts and statistics on event types, patient, document, and sections.

Event type	Patients	%	Documents	Documents per patient	Sections	Sections per document
Limited Evaluation	134,599	97.4	2,694,552	20.02	20,960,884	7.78
Miscellaneous	134,640	97.4	3,597,121	26.72	8,574,471	2.38
Test-Oriented Miscellaneous	124,985	90.4	1,417,670	11.34	3,830,922	2.70
Consult	124,492	90.1	1,211,562	9.73	8,213,886	6.78
Multi-system Evaluation	123,185	89.1	683,494	5.55	6,701,968	9.81
Subsequent Visit	115,336	83.4	1,881,387	16.31	9,096,844	4.84
Therapy	111,648	80.8	1,375,248	12.32	4,720,783	3.43
Supervisory	111,325	80.5	653,967	5.87	2,865,751	4.38
Hospital Summary of Care	68,146	49.3	170,054	2.50	1,720,427	10.12
Specialty Evaluation	65,306	47.2	169,673	2.60	1,360,788	8.02
Hospital Admission Note	46,499	33.6	161,802	3.48	1,238,223	7.65
Express Care Visit	45,705	33.1	112,465	2.46	739,689	6.58
Progress Note	45,211	32.7	607,509	13.44	1,859,236	3.06
Post Anesthesia Assessment	38,206	27.6	67,883	1.78	137,397	7.78

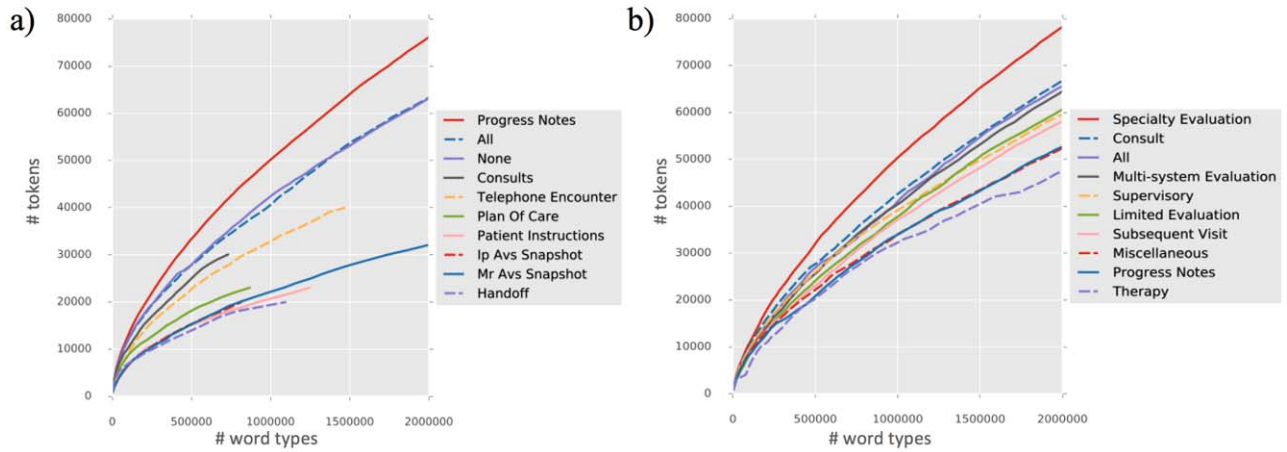

FIG. 4. Type-token plots for: (a) the top 10 note types in OHSU data; (b) the top 10 event types in Mayo Clinic data. "None" means no note type was listed. [Color figure can be viewed at wileyonlinelibrary.com]

comprehensive nature of the information these documents provide. Documents with types like Miscellaneous and Test-Oriented Miscellaneous are used to record brief information on a patient's care without a face-to-face interaction. Although frequently appearing in a patient's records, they have less than three sections per document in average, which is fewer than the number of sections in other documents. A *Limited Evaluation* is defined as an interaction focused on certain components of the patient's health history and certain body systems. It covers all noncomprehensive evaluation of a patient, and has 7.73 sections per document in average. With both a large number of documents per patient and a large number of sections per document, the number of total sections in Limited Evaluations is significantly larger than other event types.

