Evaluation of Patient-Level Retrieval from Electronic Health Record Data for a Cohort Discovery Task

Steven R. Chamberlin[1*], Steven D. Bedrick[1,2], Aaron M. Cohen[1], Yanshan Wang[3], Andrew Wen[3], Sijia Liu[3], Hongfang Liu[3], William R. Hersh[1]

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR
[2]Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR
[3]Division of Digital Health Sciences, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[*]Correspondence:
Steven R. Chamberlin
chambest@ohsu.edu

Abstract

Objective

Growing numbers of academic medical centers offer patient cohort discovery tools to their researchers, yet the performance of systems for this use case is not well-understood. The objective of this research was to assess patient-level information retrieval (IR) methods using electronic health records (EHR), and to investigate the interplay between commonly used IR approaches and the cohort definition structure.

Materials and Methods

Using the Cranfield IR evaluation methodology, we developed a test collection based on 56 test topics characterizing patient cohort requests for various clinical studies. Test collection data was derived from patient records originating from OHSU's EHR data warehouse. Automated IR tasks were performed, varying four different parameters for a total of 48 permutations, with performance measured using B-Pref. We subsequently created 56 structured Boolean queries for the 56 topics for performance comparisons. Finally, we designed 59 taxonomy characteristics to classify the structure of the 56 topics. Six topic complexity measures were derived from these characteristics for further evaluation using a beta regression simulation.

Results

The best-performing word-based automated query parameter settings achieved a mean B-Pref of 0.167 across all 56 topics. The way a topic was structured (topic representation) had the largest impact on performance. Performance not only varied widely across topics, but there was also a large variance in sensitivity to parameter settings across the topics. Structured queries generally performed better than automated queries on measures of recall and precision, but were still not

able to recall all relevant patients found by the automated queries. We also found strong performance associations with the six complexity measures created from the topic taxonomy, and interactions with automated query parameter settings.

Conclusion

While word-based automated methods of cohort retrieval offer an attractive solution to the labor-intensive nature of this task currently used at many medical centers, we generally found suboptimal performance in the methods tested for this study. Some of the characteristics derived from a query taxonomy could lead to improved selection of approaches based on the structure of the topic of interest. Insights gained here will help guide future work to develop new methods for patient-level cohort discovery with EHR data.

Key Words

Information retrieval
Patient cohort discovery
Electronic health record
Structured queries
Topic taxonomy

1. Introduction

Many academic medical centers offer patient cohort discovery to their researchers to facilitate clinical research, usually including electronic health record (EHR) data (1). However, the performance of systems and algorithms for this EHR use case is not well-studied. It has been shown that typical review of patients for study eligibility is a labor-intensive task, and that automated preprocessing of lists of patients may reduce human time and effort for selection of cohorts (2-4).

One challenge for evaluating this use case is the lack of test collections that include data, clinical study descriptions, and relevance judgments for retrieved patients, a problem that has hindered many types of research using EHR data, even in the modern era of ubiquitous EHR adoption (5). A major barrier has been the challenge of protecting privacy of the patients from whom the records are from and institutional hesitancy to making such data widely available for informatics research, even in de-identified form (6). This is especially so for use cases involving processing of textual data within records, especially those used on the scale of information retrieval experiments where corpora of thousands to millions of patient records are typically desired.

There are two EHR record collections that have been publicly available, one from the University of Pittsburgh Medical Center (UPMC) (7) and the other the Medical Information Mart for Intensive Care-III (MIMIC-III) from the Massachusetts Institute of Technology (8). Among the uses of the UPMC corpus has been a cohort retrieval for clinical research studies task in a

challenge evaluation as part of the annual Text Retrieval Conference (TREC). The TREC Medical Records Track ran in 2011 and 2012, attracting 29 and 24 academic and industry research groups respectively (9, 10). Using the University of Pittsburgh collection containing 17,264 encounters containing 93,551 documents (some of which included ICD-9 diagnosis codes, laboratory results, and other structured data), a total of 34 and 47 topics respectively by year were developed and relevance judgments performed based on pooled results from participating research groups using the "Cranfield paradigm" common to information retrieval (IR) evaluation research (11). The judgments were performed by physicians enrolled in biomedical informatics educational programs.

Methods found to lead to improved retrieval performance included several domain-specific enhancements on top of word-based queries, including vocabulary normalization specific to the clinical domain, synonym-based query expansion from medical controlled terminology systems such as the Unified Medical Language System (UMLS) Metathesaurus, and recognition of negation (12). Follow-on research with the test collection found continued improvement in performance from approaches such as query expansion for additional clinical and other corpora (13) as well as use of learning-to-rank methods (14).

One limitation of the TREC Medical Records Track was a limitation of the UPMC corpus, which was retrieval at the encounter (e.g., hospital or emergency department visit) and not the patient level. This was due to the de-identification process that broke the links across encounters, a process that also obscured various protected health elements, such as dates, geographic locations, and provider identifiers. Encounter-level retrieval data sets prohibit applying expert judgement and therefore evaluation at the patient level, which is the goal of cohort retrieval. Nonetheless, the TREC Medical Records Track did provide a data set for information retrieval and biomedical informatics researchers to compare different approaches to identifying patient cohorts for recruitment into clinical studies. Unfortunately, the UPMC corpus has been withdrawn from public use (Wendy Chapman, personal communication).

Outside of the TREC Medical Records Track, few other evaluations of cohort retrieval have been carried out and published. Some are limited by being document- or encounter-based, or focus on broadly defined cohorts that may be too general for the clinical research recruitment use case. One analysis using the MIMIC-III corpus looked at two straightforward clinical situations and found accurate retrieval with both structured data extraction and the use of natural language processing (NLP) (15). Another recent approach employed word embeddings and query expansion to define patient cohorts, although used only structured EHR data (16). The 2018 National NLP Clinical Challenges (n2c2) had a shared task devoted to cohort selection for clinical trials but focused on the complementary task of finding attributes of clinical trials as opposed to patient retrieval (17) (replace with overview paper when published).

Another thread of work has focused on making querying easier to carry out, typically through development of natural language or other structured interfaces to the patient data (18-21). Other approaches focus on normalizing semantic representation of patient data within the EHR itself (22) and applying deep learning to non-topical characteristics of studies and researchers (23). A related area to cohort discovery is patient phenotyping, one of the goals of which is to identify patients for clinical studies (24-26). However, the cohort discovery use case has some

differences, as some studies have criteria beyond phenotypic attributes, such as age, past treatments, diagnostic criteria, and temporal considerations.

In 2014, Oregon Health & Science University (OHSU) and Mayo Clinic launched a project to use raw (i.e., not de-identified) EHR data to perform research in parallel (i.e., able to share methods and systems but not data). The OHSU data set has been previously described (27), and this paper reports the first results using this data set along with evaluation at the patient level. The Mayo Clinic has reported some of its work, although its retrieval output and relevance judgments were at the encounter level and not the patient level (28).

2. Materials and Methods

The initial overall goal of this work was to assess and compare different approaches to patient-level retrieval by developing a "gold standard" test collection consisting of the three usual components of a Cranfield-style IR collection (11): records – in this case patient-level medical records, topics – representations of cohorts to be recruited for clinical studies, and relevance judgements – expert determination of which records were relevant to which topics. Our initial plan was to develop the test collection and apply the methods found to work effectively by research groups in the TREC Medical Records Track. However, upon finding the results for numerous topics applied to this data performed sub-optimally, we also focused on additional methods, namely developing a taxonomy of attributes of the topics and the use of structured Boolean queries with additional relevance judgments on a subset of topics.

2.1 Record Collection

As noted in our earlier paper, the patient records originated from OHSU's Epic (Verona, WI) EHR and were transformed and loaded (without any modification of the underlying structured and textual data) to a research data warehouse (27). The study protocol to use the records was approved by the OHSU Institutional Review Board (IRB00011159). To be included in the corpus, patients had to have at least three primary care encounters between January 1, 2009 and December 31, 2013, inpatient or outpatient, with at least five text note entries. This was done to ensure that records would more likely be comprehensive of their care as opposed to a patient referred to the academic medical for a single consultation.

Both structured and unstructured data were included in the collection. Document types included demographics, vitals, medications (administered, current, ordered), hospital and ambulatory encounters with associated attributes and diagnoses, clinical notes, problem lists, laboratory and microbiology results, surgery and procedure orders, and result comments. A unique medical record number was used to link the different document types, and each document type could contain multiple data fields. The collection contained a total of 99,965 unique patients and 6,273,137 associated unique encounters. It originated in a relational database but was extracted into XML format for loading into the open-source IR platform Elasticsearch (v1.7.6) for our experiments.

2.2 Topics

The 56 topics used for this research were developed from five sources by OHSU and Mayo Clinic as described in our previous paper (27). From OHSU, 29 topics were selected from research study data requests submitted by clinical researchers to the Oregon Clinical and Translational Research Institute (OCTRI). From Mayo Clinic, topics were modeled after two patient cohorts found in the Mayo Research Data Warehouse, five patient cohorts in the Phenotype KnowledgeBase (PheKB), nine patient cohorts in the Rochester Epidemiology Project (REP), and 12 patient cohorts based on presence of quality measures from the National Quality Forum (NQF).

Each topic was expressed at three levels of detail, with the complete list in Supplementary Appendix 1:

A. Summary statement – 1-3 sentences
B. Illustrative clinical case
C. Brief summary plus structured inclusion and exclusion criteria for demographics, diagnoses, medications and other attributes

2.3 Initial Runs

As is typically done in Cranfield-style IR experiments, we performed a number of different runs consisting of the text of the topic representation submitted to the ElasticSearch system, which generated ranked output that we limited to 1000 patients per topic. We varied different parameters for different runs by topic representation, text subset, aggregation method, and retrieval model. For the latter, we used a number of common ranking approaches implemented in ElasticSearch and known to be successful both in the TREC Medical Records Track and IR systems generally:

- BM25, also known as Okapi (29)
- Divergence from randomness (DFR) (30)
- Language modeling with Dirichlet smoothing (LMDir) (31)
- Default Lucene scoring, based on the term frequency-inverse document frequency (TF*IDF) model (32)

We performed 48 runs representing all permutations of the following query parameters as described in our previous paper. These representations formed the basis for all queries created for this paper, both manual and automated and include (further referencing in this paper by underlined text):

1. Topic Representation – A (summary statement), B (clinical case), or C (detailed criteria)
2. (Text Subset – only clinical notes or all document types (including structured data reporting as text)
3. Aggregation Method – patient relevance score calculated by summation (sum) of all documents or by maximum (max) value
4. Retrieval Model – BM25, DFR, LMDir, or Lucene

2.4 Relevance Assessment

The relevance assessments were carried out based on the principles discussed in our previous paper (27). The initial pools for relevance judging were generated in a similar manner to TREC

challenge evaluations, where results from different runs (described in Section 2.3) were pooled by selecting from all runs for a given topic the top 15 ranked patients and then randomly selecting 25% of the next 85 (21 patients) and 1% of the next 900 (9 patients). The process of relevance judging used the locally developed Patient Relevance Assessment Interface (PRAI) (27). This system tracked the judgements in a PostgreSQL database and interfaced with the EHR data that was loaded into Elasticsearch. Patient pools for topics were selected for judging and loaded into PRAI, where all document types could be searched by medical experts to determine patient-level relevance for the topic. Document-level sub-relevance could also be assigned in the system. Patients could be assigned one of three levels of relevance: definitely relevant, possibly relevant, or not relevant. For retrieval performance metrics, both definitely and possibly relevant patients were considered relevant, since the use case motivating aimed to identify patients who were likely to be candidates for inclusion in clinical studies, and the number of definite plus possibly relevant patients was typically not vastly larger than would be desired for a clinical trial.

