On Developing Resources for Patient-level Information Retrieval

Stephen Wu[1], Tamara Timmons[1], Amy Yates[1], Meikun Wang[1], Steven Bedrick[1], William Hersh[1], Hongfang Liu[2]

[1]Oregon Health & Science University
Portland, OR
E-mail: wst@ohsu.edu

[2]Mayo Clinic
Rochester, MN
liu.hongfang@mayo.edu

Abstract

Privacy concerns have often served as an insurmountable barrier for the production of research and resources in clinical information retrieval (IR). We believe that both clinical IR research innovation and legitimate privacy concerns can be served by the creation of intra-institutional, fully protected resources. In this paper, we provide some principles and tools for IR resource-building in the unique problem setting of patient-level IR, following the tradition of the Cranfield paradigm.

Keywords: clinical information retrieval, privacy, information retrieval evaluation

1. Introduction

Privacy concerns have often served as an insurmountable barrier for the production of research and resources in clinical information retrieval (IR). Restrictions on the use of patients' private health information are 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 intra-institutional, fully protected resources. Thus, we have begun creating such resources at two sites: Oregon Health & Science University (OHSU) and Mayo Clinic. In future work, other institutions may choose to adopt these resource development practices; additionally, access could be provided through a privacy-preserving interface, such as with Evaluation-as-a-service (1).

In this paper, we provide some principles and tools for resource-building in the unique problem setting of patient-level IR. We follow the long tradition in IR of test collections and challenge evaluations, specifically structuring our resources for Cranfield-style IR evaluations (2, 3). Cranfield evaluations require (i) a test collection, (ii) a set of test topics, expressed as queries, and (iii) judgments of whether documents are relevant, for each query.

Each of these 3 collection components needs to be re-envisioned for its role in patient-level IR. In the remainder of our paper, we will walk through the 3 components and provide *principles* for their design, highlighting how they differ from a "traditional" test collection. We will also describe the *implementation* of these principles in our own multi-institutional resource-building project. We also introduce the Patient Relevance Assessment Interface (PRAI), a tool for producing relevance judgments for patient-level IR.

2. Related Work

The Pittsburgh NLP Repository, distributed exclusively for the 2011-2012 TREC Medical Records Tracks (4, 5), sought to address this problem 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. While this was an important step forward in that it moved beyond document retrieval, it did not (and could not) fully embrace patient-level IR.

Other research resources and challenges have taken alternative approaches rather than de-identifying protected health information. The CLEF eHealth tasks since 2013 (6) have included patient data (discharge summaries) paired with queries mimicking those a lay person might search for on the web. This problem setting bears more similarities to traditional web search, and differs from ours in that it focuses on consumer health -- the patient, not the provider, is seeking health information. Multimodal and image search in medical records, e.g., the EU-sponsored VISCERAL project (www.visceral.eu), can somewhat sidestep these problems by preprocessing the identified text and giving term lists, essentially providing an immutable upstream NLP process. None of these publicly available corpora are able to support the connectedness and granularity of information that is possible in the intra-institutional corpora that we propose here.

3. Patient-Level Test Collections

3.1 Unit of Retrieval: Patient Electronic Health Records (EHRs)

Patient-level IR is about finding patients, not just documents. Any source of information might be relevant to a patient's medical history or condition – e.g., lab values and clinical notes in the EHR (or more broadly, a wearable fitness device and a patient's social media presence).

In our initial work at OHSU, we limit ourselves to patient-level retrieval based on EHRs. Patient-level structure is provided by linking all documents for each patient by his or her patient ID; documents are clinical text and other structured data generated by medical professionals documenting patient encounters. Thus, there are typically many documents per patient.

The data provided was initially collected from Epic, OHSU's electronic health record, and stored in the corresponding Clarity database. Selected medically relevant tables were then extracted and provided to the research team as XML.

1.1 Shareability: Intra-institutional collections

Data is protected by laws of the Health Insurance Portability and Accountability Act (HIPAA), and therefore cannot be shared outside the secure servers of a healthcare institution. Test collections should therefore be built within institutions, and shareability achieved at a different point in the research life cycle.