Type-Token Plots

Both corpora include a significant amount of categorized textual data: OHSU notes are subdivided by note types, whereas Mayo Clinic documents are subdivided by event types. It is natural to ask if these text subcategories constitute medical sublanguages confined to a narrow subject matter (Friedman, Kra, & Rzhetsky, 2002), or if they instead exhibit robust lexical diversity. We attempt to answer this question using type-token plots (Cohen, Baumgartner, & Temnikova, 2016); this type of plot simulates reading a text sample and comparing the total number of observed tokens (x-axis) and the number of unique tokens (y-axis). If a text subcategory were to only contain unique terms, the type-token plot would be a straight line with a slope of 1, so any "bending" of the line indicates reuse of previously seen vocabulary. If a line tends to flatten out, there is only a "closed" set of words used in that type of text. If the line continues growing, the text sample is from an "open" domain and shows greater lexical diversity.

Figure 4 shows type-token plots for the clinical text (a) at OHSU and (b) at Mayo Clinic, with only the most frequent nine text subcategories at each institution, plus an "All" aggregation (i.e., includes documents from all subcategories in that institution). It is immediately apparent from Figure 4a that lexical closure properties at OHSU depend on the note types. Progress Notes are similar in slope to the general-domain British National Corpus (not pictured). In

DOI: 10.1002/asi

Figure 4b, Mayo Clinic data is seen to exhibit strong lexical diversity across multiple sections. Here, it is Specialty Evaluations and Consults that have the most lexical diversity.

Comparing Figure 4a with 4b, we observe several interesting phenomena.

- First, the "All" aggregations have very similar curves at both institutions. Regardless of the subcategorizations, this implies that clinical text at different institutions, even across significantly different EHR data structures, still deals with a similar breadth of knowledge. This curve has (not pictured) a smaller slope than that of the general-domain British National Corpus, but a larger slope than biomedical literature in the GENIA corpus (Cohen, Baumgartner, & Temnikova, 2016).
- Aside from the "All" curves, the content of the text subcategories at the two institutions do not seem to correspond tightly. For example, whereas Progress Notes show the greatest lexical diversity at OHSU, they show some of the least lexical diversity at Mayo Clinic. This can be partially explained by how the subcategories are used: Mayo Clinic "Progress Notes" are inpatient "Patient Progress" notes, whereas OHSU's "Progress Notes" include both inpatient and outpatient data.
- There appears to be more subcategory types at OHSU that have relatively low lexical diversity. Handoffs, After Visit Summary (AVS) Snapshots, and Patient Instructions do not show closure (plateaus) at this sample size, but they do have a smaller amount of lexical diversity compared to the other types and compared to the Mayo Clinic subcategories.
- Also, the number of text samples from subcategories is smaller at OHSU (cf. short lines). There is only one note type (or the "None" type) for each of the 3,491,659 OHSU clinical notes, whereas there are multiple document types at Mayo Clinic and each document has multiple section types.

There may be implications here for how to design EHR systems that minimize a clinician's cognitive load, are sufficiently expressive for broad-based clinical care, and efficiently serve EHR use cases in research and practice. However, we should be careful in drawing these types of conclusions without further investigation. For example, lower lexical diversity in some of OHSU's records might mean: (a) the Epic EHR structure removes ambiguity, allowing clinicians to treat some of their notes as an easy-to-document sublanguage; (b) the Epic EHR structure introduces unnecessary ambiguity, requiring a clinician to regularly add templated text; (c) auto-complete features in Epic are more available or advanced; (d) these OHSU note types are more likely to be cut-and-pasted. It is beyond the scope of this article to examine each of these highly divergent hypotheses.

Simulated Competition Run Analysis

We now turn to analysis of the simulated competition strategy presented in the section on Relevance Assessments. In this setting, we must decide how to produce relevance assessments, rather than how to evaluate systems against existing relevance assessments. Again, the key desiderata were to have diversity and relevance amongst the multiple runs. To this end, we provide descriptive analysis of the runs generated by the parameter settings at OHSU.

Figure 5 shows 48^2 similarity scores between each of pair of runs (ranked patient lists). The similarity scores are asymmetric comparisons between lists l_1 and l_2: we considered the top $k=200$ items in l_1 to be relevant (as in pseudo-relevance feedback), and calculated the mean average precision (MAP) for l_2. Our metric measures IR-meaningful diversity between runs, but cannot necessarily measure relevance. The scores are presented as a heatmap: darker areas (e.g., the diagonal) show greater correlation between runs, and less diversity; lighter areas show greater diversity, and possibly less relevance. The heatmap is laid out with categorical variables alternating values in a fixed pattern, allowing for visual inspection of patterns between the runs.