2.5 Assessment of Initial Retrieval Results

We used the trec_eval program to generate retrieval results for the 48 runs. Because our queries did not exhaustively assess all possible approaches to retrieval, we opted to use the B-Pref measure for results, based on its common usage for IR evaluation when relevance judging is considered to be incomplete (33). B-Pref is a measure of how many relevant patients were retrieved, in the ranked lists, ahead of the non-relevant patients. This metric is on the interval [0,1]. The distribution of B-Pref was evaluated across all 56 topics for each of the 48 runs separately. The intent of this analysis was to assess differences in performance between run parameter settings, and variance within each setting across the 56 topics. We also evaluated the distribution across all of the 48 runs for each of the 56 topics separately to assess differences across topics and variance within each topic across the 48 runs.

We also evaluated the retrieval overlap of the 48 runs for all 56 topics combined by calculating the number of patients retrieved by one run (unique combination of the four retrieval parameters) who were also found in a different run, calculated as a percentage. This assessed the ability of the different parameters to retrieve non-overlapping populations. As each run returned a ranked list of the top 1000 patients, and this was repeated for all 56 topics, returning up to 56,000 patients. If a patient was retrieved for more than one topic for a specific run, they were only counted once in the combined list. To assess patterns of overlap we created a 48x48 heatmap. Rows represent the base run, or denominator, and each column is the percent of the row run that is found in the column run.

As noted in the Results section, the results from these runs were substantially lower than comparable methods applied in the TREC Medical Records Track. This led us to perform additional methods described in the rest of this section that included:

1. Development of a topic taxonomy to assess whether characteristics of topics may be associated with variable retrieval performance
2. Use of structured Boolean queries on a subset of topics

2.6 Topic Taxonomy

To explain and predict the performance of the word-based queries, we created a topic taxonomy composed of 59 features. Three of the authors, who were trained clinically (SC, AC, WH), iteratively developed a list of features that covered inclusion or exclusion of medical diagnoses and classifications, medications, procedures, lab tests, clinician information, patient demographics, information about the clinical setting, temporal measures and other aspects. Each of the 56 topics were then classified by these 59 features by the same three individuals. Fleiss Kappa was used to test interrater reliability (34).

We wanted to examine the association between the query performance, as measured by B-Pref, and the 59 taxonomy characteristic classifications of the 56 topics. To do this we did an exploratory data analysis by comparing one heatmap, clustered by query performance, to a second heatmap, clustered by characteristic assignment, with the topic clustering maintained from the first heatmap. The first heatmap used B-Pref as the statistic and clustered run parameter sets by topic. The second heatmap used the level of rater agreement (0-3) as the statistic and clustered by the taxonomy characteristic while maintaining the topic clusters found in the first performance-based heatmap. These heatmaps were compared for pattern similarities between performance clustering and taxonomy clustering.

We next created six binary features by grouping some of the 59 taxonomy characteristics into categories. At least one reviewer had to identify the taxonomy characteristics used to define these features as relevant to the topic. The purpose of these features was to classify query structural complexity for the 56 topics, and to separate information about complexity from information about content. The determination of these components was based on our experience and knowledge of how these topic characteristics impact the complexity of query design. The overall goal was to investigate the relationship between these six taxonomy features and query performance, and to test for any performance related interactions between these features and the four word-based query parameters (topic representation, text subset, aggregation method, and retrieval model). As mentioned above, the permutations of these four parameters created the 48 runs. These interactions capture the relationship between interventions (word-based query parameters) and inherent topic structure related to complexity (binary taxonomy features).

We used a beta regression model for this investigation. This model included the four word-based query parameters (described in Section 2.3), the six binary taxonomy features and all first-order interactions between the parameters and the features. The response variable was B-Pref. Due to data limitations we felt that model coefficients and tests of significant might not be generalizable. We instead chose to use this model to predict B-Pref on all possible permutations of values of the parameters and features, and to investigate the patterns of the predicted B-Pref in this parameter/feature space. Since this simulated data contained all possible combinations of values of the four word-based parameters and the six binary taxonomy features, there were a total of 3,072 entries. Using the simulated data, we estimated the effect of the six binary taxonomy features and the effect of the Topic Representations. We also used this simulated data for an exploratory data analysis, using a heatmap, to assess more complex interactions between the parameter space (interventions) and the binary feature space (inherent topic structure).

A beta regression mean model was selected because the response variable, B-Pref, is continuous, restricted to the unit interval [0,1], and asymmetrically distributed. The logit link function was

used for these analyses. The regression was done with R (v3.3.1) using the package betareg (v3.1-2).

2.7 Structured Queries

Because the word-based query methods that had worked well for the TREC Medical Records Track performed less successfully with this data, we constructed structured Boolean queries for the 56 topics in an iterative manner by one of the authors with clinical experience (SRC). These queries were based on Topic Representation C, which contained structured data and some free text. These queries were allowed to search all document types that we loaded into Elasticsearch. Since these were structured queries, we did not rank patients returned, and all patients returned were kept in the final query results. Patients retrieved could have been part of the word-based retrieval pools and thus known to be relevant or not, or have not been judged. In addition, some patients not retrieved by the structured queries could have been relevant from retrieval and judgments in the word-based pools. Standard to their definitions, recall for each query was measured as patients retrieved and relevant / patients known to be relevant and precision was measured as patients retrieved and relevant / patients retrieved. We also measured patients retrieved who had not been judged for relevance in the initial pool.

2.8 Additional Relevance Assessment for Ten Selected Topics

As we discovered that a number of patients retrieved by the structured queries had not been retrieved by the word-based queries and therefore not judged, we selected ten topics for additional relevance judging of patients returned by the structured queries. These included topics 2, 7, 9, 17, 32, 33, 42, 44, 48, and 52. To build on previous work done in our group, we used five topics that had been selected randomly for this previous research (35), while the second five topics were selected for diversity in taxonomy characteristics and also to represent all five of our sources for topic definitions (OHSU, Mayo, PheKP, REP, NQF). The second five were also selected based on a higher likelihood to be seen in clinical practice (based on clinician judgement), as compared to other topics in the list of 56.

For these ten topics our intent was to judge the entire list of patients retrieved by the structured queries. To compare the structured queries to the word-based queries we used simple precision and recall. B-Pref was not an appropriate measure since the judged structured query patient pools were not ranked. For recall, we combined the relevant patients found in both the structured judged pools and the word-based judged pools. We counted patients judged as definitely or possibly relevant as relevant for all analyses. We also measured relevant patients retrieved in the word-based queries but not in the structured queries.

3. Results

Results are discussed in this section in three parts: word-based query results, topic taxonomy analysis, and structured query results.

3.1 Word-based Query Results

Per the usual Cranfield approach, we performed standard batch runs for the 48 permutations of topic representation, subset, aggregation method, and retrieval model. For relevance judging, the results were pooled by topic. Relevance assessing of patients was done by a physician who took around 30-40 hours per topic. Table 1 shows the number, source, summary, and distribution of relevance judgments for each topic. One topic had no definite or possibly relevant patients and was excluded from further analysis (25). We used the standard trec_eval program to include each topic for each run, along with the relevance judgments, to generate retrieval results for each run.

Num	Source	Summary	Pool	Def Rel	%	Poss Rel	%	Not Rel	%
1	OHSU	Pregnant women w/o psychiatric disorder	721	92	12.8%	9	1.2%	620	86.0%
2	OHSU	Adults with IBD who haven't had GI surgery	684	63	9.2%	4	0.6%	617	90.2%
3	OHSU	Adults with a Vitamin D lab result	774	373	48.2%	0	0.0%	401	51.8%
4	OHSU	Postherpetic neuralgia treated with topical and systemic medication	685	11	1.6%	3	0.4%	671	98.0%
5	OHSU	Children seen in ED with oral pain	726	2	0.3%	0	0.0%	724	99.7%
6	OHSU	3rd trimester prenatal visit with midwife or Ob/Gyn	743	173	23.3%	0	0.0%	570	76.7%
7	OHSU	Hereditary hemorrhagic telangiectasia	695	15	2.2%	0	0.0%	680	97.8%
8	OHSU	Breast cancer and high risk of BRCA mutation	624	100	16.0%	9	1.4%	515	82.5%
9	OHSU	Children with focal epilepsy with partial seizures	687	31	4.5%	13	1.9%	643	93.6%
10	OHSU	Non-smokers with CAD and no DM	682	26	3.8%	2	0.3%	654	95.9%
11	OHSU	Pregnancy with preterm delivery	708	47	6.6%	0	0.0%	661	93.4%
12	OHSU	Children with autism	641	61	9.5%	0	0.0%	580	90.5%
13	OHSU	Renal impairment and daptomycin	673	27	4.0%	4	0.6%	642	95.4%
14	OHSU	Adults with cardiac arrest and CPR who died in ICU	647	27	4.2%	0	0.0%	620	95.8%
15	OHSU	Rheumatoid arthritis and positive anti-CCP	698	51	7.3%	10	1.4%	637	91.3%
16	OHSU	Gestational anemia and postpartum hemorrhage	667	61	9.1%	0	0.0%	606	90.9%
17	OHSU	RA on MTX w/o biologic DMARD	704	20	2.8%	0	0.0%	684	97.2%
18	OHSU	RA on conventional DMARD w/o hepatitis	749	9	1.2%	0	0.0%	740	98.8%
19	OHSU	Children taking an understudied drug	780	17	2.2%	0	0.0%	763	97.8%
20	OHSU	Osteoarthritis w/o rheumatoid or psoriatic arthritis	690	67	9.7%	0	0.0%	623	90.3%
21	OHSU	Premature infants with ALT or AST lab	772	103	13.3%	0	0.0%	669	86.7%
22	OHSU	Pediatric stroke with endovascular procedure	681	0	0.0%	0	0.0%	681	100.0%
23	OHSU	Adults with quadriplegia	721	47	6.5%	0	0.0%	674	93.5%
24	OHSU	Children w/anemia and height and weight measurements.	755	60	7.9%	0	0.0%	695	92.1%
25	OHSU	LASIK w/acuity and corneal sensitivity measured pre- and post-op	711	0	0.0%	0	0.0%	711	100.0%
26	OHSU	Adults with lab result for anti-tTG Ab or antigliadin Ab	687	198	28.8%	0	0.0%	489	71.2%