We have built preliminary test collections at OHSU and Mayo Clinic, while the tooling and infrastructure has been provided by OHSU. We primarily list statistics and details from the OHSU corpus and tooling.

1.2 Significance: Collections represent patient populations

Since a test collection is a collection of patients, it represents a *patient population*. Thus, in patient-level IR, the choice of which patient records to include is of potential medical and public health significance. Further, boolean retrieval results may be considered *patient cohorts*.

At OHSU, patients were included in the pool if they had inpatient or outpatient encounters with primary care departments (Internal Medicine, Family Medicine, or Pediatrics), with 3 or more encounters and 5 or more text entries, between 1/1/2009 and 12/31/2013. This resulted in a pool of 99,965 unique patients and 6,273,137 unique encounters.

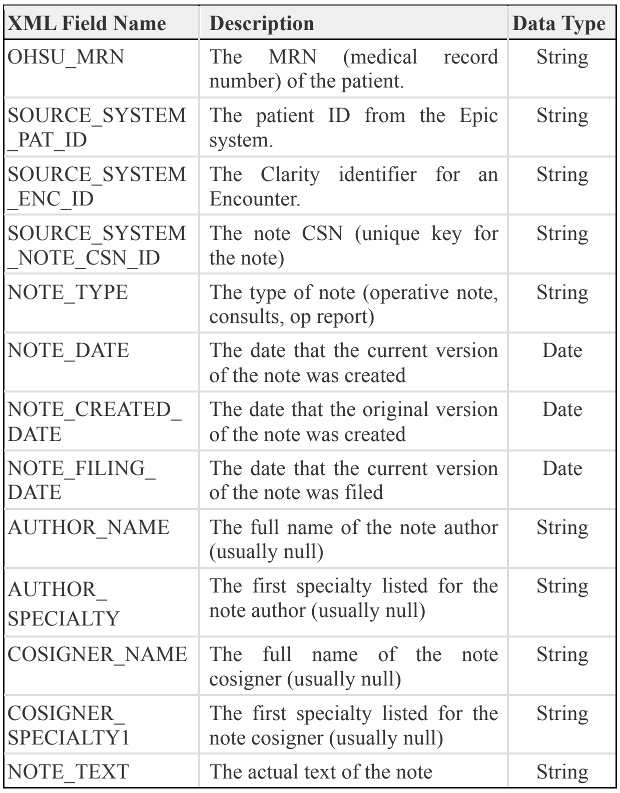
The patient population at Mayo Clinic is still under development, but a preliminary set has 15,486,886 notes corresponding to 138,228 patients, spanning a period of 15 years (1998–2013), and covering both inpatient and outpatient data.

1.3 Distributedness: Relevant evidence is dispersed across diverse document types

As mentioned in Section 2.1, a natural consequence to looking at whole patients is that diverse types of data records must be considered.

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.

Within each of these documents there are multiple fields that contain data such as medical record number, note text, lab results, or diagnosis. For example, Figure 1 shows the document structure for the provider Notes; fields are either in string format or in date format.



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 system.	String
SOURCE_SYSTEM_ENC_ID	The Clarity identifier for an Encounter.	String
SOURCE_SYSTEM_NOTE_CSN_ID	The note CSN (unique key for the note)	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

Figure 1: Clinical Notes document structure

The number of records of each type varied widely. There were 10,111,930 clinical notes (approximately 100 notes per patient), but 31,997,402 current medications (~300 medications per patient), versus 31,889 surgeries (~1 surgery for every 3 people).

2. Patient-level Test Topics

2.1 Sources: Diverse, practical topics sources

We have adopted the task of cohort identification for our patient-level IR evaluation topics. Use cases for such cohorts include research study recruitment, preliminary screening for a later manual review, evidence-based clinical care, and characterization of population health in epidemiological studies; in development of test topics, we aimed to reflect this diversity of use cases and represent real-world information needs.

A total of 56 test topics were developed based on defined patient cohorts drawn from 5 sources, illustrating a variety of use cases. Cohort descriptions from these sources, generally composed of eligibility criteria, serve as models for test topics.