Some observations from Figure 5:

- First, retrieval models correspond to 4×4 blocks. The fact that most of the heatmap is checkered with cohesive 4×4 blocks indicates that very little diversity is introduced with the 4 different retrieval models. The only exception is that BM25 diverges from the others when only text notes are considered (resulting in some 3×3, 3×4, or 4×3 blocks solid color).
- Zooming out, the aggregation strategy is represented by 8×8 blocks (assembled from four 4×4 blocks, or four quadrants comparing the aggregation strategies). The top-left quadrant (max aggregation) and bottom right quadrant (sum aggregation) always show higher similarities; the top-right and bottom-left quadrants (comparing between aggregations) are less similar. Thus, the different aggregation strategies are unique versions of each run, and add some diversity to the runs. Although this difference disappears if we consider the heatmap for $k=1000$ (not pictured), this means that the rankings between sum and max aggregations are different. Significantly different rankings will change how a sampled pooling strategy selects patients to judge.
- Comparisons between subsets of notes are shown within 16×16 blocks. Again, considering these blocks to have 4 quadrants, the top-left (notes only) and bottom-right (all data) are not easily distinguished from the top-right and bottom-left (mixtures), except along the main diagonal of the heatmap. Thus, there is relatively little diversity introduced by subsetting the data to use only notes. The one exception is when the retrieval model is BM25, as previously mentioned.
- Finally, the whole heatmap can be split into nine sections, corresponding to comparisons between each of the three topic representations. Topic representations largely correlate with other runs with the same topic representation. Thus, these different representations add diversity to the pool.

From the foregoing analysis, we might determine that we could eliminate unnecessary runs and maximize diversity by comparing all three topic representations; two aggregation strategies; and BM25 with notes only vs. LMDirichlet with all data. This would lead to a set of 12 runs with a higher