27	OHSU	Young adults with high A1c	731	34	4.7%	0	0.0%	697	95.3%
28	OHSU	Pregnancy complication with lab results and no HIV or hepatitis	755	90	11.9%	0	0.0%	665	88.1%
29	OHSU	Adults with thyroid surgery or ablation w/o CVD or ischemic heart disease	696	21	3.0%	0	0.0%	675	97.0%
30	PheKB	Possible acute drug-induced liver injury	653	3	0.5%	4	0.6%	646	98.9%
31	PheKB	Peripheral arterial disease (PAD)	700	87	12.4%	0	0.0%	613	87.6%
32	PheKB	ACE inhibitor-induced cough	700	40	5.7%	0	0.0%	660	94.3%
33	PheKB	Children with ADHD on CNS stimulant	732	112	15.3%	0	0.0%	620	84.7%
34	PheKB	WBC differential and no h/o splenectomy, dialysis, or HIV	685	231	33.7%	0	0.0%	454	66.3%
35	NQF	Breast cancer screening mammogram	713	240	33.7%	33	4.6%	440	61.7%
36	NQF	Children with dental decay	734	78	10.6%	0	0.0%	656	89.4%
37	NQF	Inpatient falls with injury	735	10	1.4%	0	0.0%	725	98.6%
38	NQF	Adolescent immunization w/ meningococcal and Tdap/Td	762	90	11.8%	0	0.0%	672	88.2%
39	NQF	ED admission for appendix perforation or abscess	757	12	1.6%	0	0.0%	745	98.4%
40	NQF	Death from acute MI while inpatient	692	5	0.7%	4	0.6%	683	98.7%
41	NQF	Prostate cancer and external beam radiation tx w/adjuvant GnRH agonist/antagonist	668	18	2.7%	0	0.0%	650	97.3%
42	NQF	Elderly patients with dementia on antipsychotic medication	731	24	3.3%	0	0.0%	707	96.7%
43	NQF	Chronic steroid therapy and osteoporosis prevention	715	8	1.1%	1	0.1%	706	98.7%
44	NQF	COPD with potentially avoidable complication	680	38	5.6%	0	0.0%	642	94.4%
45	NQF	Patients w/coronary stent and 12 mo. antiplatelet tx	650	159	24.5%	43	6.6%	448	68.9%
46	NQF	Children with sickle cell anemia and transcranial doppler U/S	689	9	1.3%	0	0.0%	680	98.7%
47	REP	Prostate cancer on biopsy with Gleason score	602	184	30.6%	0	0.0%	418	69.4%
48	REP	Stroke after first MI	698	5	0.7%	0	0.0%	693	99.3%
49	REP	Nephrolithiasis prophylaxis with thiazide diuretic	518	5	1.0%	0	0.0%	513	99.0%
50	REP	Bicuspid aortic valve on echocardiography	450	163	36.2%	26	5.8%	261	58.0%
51	REP	Adults with HCV on lab testing	769	77	10.0%	0	0.0%	692	90.0%
52	REP	Cataract surgery and prior SSRI use	737	23	3.1%	13	1.8%	701	95.1%
53	REP	Vitamin D-deficiency rickets	776	3	0.4%	0	0.0%	773	99.6%
54	REP	Colonic diverticular disease w/o IBD or colon cancer	681	112	16.4%	22	3.2%	547	80.3%
55	Mayo	Functional status in knee-related PT	706	38	5.4%	13	1.8%	655	92.8%
56	Mayo	Fall risk screening in elderly patients	727	62	8.5%	54	7.4%	611	84.0%

Table 1. The 56 topics with number, source, summary, and pool size, as described in the text. Also shown are number and percentage for definitely relevant, possibly relevant, and not relevant from the initial relevance assessment process.

The highest overall performing run was b.notes.max.LMDir, with a mean B-Pref of 0.167. Very close to this run were two variations of the Retrieval Model: b.notes.max.DFR, and b.notes.max.Lucene, although b.notes.max.BM25 scored lower. At the other end of performance, the a.notes.sum.BM25 run had a mean B-Pref of 0.106. Figure 1 depicts the median and distribution of B-Pref for all 48 runs across all 56 topics (Figure 1).

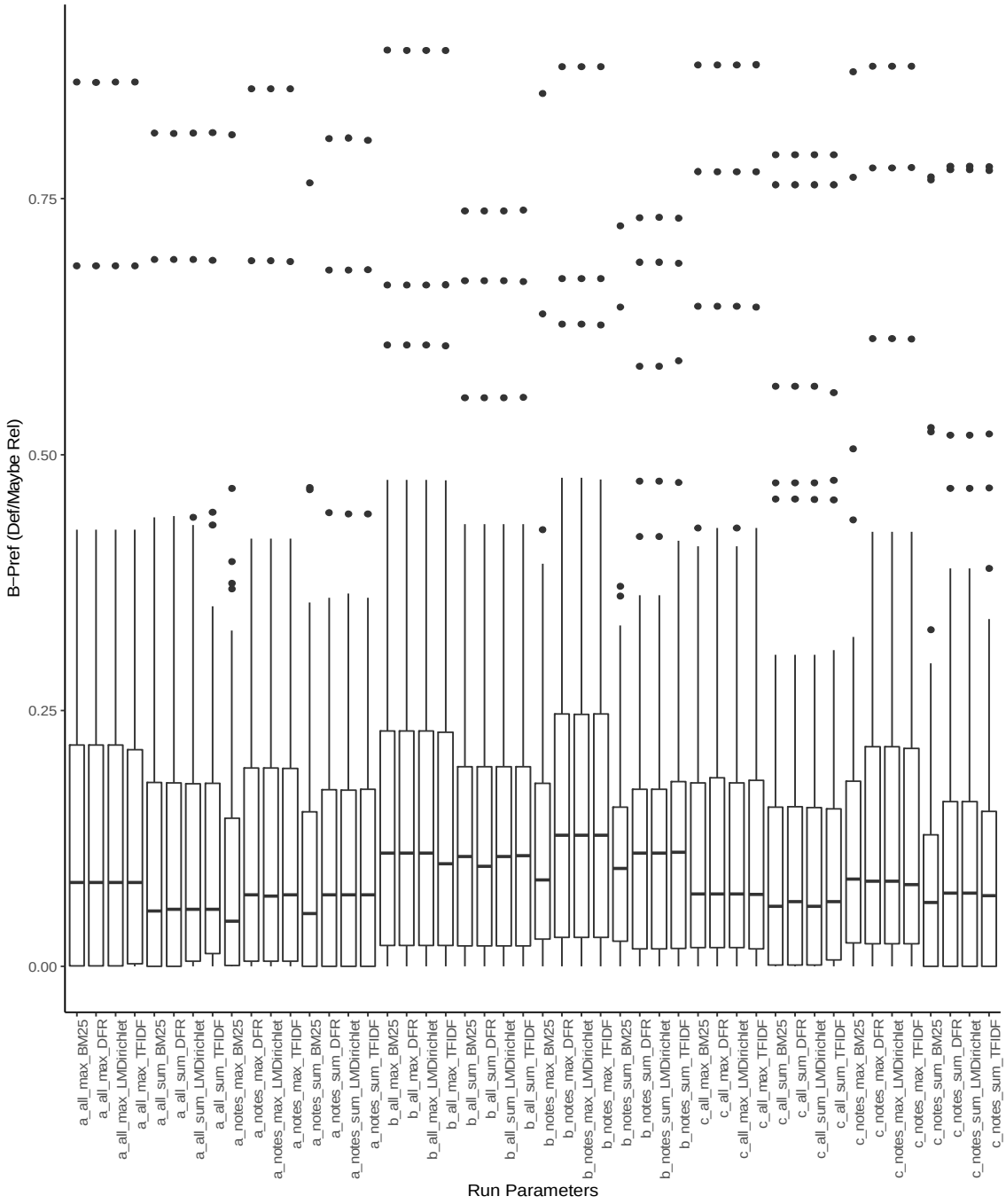

Figure 1. B-Pref distributions for topics within each run. Box ends represent the upper and lower quartile values and whiskers extend 1.5 times the interquartile range. Data points beyond the end of the whiskers are values for individual topics outside the whiskers. The parameter settings are ordered hierarchically first by Topic Representation (A, B, C), then Text Subset (all, notes), then Aggregation Method (max, sum) and finally the Retrieval Model (BM25, DFR, LMDirichlet, TFIDF).

There were several performance grouping patterns seen within the four parameters. Overall, Topic Representation B performed better than the other two representations. This representation was only comprised of text, but includes a detailed individual case description, along with summary description. There was a tendency for the Retrieval Model BM25 to perform poorer than the other models, primarily with the Text Subset notes, which was comprised of a more limited use of available document types. Text Subset did not seem to affect the performance of the other models to the same extent. In fact, there was little difference in performance across the other three models in any subset of parameter settings. Finally, there was a trend for the Aggregation Method sum to have lower performance than the method max.

As is commonly seen in IR experiments, the distribution of topics was spread widely. The highest mean B-Pref was for topic 50 (0.895), while 11 topics had essentially a mean B-Pref of 0.0 (i.e., most runs retrieved no relevant patients) (Figure 2). Two topics consistently had the top two highest values for B-Pref for all parameter combinations within Topic Representation A and C, topics 50 and 28. For Topic Representation B, topic 50 was also consistently in the top two extreme B-Pref values along with topic 47. These topics did not have complex temporal conditions, medication requirements or surgery inclusions or exclusions, and only required relatively straightforward inclusion/exclusion lists of medical conditions, lab and radiology tests, and demographics.