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 since it specified retrieval of clinic notes rather than an individual patient. Additional topics were modeled after cohorts from the Phenotype KnowledgeBase (PheKB) (7 topics), Rochester Epidemiology Project (REP) (9 topics), and National Quality Forum (NQF) (12 topics). Finally, Mayo Clinic provided cohort descriptions from its own RDW to create 2 topics. Cohorts with similar characteristics were merged during topic development to avoid redundancy (1 OHSU/REP topic, and 2 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.

2.2 Format: Diverse topic representations

Including different representations of topics allows for hypothetical queries from different use cases.

We provide the 56 topics in 3 different formats for possible queries: 1) summary statement; 2) brief summary and clinical – a shorter summary statement plus a mock clinical case incorporating a patient and scenario which typify the topic criteria; 3) 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 2:

a. Adults under age 65 with rheumatoid arthritis who have cyclic citrullinated peptide antibodies >40

b. Adults under age 65 with rheumatoid arthritis with positive cyclic citrullinated peptide antibodies.

i. 58 year old female presents with morning stiffness and joint pain in her hands, especially her fingers, which improves but does not remit fully after approximately 30 minutes. 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 Anti-CCP Ab level is 45.

c. Adults under age 65 with rheumatoid arthritis with positive cyclic citrullinated peptide antibodies.

i. Demographics: 18-64 years old, alive, not on genetic opt-out list
ii. Encounter: date within last 2 years
iii. Inclusion: rheumatoid arthritis diagnosis, Anti-CCP antibody level >40 (“CYCLIC CITRUL PEPTIDE AB, IGG”)

Figure 2: 3 representations for Topic 15

3. Patient-level Relevance Judgments

3.1 Judges: Medical expert relevance judges

Relevance judgments are the rate-limiting portion of IR resource development, since they must be done by human assessors. Furthermore, chart reviews require specialized medical knowledge to be meaningful, and patient privacy concerns prevent the sharing of documents from a patient’s EHR.

Thus, in our work, crowd-sourcing relevance judgments was not an option; rather, we have employed intra-institutional medical experts to take on the costly step of chart review to make patient-level relevance judgments.

3.2 Task and tools: Relevance assessment as Chart review

With a test collection of patients, the assessment of relevance to a query is equivalent to the problem of manual medical record (chart) review. Unlike traditional IR assessment (on a single document), many pieces of information (possibly thousands of text and/or structured data documents corresponding to a single patient) must be considered in making a judgment on relevance.

Chart review in medical research is often done with a clinician’s full EHR interface, elaborate spreadsheets, and manual record-keeping. We found that this patient-level IR setting required a new tool for relevance assessment, to which we devote the rest of this paper.

Patient Relevance Assessment Interface

To support this 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 PostgreSQL database for tracking judgments and to Elasticsearch for retrieving patient data.

PRAI Interface and Usage

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.

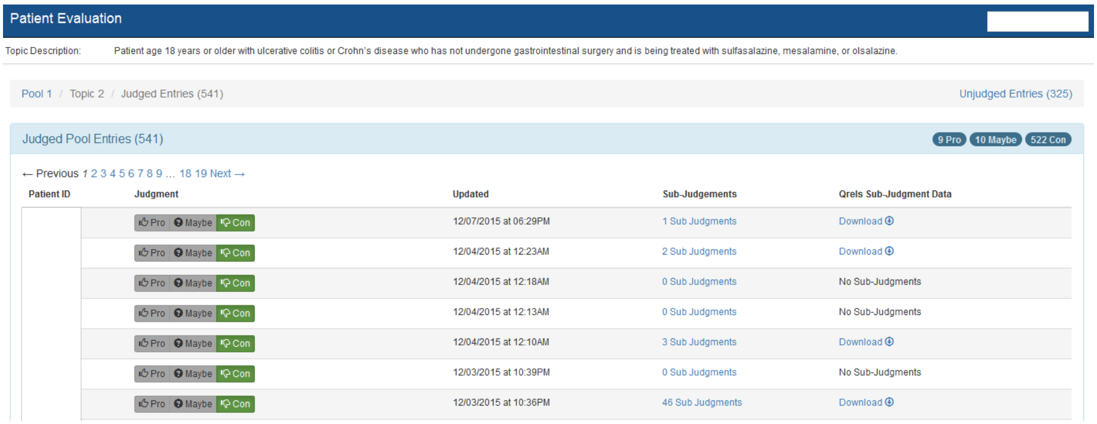
PRAI allows users to record patient-level relevance judgments for a given topic and patient (see Figure 3). It also introduces the ability to perform “Sub Judgments” (document-level judgments, see Figure 4), whereby a single piece of data is marked as providing evidence in the overall judgment for the patient. A given sub judgment may concern criteria for patient inclusion or exclusion, and may support or contradict 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.