DOI: 10.1002/asi

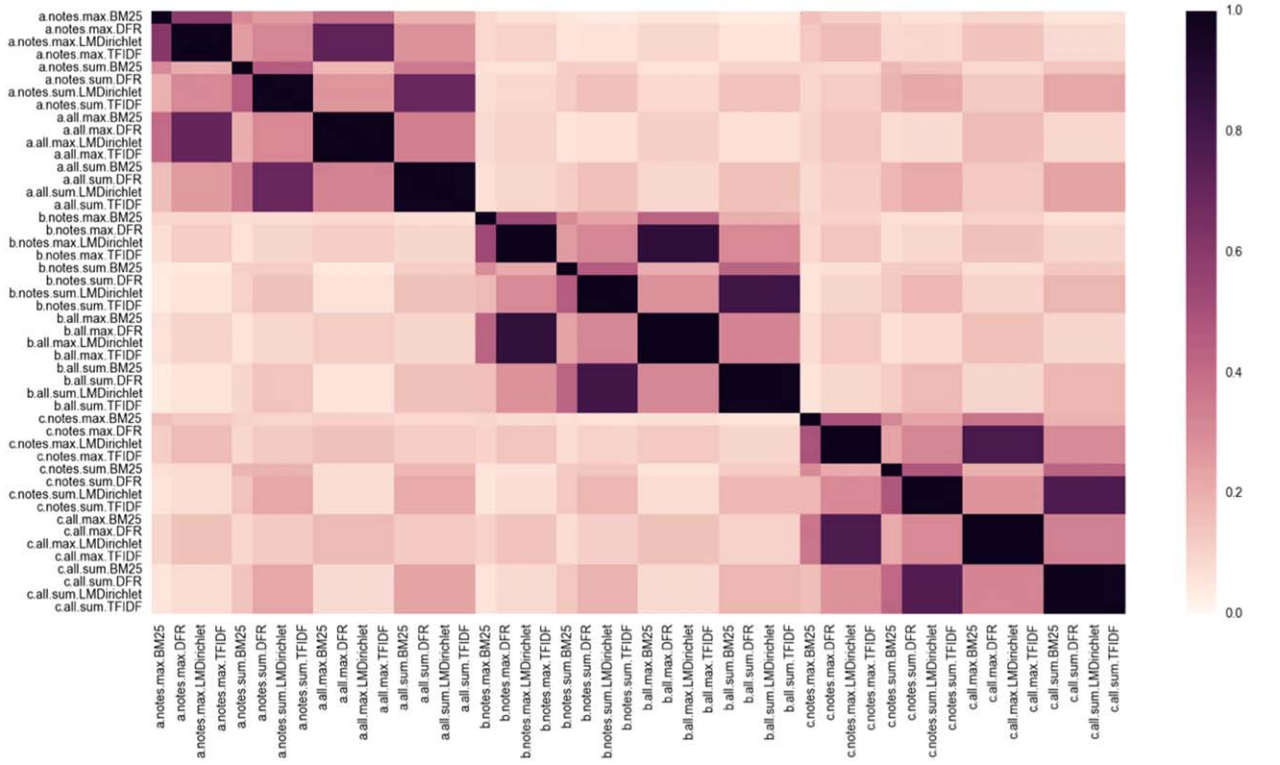

FIG. 5. Similarity (as measured by pseudo-relevance map of top 200 items) for the 48 simulated competition runs at OHSU. Darker areas represent higher similarity. Periodic bands of color show regularities in results. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4. Counts and time (in minutes per patient) for preliminary relevance assessments on five topics.

ID	Description	# in pool	# rel	min/pt
1	non-smoking women in 3rd trimester of pregnancy without a DSM-IV axis 1 diagnosis	1161	71	2.0
2	adults with IBD being managed medically	866	10	4.0
3	adults with a measured vitamin D (25-hydroxycholecalciferol) level	833	722	1.0
4	adults with post-herpetic neuralgia using Qutenza (capsaicin 8% patch)	714	0	2.3
6	pregnant women in 3rd trimester seen in outpatient women's health clinic	767	119	4.2

degree of internal diversity. However, it is unclear whether, in conjunction with a sampling method, such a selection would positively or negatively affect the ability of the simulated competition to return relevant articles with high confidence. Though it is outside the scope of this article to explore strategies for selecting which patients on which to do relevance assessments, active learning strategies based on simulated competitions are an open area for future work.

Preliminary Estimates of Assessment Efficiency

A set of relevance judgments on full patient pools was performed for five topics. The amount of time spent on judging each topic's pool is listed alongside the size of the pool in Table 4.

It should be noted that the average times listed in Table 4 do not scale linearly with the amount of patient information available, when compared with the TREC-Med relevance assessment process. In TREC-Med, assessors were responsible for judging patient visits (such as OHSU's Encounters) and spent about 1 minute per visit. For each patient at OHSU, there was 37.62 times the number of visits (Encounters), yet assessments took from 1–4 times as long. Table 4 is reassuring in that the move from visit-level relevance to patient-level relevance is not infeasible.

However, based on these relevance judgments, we also expect the average time required per patient-level judgment to exhibit significant variation between topics. The main variable that was noted (according to the qualitative judgment of relevance assessors) to affect assessment speed was whether the information required to evaluate topic criteria was present in structured data or as free text.

Conclusion and Future Work

We have considered at length the relatively novel problem of patient-level IR, and discussed some principles for developing resources in this domain. In addition, we have included details on an implementation of these concepts at OHSU and Mayo Clinic: a patient-level test collection, diverse topics, the PRAI web interface for chart review-based relevance judgments, and a simulated competition

DOI: 10.1002/asi

pooling strategy. Finally, we have shown some analysis on the characteristics of the parallel IR collections at OHSU and Mayo Clinic.

Future work includes making it easier to replicate our study. Given the complicated relationship between OHSU and Mayo Clinic data, mapping EHR data to a common data model (e.g., from the Observational Health Data Science Initiative [OHDSI] or the National Patient-Centered Clinical Research Network [PCORnet]) will enable greater interoperability and ease adoption of collection creation tools. Also, building off our experience with using the chart review interface with Mayo Clinic data, we plan to eventually release PRAI as an open source project. Furthermore, we plan to explore and extend an Evaluation-as-a-Service approach to shareability (Hanbury et al., 2015), scaling to a distributed collection across multiple sites and running federated shared tasks in this setting.