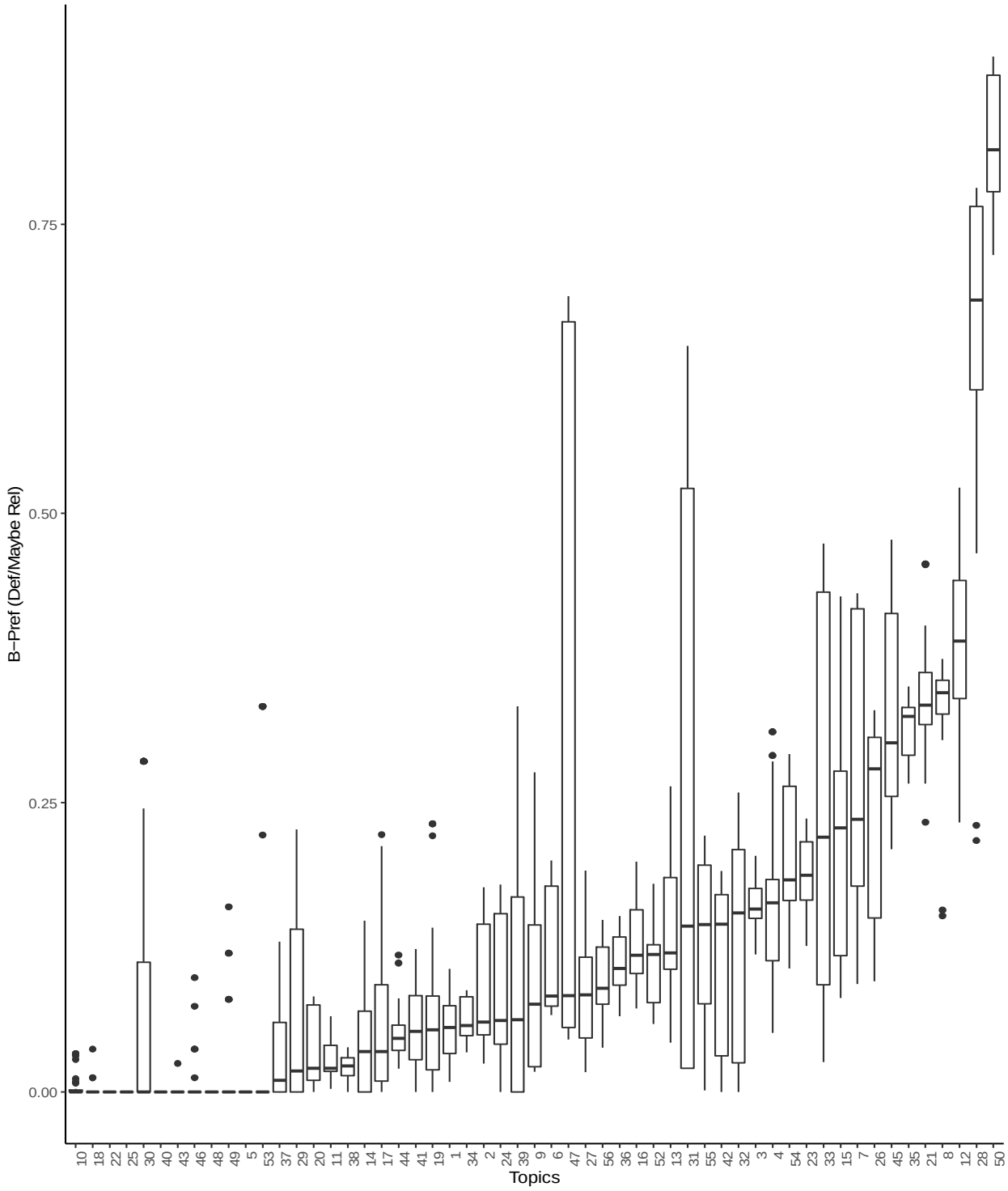

Figure 2. B-Pref distributions for parameter combinations within each topic. Box ends represent the upper and lower quartile values and whiskers extend 1.5 times the interquartile range. Data points beyond the end of the whiskers are values for parameter combinations outside the whiskers. Boxplots are ordered by median B-Pref values.

B-Pref distributions of the 48 parameter combinations (runs) within each topic varied widely in range and shape (Figure 3). Topics 31 and 47 were distinctive, showing much greater variation in performance across parameter settings than the other topics. This variation was entirely due to large differences between Topic Representations. There was very little performance variation for these two topics across the other parameter combinations within each representation. Topic 31 performed much better with the Topic Representation C, which used a combination of structured and unstructured data for the query. The average B-Pref for topic 31 for this representation was 0.60, and for representations A and B was 0.08. Topic 47 performed much better with the B representation, which is a text-only query with a clinical case study added to a summary statement. The average B-Pref for topic 47 for this representation was 0.67, and for representations A and C was 0.07. Other topics did not show as large a performance difference across parameter combinations as topics 31 and 47, but there is still a varying sensitivity. Topics 31 and 47 also demonstrate that query structure could guide optimal parameter combination selection.

We also measured the overlap of patients retrieved across the different runs. Figure 3 shows a pairwise heat map of the runs, with darker color representing a higher overlap. The least overlap occurred between different Topic Representations, particularly for patients found in representation B that were also in C. Within the same Topic Representation for Text Subset all, identical populations were retrieved by all settings of the Aggregation Method and Retrieval Model. However, within the same Topic Representation, there was increased diversity (less overlap) between the Text Subset categories. Interestingly, the Retrieval Model BM25 shows a pattern distinct from the other models. Within a Topic Representation, the overlap of this model seemed to have an interaction with Text Subset, but not the Aggregation Method. In fact, there did not appear to be differences across Aggregation Method within the same Topic Representation.

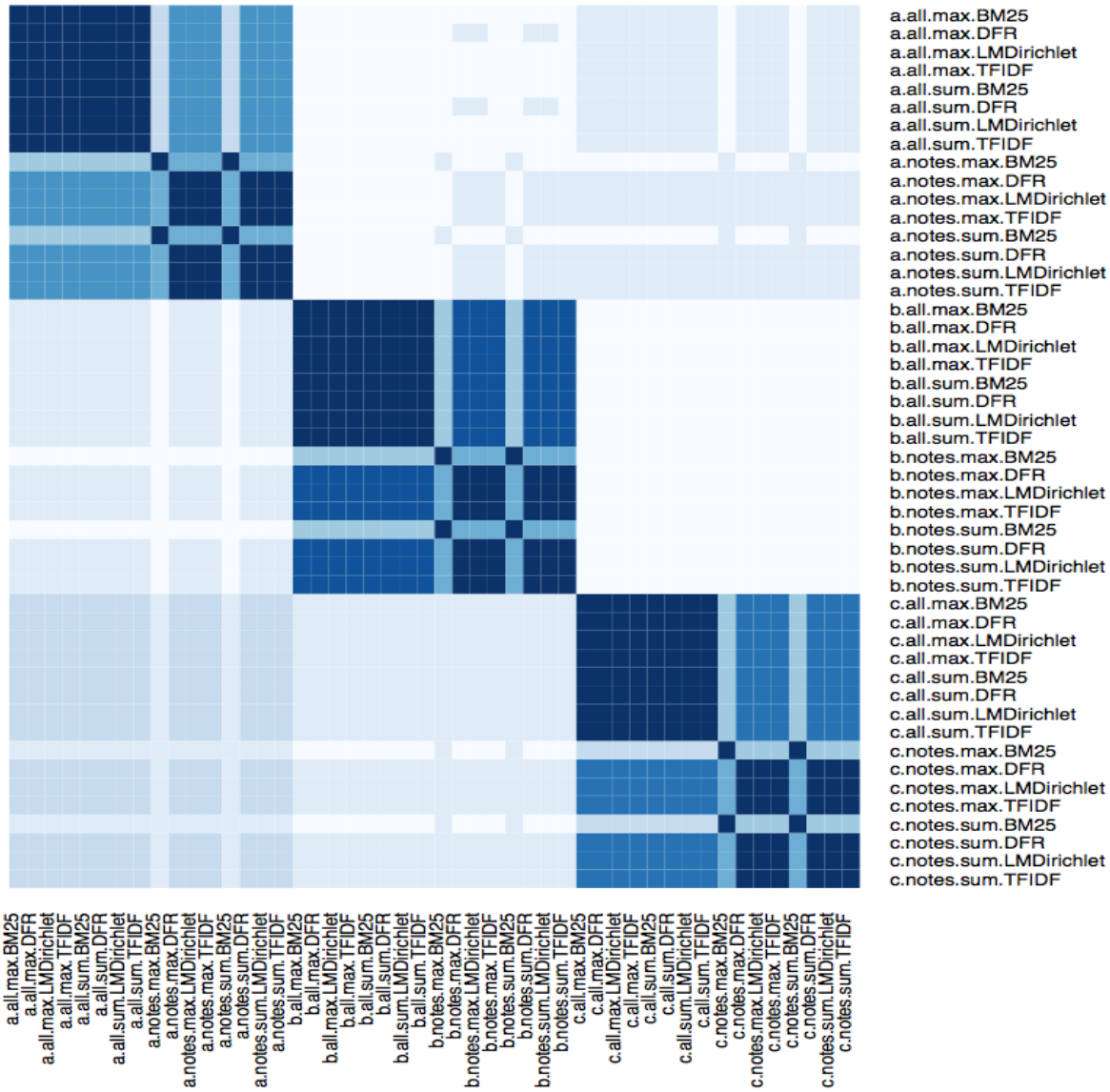

Figure 3. Overlap patterns illustrated between parameter settings for the top 1000 retrieved per topic, all topics combined. Color intensity represents percentage of patients in the row parameter setting also retrieved in the column parameter setting. Darker colors are higher percentages. The parameter settings are ordered hierarchically first by Topic Representation (A,B,C), then Text Subset (all, notes), then Aggregation Method (max, sum) and finally the Retrieval Model (BM25, DFR, LMDirichlet, TFIDF). Query performance is not considered in this figure.

3.2 Taxonomy Analysis

Table 2 lists the 59 elements of the taxonomy, while Table 3 lists the six binary features derived from the 59 taxonomy elements. We found moderate, substantial or almost perfect agreement by

Fleiss kappa on 50 of the 56 topics, rated by the three clinically trained raters for the 59 query taxonomy characteristics (Figure 4).

Medical condition explicitly present?
Systems: CV
Systems: GI
Systems: Neuro/Psych
Systems: Dental
Systems: Accident
Systems: Ophthalmic
Systems: Metabolic
Systems: Infectious
Systems: GU
Systems: Pulmonary
Systems: Hematology
Systems: Reproductive
Systems: Oral
Systems: Skin
Systems: Cancer
Systems: Breast
Systems: Immunology
System: Endocrine
System: Musculoskeletal
Age: none
Age: adult
Age: child
Age: senior
Sex: none
Sex: female
Sex: male
Medical diagnosis (exc)? ICD or Text
Medical diagnosis (inc)? ICD or Text
Medication (inc)?
Medication (exc)?
Labs?
Labs? (if yes) Values also?
Imaging?
Imaging? (if yes) Interpretation also?
Physical exam?
Physical exam? (if yes) Values also?
Procedures/Surgeries (inc)?
Procedures/Surgeries (exc)?
Location? (none)
Location? (inpatient)

Location? (outpatient)
Location? (emergency)
Location? (ICU)
Provider Type
Visit Type
Visit Specialty
Status? (deceased)
Status? (living)
Temporal: sequence of events
Temporal: age at first diagnosis
Temporal: time with diagnosis
Temporal: time on medication
Temporal: date of procedure
Temporal: age at encounter
Temporal: date of procedure/lab
Free text descriptions required, simple term?
Free text descriptions required, complex concept?
Number of encounters
'Complexity': number of required simple conditions (ICD, CPT, Meds, Labs) + temporal checks + simple free text requirements + complex free text requirements

Table 2. Taxonomy all features.

Temporal	The topic included a temporal component
Text	The structured query version of the topic also required unstructured data
Medication	The topic included or excluded a medication list.
Procedure	The topic included or excluded surgical procedures.
Additional	The topic included additional information about labs, imaging or physical exams.
Condition	The topic explicitly included information about a specific medical condition.

Table 3. Taxonomy binary features.

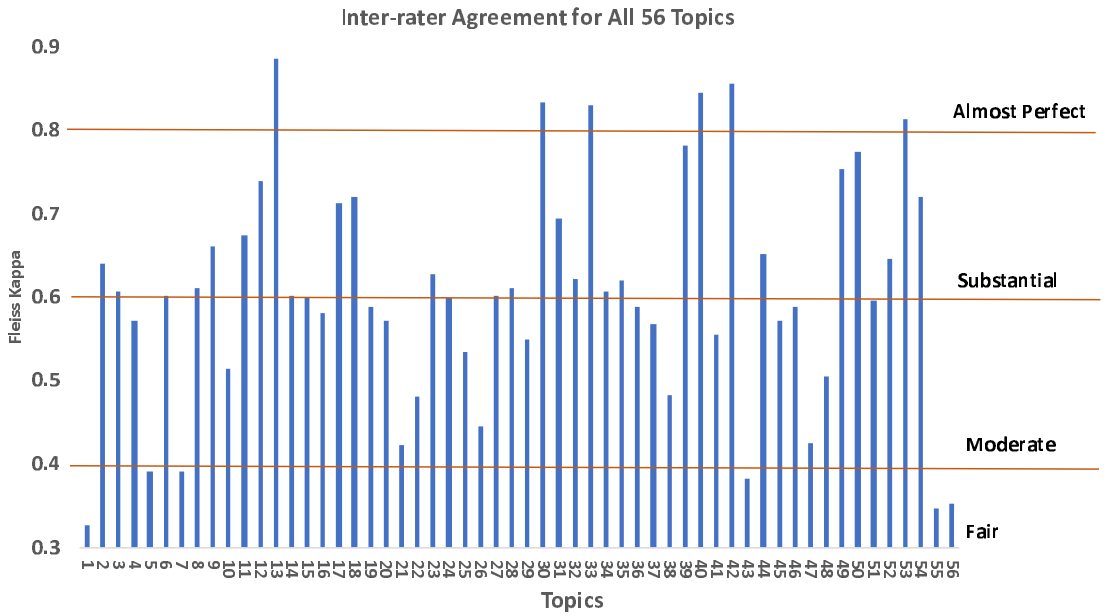

Figure 4. Interrater agreement for the 59 taxonomy characteristics applied to 56 topics. Fleiss Kappa was calculated for each topic based on agreement between three clinical trained raters on the 59 taxonomy characteristic assignments.

We next used heatmaps to investigate the relationship between word-based query performance and assignment to the 59 taxonomy characteristics (Figure 5). Topic clusters, based on B-Pref performance (left heatmap), were maintained for taxonomy characteristics (right heatmap), but column clustering was allowed for this heatmap. Performance-based clustering of topics can be seen for the B-Pref heatmap, but there do not appear to be similar patterns found in the taxonomy heatmap. For this reason, the next analysis was designed to focus specifically on the subset of query taxonomy characteristics more related to complexity, and less on content. To do this we derived six binary features from the 59 taxonomy characteristics.

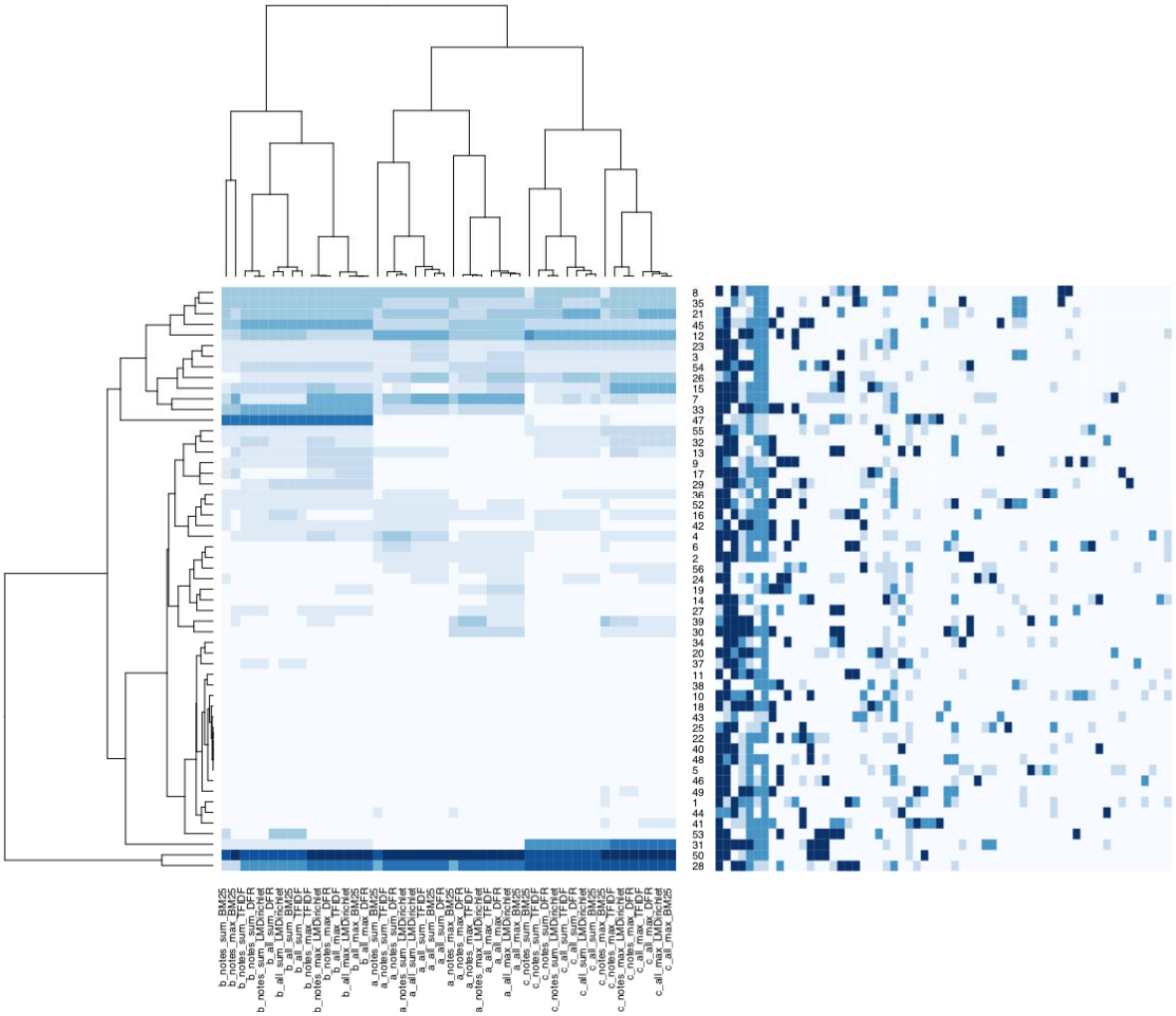

Figure 5. Comparison B-Pref Performance and Taxonomy Characteristic Clustering. The heatmap on the left are the topics by word-based query parameter settings clustered for B-Pref performance. Darker colors represent higher B-Pref values. The heatmap on the right maintains topic order from B-Pref clustering on the left (rows), but the columns, which represent the 59 taxonomy characteristic assignments, are clustered. Darker colors on the right heatmap represent the level of interrater agreement, but the lightest color means no association between the characteristic and the topic.

The first binary feature was positive if there was a temporal component in the topic ('Temporal', y/n). The 56 topics contain a variety of temporal conditions, including age at first diagnosis, time with diagnosis, chronological order of disease onset for several diagnoses, and medication use before or after first diagnosis. The second binary feature was positive if the topic could not be written exclusively with structured data, required some free text ('Text', y/n). The third binary feature was positive if the topic required a medication list check, either exclusions or inclusions or both ('Medication', y/n). The fourth binary feature was positive if there was a procedure in the

topic. This includes any surgical or non-surgical procedure ('Procedure', y/n). The fifth binary feature was positive if additional values were required from lab tests, imaging, or physical exams. ('Additional', y/n). And finally, the sixth binary feature was positive if the topic required a specific disease diagnosis or diagnoses. Some topics were defined for cohorts who only received certain screening tests without an explicit disease requirement ('Condition', y/n).

For our data, the beta regression modeling did show that five of the six binary taxonomy features were associated with poorer performance, as measured by B-Pref. One feature, 'text', was associated with better performance. Features associated with poorer performance were designed to capture increased topic complexity, so this result is not surprising. The feature 'text' captures the ability of purely structured data to describe a medical topic, with or without added free text. Our result indicates that topics that require text, in addition to structured data, might perform better. And there were notable interactions between the taxonomy features and the run parameters, particularly between the feature 'temporal' and Topic Representation. Interestingly this analysis did not point to any notable interactions between the four word-based parameters. But it is not clear if these results are generalizable due to the specific nature of our 56 topic descriptions.

We used the beta regression model, containing the four word-based parameters, six binary taxonomy features and the interactions between the parameters and features, to predict B-Pref with a simulated dataset. This dataset contained all possible permutations of the ten predictors. We varied each of the six binary taxonomy flags independently, while holding all other values constant, to estimate the impact of these flags. We also did this for Topic Representation (Figure 6). We again saw that five of the six binary taxonomy features were associated with poorer performance, and the feature 'text' was associated with improved performance. We also saw Topic Representation B associated with improved performance.

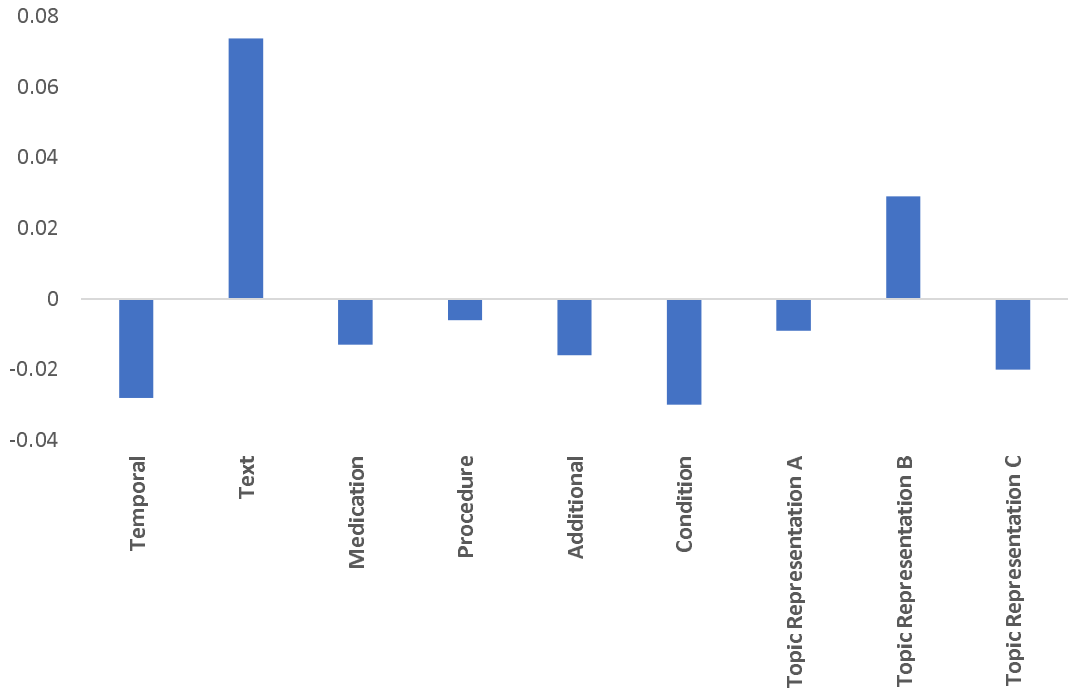

Figure 6. B-Pref prediction changes on simulated data for selected characteristics. Values represent change in predicted B-Pref using a simulated dataset with all permutations of the four word-based query parameters and the six binary taxonomy flags. All other values are held constant while varying the characteristic listed.

We also created a heatmap of the predicted B-Pref values generated from the simulated data (Figure 7). The x axis contained all possible permutations of the four word-based query parameters and the y axis contained all possible permutations of the six binary taxonomy features, and hierarchical clustering was done in both dimensions. Clear patterns of performance clustering can be seen, particularly around the combinations of three of the binary taxonomy features, temporal, text and condition. These three features are conceptually different from the other three (medication, procedure, additional) in that the latter are simple additions of information but the former represent more complex topic structural aspects. In addition, within specific combinations of these flags there are also clear variations in performance across different word-based parameter settings. In the bottom horizontal cluster, the best performance for topics without a temporal, text and condition component is seen with a completely different set of parameter settings than for topics with all three of these structural components. This performance pattern is an example of a possible interaction between the parameters and features, which could help guide the selection of parameters to optimize retrieval results.

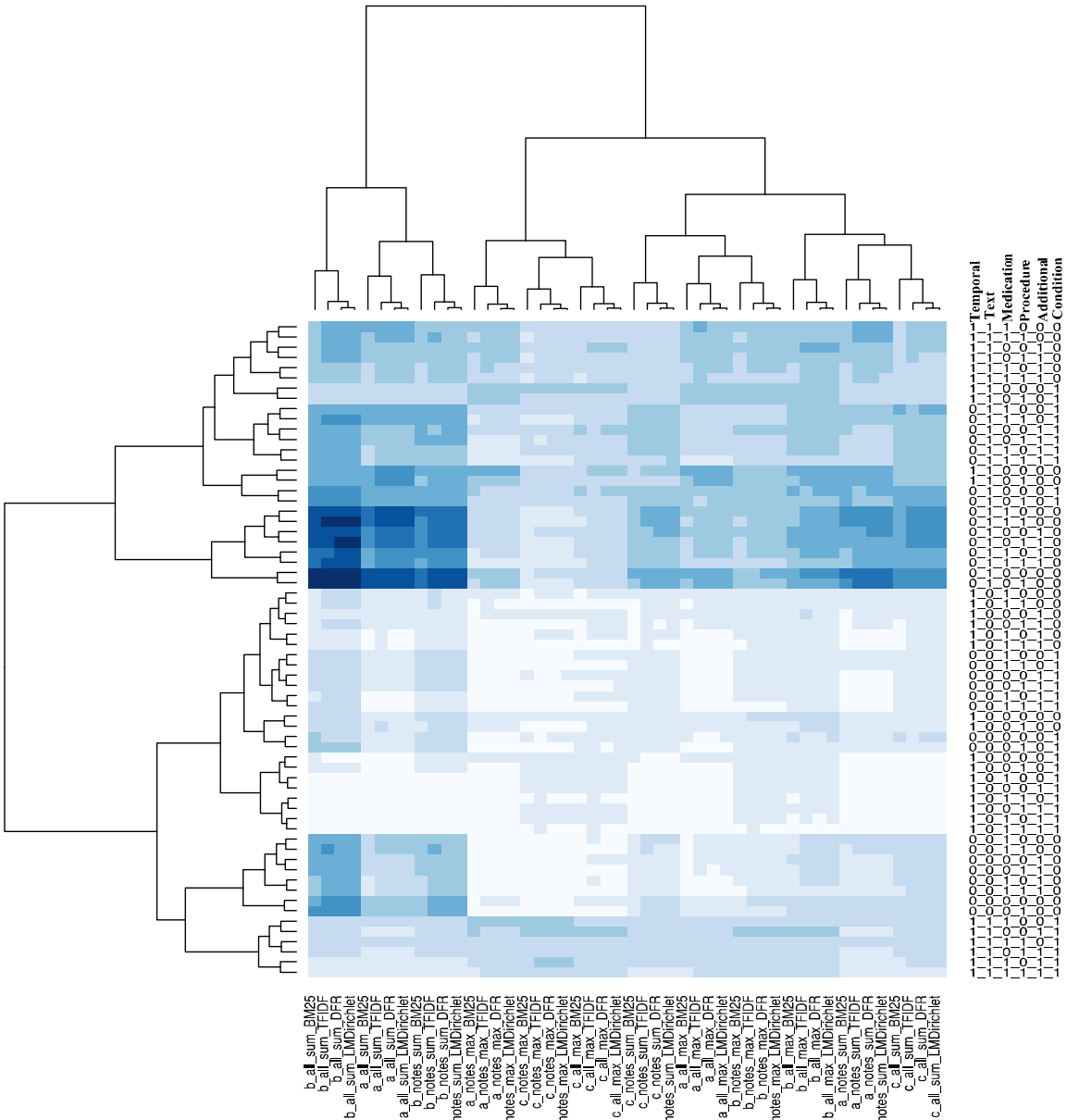

Figure 7. B-Pref predictions on simulated data. Darker values represent higher predicted B-Pref. Clustering was done on both axes. The x axis represents all permutations of the four word-based query parameters and the y axis represents all permutations of the six binary taxonomy features.

3.3 Structured Queries

For each topic, we calculated simple recall and precision on the output of each structured Boolean query (the system could not rank Boolean output) using the relevance judgments for the word-based query pools (Section 3.1). As with the word-based queries, a patient was considered relevant if rated definitely or possibly relevant. Table 4 shows an example structured Boolean

query for Topic 7. Recall for the structured queries varied widely across topics (Figure 8). There was 100% recall of word-based query relevant patients on 8 of the 56 topics, greater than 50% recall on 35 of the 56 topics, less then 50% recall on 13 of the 56, one topic (48) with no recall of relevant patients, and two topics with no relevant patients (22, 25). All of the topics with 100% recall came from OCTRI data requests. For topics with recall less than 100% but greater than 50%, about half came from OCTRI requests and half from the other sources. For the 13 topics with less than 50% recall, 9 came from the other sources. The structured queries were not able to return relevant patients as well as the word-based queries for some topics.

```
(demographics.BIRTH_DATE: Range[1913-01-01, 1995-12-31])

AND

(
encounter_diagnoses.DX_ICD=448.0
OR
hospital_encounters.ADMITTING_DX_ICD_CODE=448.0
OR
hospital_encounters.BILL_DISCHARGE_DX_ICD_CODE=448.0
OR
hospital_encounters.hospital_encounters.BILL_DX2_ICD_CODE=448.0
OR
hospital_encounters.BILL_DX3_ICD_CODE=448.0
OR
hospital_encounters.BILL_DX4_ICD_CODE=448.0
OR
hospital_encounters.ENCOUNTER_DIAGNOSIS=448.0
OR
problem_list.DX_ICD=448.0
OR
notes. NOTE_TEXT  contains " Hereditary hemorrhagic telangiectasia"
OR
notes. NOTE_TEXT  contains "Osler-Weber-Rendu"
 )
```

Table 4 – Structured Boolean query for Topic 7: Adults 18-100 years old who have a diagnosis of hereditary hemorrhagic telangiectasia (HHT), which is also called Osler-Weber-Rendu syndrome.

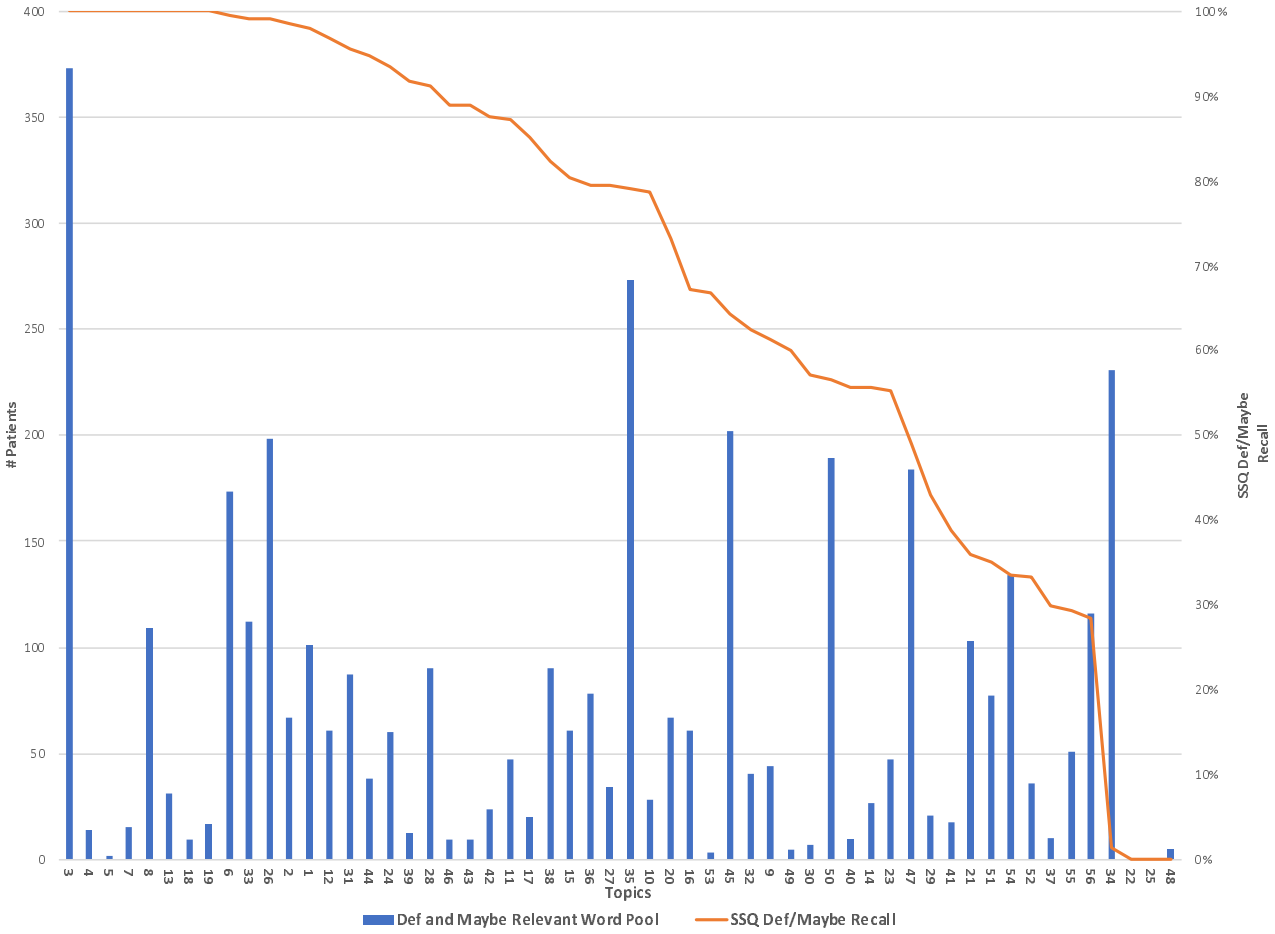

Figure 8. Recall for structured queries, ordered by recall for each topic (red line). The bars represent the total number of relevant patients found in each topic judged pool.

Precision likewise varied widely across topics (Figure 9). Structured query precision is represented by the red line and word-based query precision, at the topic pool level, is represented by the blue line. The structured queries outperformed the word-based queries in precision for all topics except 48. Again, topics 22 and 25 did not return any relevant patients. Three topics had 100% precision (29, 34, 46).

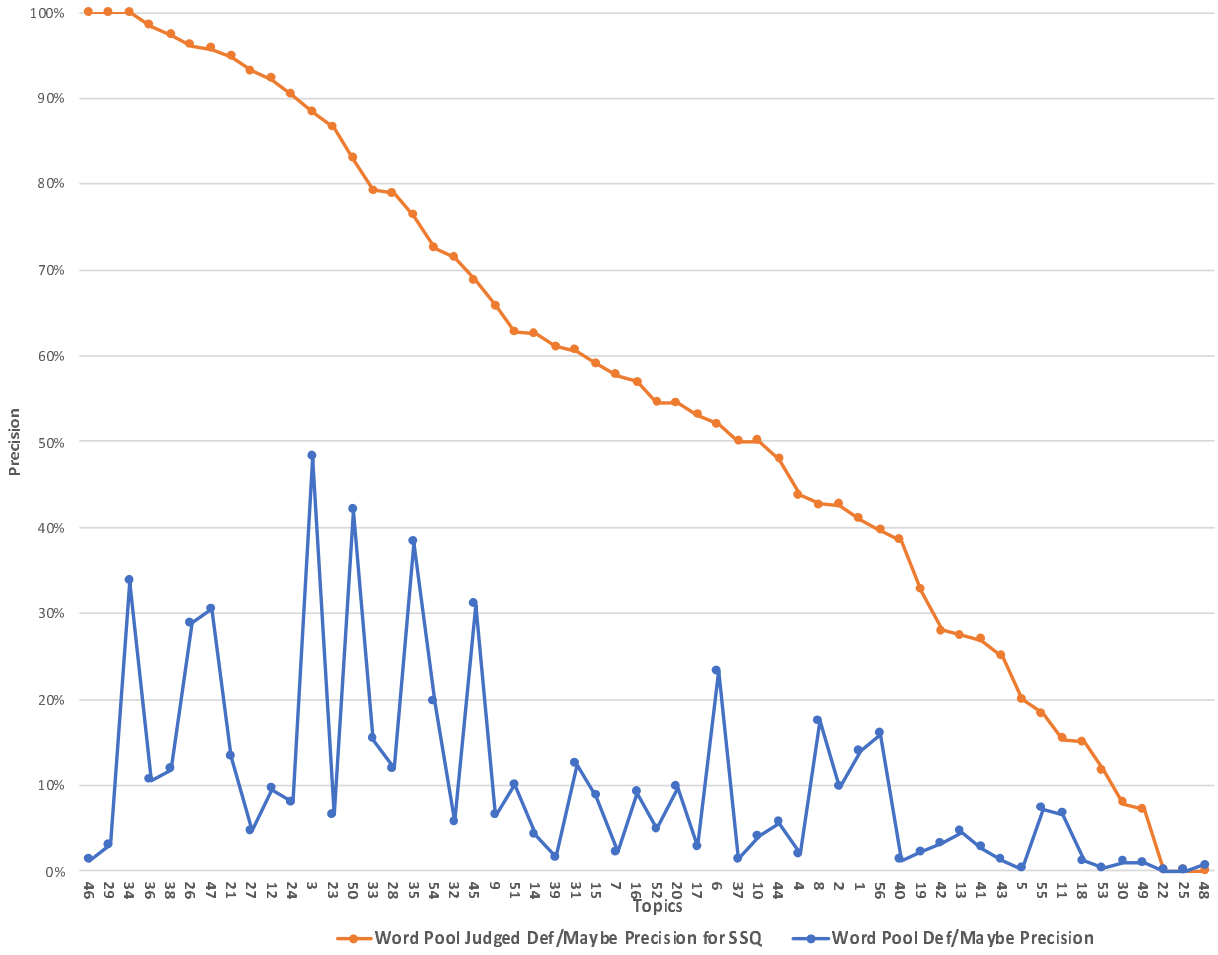

Figure 9. Precision for structured queries (red line) and word-based judged pools (blue line), ordered by structured query precision.

3.4 Topics with Expanded Relevance Judgements for the Structured Queries

Because the structured queries retrieved patients who had not been retrieved by the word-based queries, we did additional relevance judging for ten selected topics. Due to the large number of patients returned for topic 2 by the structured query (2,578), only a random sample of 750 patients was judged.

Although the structured queries had higher recall than the word-based queries for all ten topics, these queries did not achieve complete recall of all of the relevant patients for nine of the ten topics (Figure 10). The numbers of relevant patients found only in word-based queries was relatively low, compared to the total number of relevant patients (Table 5). This explains the larger number of missed relevant patients for this topic. The structured queries had higher precision for all ten topics (Figure 11). For topic 52, all patients retrieved by the structured query were judged relevant.

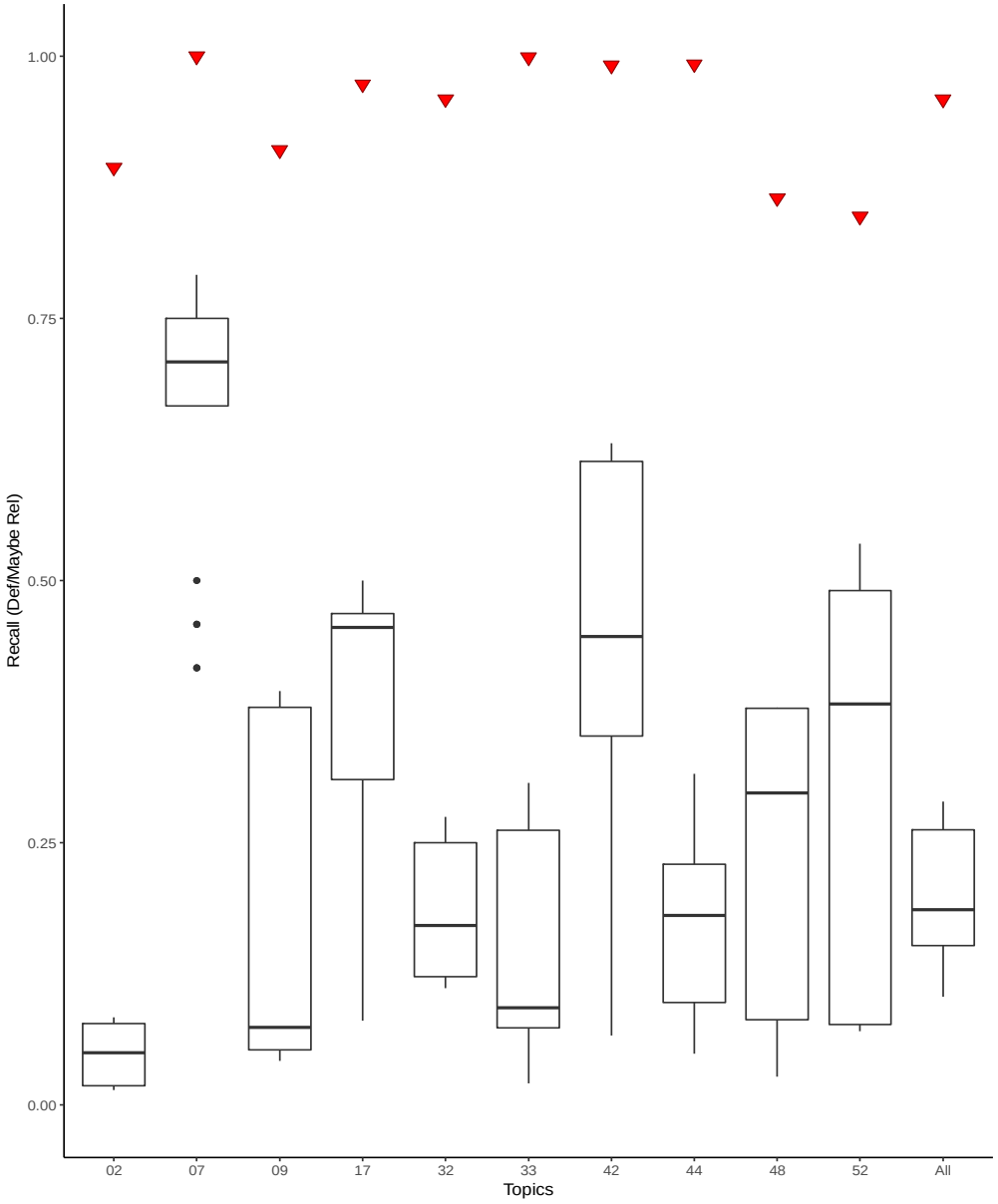

Figure 10. Recall distributions for ten selected topics on combined full structured query relevance judged pools and word-based sample relevance judged pools (using both definitely relevant and maybe relevant patients). Distributions are recall for all word-based parameter combinations, for each topic. Red triangles are the values for the structured queries using both pools. Topics are ordered by median recall for the word-based query distributions.

Topic	Structured query	Word-based	Structured query	Structured query	Recall for structured	Precision for	Structured query

	patients retrieved	query relevant	additional relevant	relevant and retrieved	query	structured query	relevant missed
2	750	67	490	438	0.89	0.58	52
7	50	15	24	24	1.00	0.48	0
9	357	44	190	173	0.91	0.48	17
17	110	20	112	109	0.97	0.99	3
32	390	40	368	353	0.96	0.91	15
33	1092	112	983	982	1.00	0.90	1
42	347	24	347	344	0.99	0.99	3
44	378	38	266	264	0.99	0.70	2
48	68	5	37	32	0.86	0.47	5
52	133	36	157	133	0.85	1.00	12

Table 5. Ten topics with additional relevance judgments for results from structured Boolean queries. The structured queries retrieved additional patients who were judged for relevance, allowing calculation of recall and precision for these queries as well as determination of numbers found by the word-based queries but missed by the structured queries.

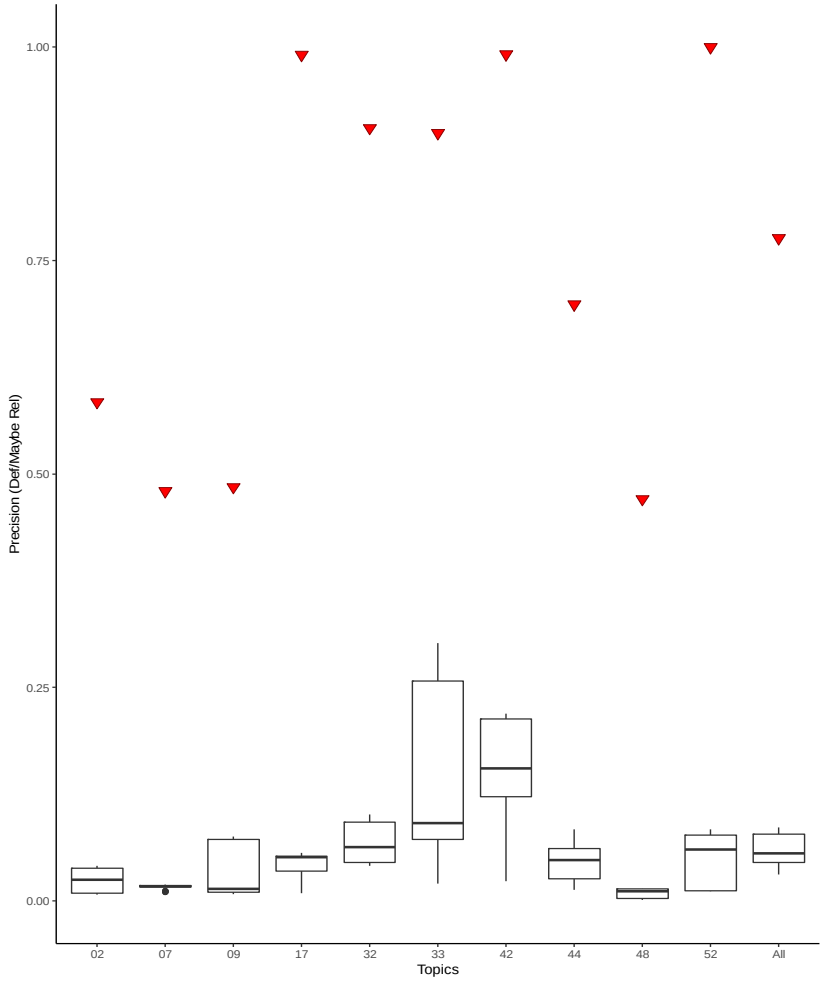

Figure 11. Precision distributions for ten selected topics on combined full structured query relevance judged pools and word-based sample relevance judged pools (using both definitely relevant and maybe relevant patients). Distributions are precision for all word-based parameter combinations, for each topic. The red triangles are the values for the structured queries. Topics are ordered by median precision for the word-based query distributions.

4. Discussion

We set out to begin this work using word-based query methods that performed well for the TREC Medical Records Track. Our results did not achieve the performance we expected (Figure 1). Overall, the best results were achieved with the Topic Representation of the illustrative clinical case formulation (B), with small further improvements for using Text Subset all and Aggregation Method max. With this combination of Topic Representation, Text Subset, and Aggregation Method, there was relatively little variation for three of the four Retrieval Model parameters, although BM25 scored worse.

Within our results, we observed variation common to IR challenge evaluations. Although the overall differences were modest, there was consistently higher values for Topic Representation B. Likewise, there was small benefit for Aggregation Method sum vs. max. For combinations of parameters, the Retrieval Model BM25 performed worse than the other three. To the extent that these results are generalizable, clinical case formulations are best query type among word-based methods for the patient cohort discovery task.

Also common to IR challenge evaluation results, reflecting the adage that means and medians can obscure variations, there was a large difference in retrieval performance by topic. As seen in Figure 2, about ten have very poor performance while two have very high performance across all retrieval methods. There is also a substantial range of performance within a number of individual topics. A final aspect common to IR challenge evaluation results is the diversity of patients retrieved by variation of the different parameters. The heat map in Figure 3 shows that Topic Representation plays a large role in this diversity.

The overall suboptimal performance of word-based methods led us to pursue two further methods to understand and improve our results. In an attempt to understand our results, we developed a taxonomy for the topics that we hoped would identify characteristics associated with the differences in results. We first developed an exhaustive 59 parameter taxonomy that did not reveal any associations. However, when we reduced the taxonomy to six binary variables, we did find association with performance. As also shown by comparable work at Mayo Clinic (28), it may be possible with further prospective analysis that query taxonomy might lead to selection of different query approaches based on characteristics of the topic.

In the effort to improve our results, we reformulated our queries using structured approaches developed iteratively. Because pure Boolean queries do not rank their output, we could not directly compare our results with the word-based queries. Instead, we measured standard recall and precision based on the relevance judgments made for patients retrieved by the word-based methods. The results for the structured queries were much better, with a median recall of 0.86 and eight topics having recall of 1.0 (Figure 7). There were likewise 13 topics with recall under 0.4 and a couple near zero. Precision was not associated with recall for the topics but did vary almost linearly from 1.0 to 0.0 across the topics (Figure 8).

One concern for the structured queries was the use of the relevance judgments only from the word-based query results. As such, we performed additional relevance judgments based on the structured query retrieval for 10 topics. This not only would give us a more realistic picture of the performance of these queries, but also identify additional patients for relevance judgment for the word-based queries. After the additional judgments, we found that the structured queries had much higher recall than the word-based queries (Figure 8) as well as much higher precision (Figure 9), which has also been found by comparable results from Mayo Clinic (35). The latter would of course be expected given the default retrieval set size of 1000 for the word-based queries, although retrieving fewer than 1000 would reduce recall from those queries.

We reviewed a sample of the relevant patients found only in word-based queries for most of the ten topics (Table 5). Most of these relevant patients missed by the structured queries were due to inconsistencies between the structured data (diagnosis or procedure codes, medication lists) and

the free text found in clinical notes. These inconsistencies, and others, were partly due to the date window limitation used to select encounter data for this test dataset. Some of the patients listed for topics 33, 42, 44 and 48 did not have the relevant diagnosis codes in any encounters, but did have the required diagnoses listed in clinical notes. Patients listed for topic 52 did not have the procedure codes for cataract surgery listed in any surgery encounters, but did have this procedure listed in clinical notes. Topics 32 and 42 both had a medication requirement that was not found in any medication data types, but was mentioned in the clinical notes. In addition, the structured query form for topic 32 also included a free text search of the clinical notes for a medication adverse event (cough associated with an ace inhibitor medication). The word-based query relevant patient had a wording for this adverse event not included in the structured query. Topic 9 required an age at first diagnosis less than four years for epilepsy. The structured query form used the earliest encounter with the associated diagnosis code for this calculation. Since the data had a limited range of visit data the age requirement could not be calculated for older patients.

Overall, our work has found that patient cohort discovery for clinical study recruitment is feasible with EHRs, although manual crafting and iteration of structured queries is usually required, which is a different result from general IR evaluation showing comparable performance for automated methods (36). There is some evidence that applying a query taxonomy might improve performance. Further work with methods such as machine learning might yield improvements, although it is not clear what features will lead to performance improvement across varying topical criteria for different queries.

There were a number of limitations to this work. Our records were limited to a single academic medical center. There are many additional retrieval methods we could have assessed, but we would not have the resources to carry out the additional relevance judgments required as those additional methods would add new patients to be judged. Finally, there is a global limitation to work with EHR data for these sorts of use cases in that raw, identifiable patient data is not easily sharable such that other researchers could compare their systems and algorithms with ours using our data.

5. Conclusions

Cohort retrieval is commonly offered by many medical centers with EHR systems but is poorly understood and, with current approaches, still labor-intensive. Automated methods can likely improve performance of systems and reduce time taken to identify definitely relevant patients, although manual crafting of Boolean queries showed much better performance in this research. Challenges to developing and evaluating IR methods for this use case include the resources required to perform relevance judgments and the nature of such highly private data that makes their comparison across different research groups difficult. Our future work will continue to develop methods that show promise and evaluate them with real-world topics and relevance judgments in our data.

Acknowledgments

This work was supported by NIH Grant 1R01LM011934. The authors also thank Jack Wiedrick, PhD for providing statistical consulting for the beta regression modeling and data simulation.

References

1. Obeid J, Beskow L, Rape M, Gouripeddi R, Black R, Cimino J, et al. A survey of practices for the use of electronic health records to support research recruitment. Journal of Clinical and Translational Science. 2017;1:246-52.
2. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Medical Informatics & Decision Making. 2015;15:28.
3. Ni Y, Kennebeck S, Dexheimer J, McAneney C, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. Journal of the American Medical Informatics Association. 2015;22:166-78.
4. Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. JMIR Medical Informatics. 2019;7(3):e14185.
5. Chapman W, Nadkarni P, Hirschman L, D'Avolio L, Savova G, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011;18:540-3.
6. Friedman C, Rindflesch T, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. Journal of Biomedical Informatics. 2013;46:765-73.
7. Chapman W, Saul M, Houston J, Irwin J, Mowery D, Harkema H, et al., editors. Creation of a repository of automatically de-identied clinical reports: processes, people, and permission. Proceedings of the American Medical Informatics Association Clinical Reserach Informatics Summit; 2011; San Francisco, CA.
8. Johnson A, Pollard T, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.
9. Voorhees E, Tong R, editors. Overview of the TREC 2011 Medical Records Track. The Twentieth Text REtrieval Conference Proceedings (TREC 2011); 2011; Gaithersburg, MD: National Institute of Standards and Technology.
10. Voorhees E, Hersh W, editors. Overview of the TREC 2012 Medical Records Track. The Twenty-First Text REtrieval Conference Proceedings (TREC 2012); 2012; Gaithersburg, MD: National Institute of Standards and Technology.
11. Cleverdon C, Keen E. Factors determining the performance of indexing systems (Vol. 1: Design, Vol. 2: Results). Cranfield, England: Aslib Cranfield Research Project; 1966.
12. Voorhees E, editor The TREC Medical Records Track. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics; 2013; Washington, DC.
13. Zhu D, Wu S, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. Journal of Biomedical Informatics. 2014;49:275-81.
14. Goodwin T, Harabagiu S. Learning relevance models for patient cohort retrieval. JAMIA Open. 2018;1:265–74.
15. Sarmiento R, Dernoncourt F. Improving Patient Cohort Identification Using Natural Language Processing. In: Anonymous, editor. Secondary Analysis of Electronic Health Records. Cham, Switzerland: Springer; 2016. p. 405-17.

16. Glicksberg B, Miotto R, Johnson K, Shameer K, Li L, Chen R, et al., editors. Automated disease cohort selection using word embeddings from electronic health records. Pacific Symposium on Biocomputing; 2018.
17. Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. Journal of the American Medical Informatics Association. 2019:Epub ahead of print.
18. Ateya M, Delaney B, Speedie S. The value of structured data elements from electronic health records for identifying subjects for primary care clinical trials. BMC Medical Informatics & Decision Making. 2016;16:1.
19. Kang T, Zhang S, Tang Y, Hruby G, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association. 2017;24:1062-71.
20. Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. Journal of the American Medical Informatics Association. 2017;24:781-7.
21. Yuan C, Ryan P, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019;26:294-305.
22. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. Journal of the American Medical Informatics Association. 2018;25:530-7.
23. Gligorijevic J, Gligorijevic D, Pavlovski M, Milkovits E, Glass L, Grier K, et al. Optimizing clinical trials recruitment via deep learning. Journal of the American Medical Informatics Association. 2019:Epub ahead of print.
24. Denny J, Bastarache L, Roden D. Phenome-wide association studies as a tool to advance precision medicine. Annual Review of Genomics and Human Genetics. 2016;17:353-73.
25. Richesson R, Sun J, Pathak J, Kho A, Denny J. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artificial Intelligence in Medicine. 2016;71:57-61.
26. Robinson J, Wei W, Roden D, Denny J. Defining phenotypes from clinical data to drive genomic research. Annual Review of Biomedical Data Science. 2018;1:69-92.
27. Wu S, Liu S, Wang Y, Timmons T, H Uppili, Bedrick S, et al. Intra-institutional EHR collections for patient-level information retrieval. Journal of the American Society for Information Science & Technology. 2017;68:2636-48.
28. Wang Y, Wen A, Liu S, Hersh W, Bedrick S, Liu H. Test collections for electronic health record-based clinical information retrieval. JAMIA Open. 2019:Epub ahead pf print.
29. Robertson S, Walker S, editors. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 1994; Dublin, Ireland: Springer-Verlag.
30. Amati G, VanRijsbergen C. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems. 2002;20:357-89.
31. Zhai C, Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems. 2004;22:179-214.

32. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management. 1988;24:513-23.
33. Buckley C, Voorhees E, editors. Retrieval evaluation with incomplete information. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2004; Sheffield, England: ACM Press.
34. Fleiss J, Levin B, Paik M. The Measurement of Interrater Agreement. Statistical Methods for Rates and Proportions, Third Edition. Hoboken, NJ: John Wiley & Sons; 2003. p. 598-626.
35. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. CREATE: cohort retrieval enhanced by analysis of text from electronic health records using OMOP common data model. arXivorg. 2019.
36. Harman D. Information Retrieval Evaluation. San Rafael, CA: Morgan & Claypool; 2011.