Preliminary relevance assessments for five topics

A first round of relevance judgments on full patient pools was performed for five topics.

The first topic, “non-smoking women in 3rd trimester of pregnancy without a DSM-IV axis 1 diagnosis” (the same topic as used for preliminary assessment), had 1161 patients in the judgment pool. Relevance judgments for this topic required an average of 2 minutes per patient. The judgment pool for Topic 2, “adults with IBD being managed medically,” included 866 patients, and about 4



Patient Evaluation

Topic Description: Patient age 18 years or older with ulcerative colitis or Crohn's disease who has not undergone gastrointestinal surgery and is being treated with sulfasalazine, mesalamine, or olsalazine.

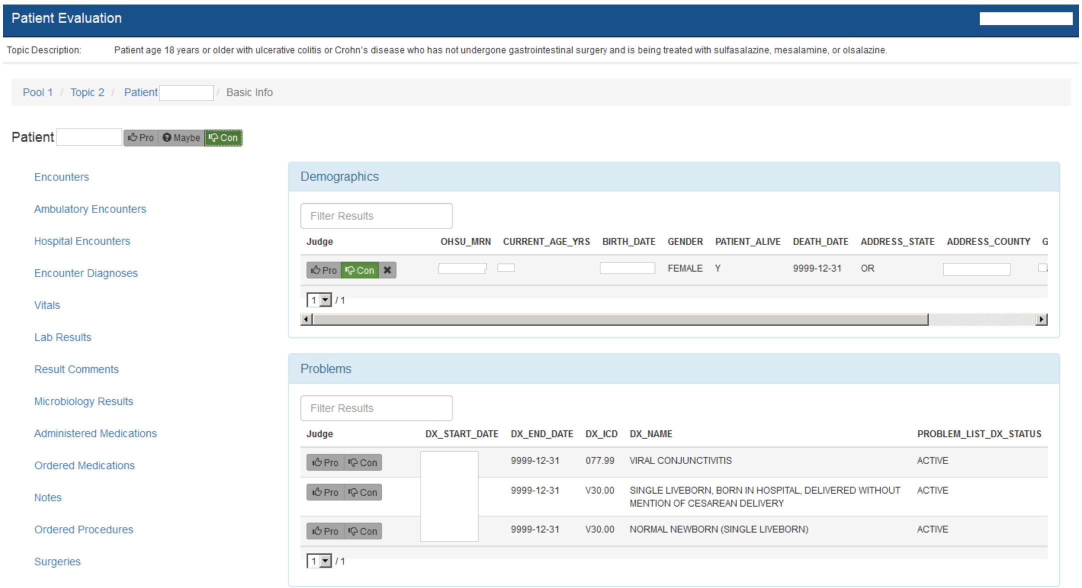
Pool 1 / Topic 2 / Judged Entries (541) Unjudged Entries (325)

Judged Pool Entries (541) 9 Pro 10 Maybe 522 Con

← Previous 1 2 3 4 5 6 7 8 9 … 18 19 Next →

Patient ID	Judgment	Updated	Sub-Judgements	Qrels Sub-Judgment Data
	Pro Maybe Con	12/07/2015 at 06:29PM	1 Sub Judgments	Download
	Pro Maybe Con	12/04/2015 at 12:23AM	2 Sub Judgments	Download
	Pro Maybe Con	12/04/2015 at 12:18AM	0 Sub Judgments	No Sub-Judgments
	Pro Maybe Con	12/04/2015 at 12:13AM	0 Sub Judgments	No Sub-Judgments
	Pro Maybe Con	12/04/2015 at 12:10AM	3 Sub Judgments	Download
	Pro Maybe Con	12/03/2015 at 10:39PM	0 Sub Judgments	No Sub-Judgments
	Pro Maybe Con	12/03/2015 at 10:36PM	46 Sub Judgments	Download

Figure 3: Patient-level judgments (IDs removed)



Patient Evaluation

Topic Description: Patient age 18 years or older with ulcerative colitis or Crohn's disease who has not undergone gastrointestinal surgery and is being treated with sulfasalazine, mesalamine, or olsalazine.

Pool 1 / Topic 2 / Patient / Basic Info

Patient Pro Maybe Con

Demographics

Filter Results

Judge	OHSU_MRN	CURRENT_AGE_YRS	BIRTH_DATE	GENDER	PATIENT_ALIVE	DEATH_DATE	ADDRESS_STATE	ADDRESS_COUNTY	G
Pro Con ✖				FEMALE	Y	9999-12-31	OR		

1 / 1

Problems

Filter Results

Judge	DX_START_DATE	DX_END_DATE	DX_ICD	DX_NAME	PROBLEM_LIST_DX_STATUS
Pro Con		9999-12-31	077.99	VIRAL CONJUNCTIVITIS	ACTIVE
Pro Con		9999-12-31	V30.00	SINGLE LIVEBORN, BORN IN HOSPITAL, DELIVERED WITHOUT MENTION OF CESAREAN DELIVERY	ACTIVE
Pro Con		9999-12-31	V30.00	NORMAL NEWBORN (SINGLE LIVEBORN)	ACTIVE

1 / 1

Figure 4: Document-level "Sub Judgments" showing the Demographics and Problem List sections

minutes per patient-level judgment. Each patient-level judgment for the 3rd topic, "adults with a measured vitamin D (25-hydroxycholecalciferol) level," was completed in just over 1 minute on average, with a pool of 833 patients. Topic 4, "adults with post-herpetic neuralgia using Qutenza (capsaicin 8% patch)," had a pool of 714 patients for judgment, taking the medical expert approximately 2.3 minutes to complete each patient-level judgment. Lastly, 767 patients were included the judgment pool for the topic "pregnant women in 3rd trimester seen in outpatient women's health clinic," and each patient-level judgment was completed in 4.2 minutes on average.

Based on our preliminary relevance judgments, we expected the average time required per patient-level judgment to exhibit significant variation between topics. Relevance assessments of these 5 topics showed over 4-fold variation in the average time per patient-level judgment between topics.

The main variable noted to affect the time required to perform a patient-level judgment was whether the information required to evaluate topic criteria was present in structured data or as free text.

4. Conclusion & 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. Additionally, we have included details on an implementation of these concepts at OHSU: a patient-level test collection, diverse topics, and the PRAI web interface for chart review-based relevance judgments.

Future work includes making it easier to replicate our study; building off of 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 may consider modern approaches to shareability such as Evaluation-as-a-Service (1).

5. Bibliographical References

1. Lin J, Efron M, editors. Evaluation as a service for information retrieval. ACM SIGIR Forum; 2013: ACM.
2. Cleverdon CW, Keen M. Aslib Cranfield research project-Factors determining the performance of indexing systems; Volume 2, Test results. 1966.
3. Voorhees E, Harman DK, National Institute of

Standards and Technology (U.S.). TREC : experiment and evaluation in information retrieval. Cambridge, Mass.: MIT Press; 2005. x, 462 p. p.

4. Voorhees E, Hersh W, editors. Overview of the TREC 2012 medical records track. The Twenty-first Text REtrieval Conference Proceedings TREC; 2012; Gaithersburg, MD: National Institute of Standards and Technology.

5. Voorhees E, Tong R, editors. Overview of the TREC 2011 medical records track. The Twentieth Text REtrieval Conference Proceedings TREC; 2011; Gaithersburg, MD: National Institute of Standards and Technology.

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