Acknowledgments

This work was supported in part by the National Institutes of Health Grant R01LM011934.

References

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W.F., Warner, C., Hwang, J.D., ... Martin, J. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association, 20, 922–930.

Amati, G., & Van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20, 357–389.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. Proceedings of SemEval (pp. 1052–1062).

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. Paper presented at the Proceedings of the 23rd annual international ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece: ACM.

Chute, C.G., Pathak, J., Savova, G.K., Bailey, K.R., Schor, M.I., Hart, L. A., ... Huff, S.M. (2011). The SHARPn project on secondary use of Electronic Medical Record data: Progress, plans, and possibilities. AMIA ... Annual Symposium Proceedings/AMIA Symposium. AMIA Symposium, 2011, (pp. 248–256). Washington, DC: AMIA.

Cleverdon, C.W., & Keen, M. (1966). Aslib Cranfield research project-Factors determining the performance of indexing systems; Volume 2, Test results.

Cohen, K.B., Baumgartner, Jr, W., & Temnikova, I. (2016). SuperCAT: The (New and Improved) Corpus Analysis Toolkit. Paper presented at the The Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two biomedical sublanguages: A description based on the theories of Zellig Harris. Journal of Biomedical Informatics, 35, 222–235. doi:10.1016/s1532-0464(03)00012-1

Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., ... Zuccon, G. (2013). ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. CLEF 2013 Online Working Notes, 8138.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., ... Mueller, H. (2014). Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. Paper presented at the Proceedings of CLEF 2014. Sheffield, UK: Conference and Labs of the Evaluation Forum.

Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., ... Zuccon, G. (2015). Overview of the clef ehealth evaluation lab 2015. Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages. Toulouse, France: Conference and Labs of the Evaluation Forum.

Hanbury, A., Boyer, C., Gschwandtner, M., & Müller, H. (2011). KHRESMOI: Towards a multi-lingual search and access system for biomedical information. Med-e-Tel, Luxembourg, 2011, 412–416.

Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G.V., Eggel, I., ... Kando, N. (2015). Evaluation-as-a-service: Overview and outlook. arXiv preprint arXiv:1512.07454.

Lin, J., & Efron, M. (2013). Evaluation as a service for information retrieval. Paper presented at the ACM SIGIR Forum, New York.

Mowery, D.L., Velupillai, S., South, B.R., Christensen, L., Martinez, D., Kelly, L., ... Savova, G. (2014). Task 2: Share/clef ehealth evaluation lab 2014. Paper presented at the proceedings of CLEF 2014, Sheffield, UK.

Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G., ... Pecina, P. (2015). CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. Paper presented at the proceedings of CLEF, Toulouse, France.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., & Gatford, M. (1995). Okapi at TREC-3. NIST SPECIAL PUBLICATION SP (p. 109).

Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.-W., Moody, G., ... Mark, R.G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical Care Medicine, 39, 952.

Stubbs, A., Kotfila, C., Xu, H., & Uzuner, O. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. Journal of Biomedical Informatics, 58 Suppl, S67–S77. doi:10.1016/j.jbi.2015.07.001

Stubbs, A., & Uzuner, O. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics, 58 Suppl, S20–S29. doi:10.1016/j.jbi.2015.07.020

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association, 20, 806–813. doi:10.1136/amiajnl-2013-001628

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., ... Jones, G.J. (2013). Overview of the ShARe/CLEF eHealth evaluation lab 2013. Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages. Valencia, Spain: Conference and Labs of the Evaluation Forum.

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., & South, B.R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association, 19, 786–791. doi:10.1136/amiajnl-2011-000784

Uzuner, O., South, B.R., Shen, S., & DuVall, S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18, 552–556. doi: amiajnl-2011-000203 [pii] 10.1136/amiajnl-2011-000203

Voorhees, E., & Harman, D.K. & National Institute of Standards and Technology (U.S.). (2005). TREC: Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

Voorhees, E., & Hersh, W. (2012). Overview of the TREC 2012 medical records track. Paper presented at the the Twenty-first Text REtrieval Conference Proceedings TREC, Gaithersburg, MD.

Voorhees, E., & Tong, R. (2011). Overview of the TREC 2011 medical records track. Paper presented at the 20th Text REtrieval Conference Proceedings TREC, Gaithersburg, MD.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval.