Overview of the TREC 2014 Clinical Decision Support Track

Matthew S. Simpson[1], Ellen M. Voorhees[2], and William Hersh[3]

[1]Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD
[2]National Institute of Standards and Technology, Gaithersburg, MD
[3]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR

1 Introduction

In making clinical decisions, physicians often seek out information about how to best care for their patients. Information relevant to a physician can be related to a variety of clinical tasks such as determining a patient's most likely diagnosis given a list of symptoms, deciding on the most effective treatment plan for a patient having a known condition, and determining if a particular test is indicated for a given situation. In some cases, physicians can find the information they seek in published biomedical literature. However, given the volume of the existing literature and the rapid pace at which new research is published, locating the most relevant and timely information for a particular clinical need can be a daunting and time-consuming task.

To make biomedical information more accessible and to meet the requirements for the meaningful use of electronic health records, a goal of modern clinical decision support systems is to anticipate the needs of physicians by linking electronic health records with information relevant for patient care. The Clinical Decision Support Track aims to simulate the requirements of such systems and to encourage the creation of tools and resources necessary for their implementation.

The focus of the 2014 track was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. In the absence of a reusable, de-identified collection of medical records, we used short case reports, such as those published in biomedical articles, as idealized representations of actual medical records. A case report typically describes a challenging medical case, and it is often organized as a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to the case.

Participants of the track were challenged with retrieving, for a given case report, full-text biomedical articles relevant for answering questions related to several types of clinical information needs. Each topic consisted of a case report and one of three generic clinical question types, such as "What is the patient's diagnosis?" Retrieved articles were judged relevant if they provide information of the specified type useful for the given case. The evaluation of the submissions followed standard TREC evaluation procedures.

In the remainder of this overview we describe the documents (Section 2) and topics (Section 3) used for the retrieval task and the evaluation (Section 4) of the retrieval results. We also include raw statistics (Section 5) summarizing the performance of the participants' submissions.

2 Documents

The target document collection for the track was an open access subset[1] of PubMed Central[2] (PMC), an online repository of freely available full-text biomedical literature. Because documents are constantly being

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] http://www.ncbi.nlm.nih.gov/pmc/

Table 1: Types of case-based topics

Type	Generic question	Ely et al.'s Classification	Frequency (%)
Diagnosis	What is the patient's diagnosis?	1.1.x.1: Diagnosis/Cause/*	21.28
Test	What tests should the patient's receive?	1.3.x.1: Diagnosis/Test/*	11.89
Treatment	How should the patient be treated?	2.1.2.x: Treatment/Drugs/Indications/*	13.61
		2.2.1.x: Treatment/General/Indications/*	5.95
All			52.72

added to PMC, to ensure the consistency of the collection, we obtained a snapshot of the open access subset on January 21, 2014, which contained a total of 733,138 articles. The full text of each article in the open access subset is represented as an NXML file (XML encoded using the U.S. National Library of Medicine's Journal Archiving and Interchange Tag Library),[3] and images and other supplemental materials are also available.

Each article in the collection is identified by a unique number (PMCID) that was used for run submissions. The PMCID of an article is specified by the `<article-id>` element within its NXML file. Although each article is represented by multiple identifiers (e.g., PubMed, PMC, Publisher, etc.), we used only PMCIDs for this task. The various identifier types are specified using the `pub-id-type` attribute of the `<article-id>` element. Valid values of `pub-id-type` that indicate a PMCID include `pmc` and `pmcid`. For example, the document identifier of an article with PMCID 3148967 might by specified in the article's NXML file as follows.

```
<article-id pub-id-type=”pmc”>
 3148967
</article-id>
```

To make processing the document collection easier for the participants, we renamed each article NXML in the collection according to the article's PMCID. For example, an article with PMCID 3148967 was given the name `3148967.nxml`.

Participants were able to obtain the document collection in one of two ways. First, participants who were only interested in indexing the text of the articles in the collection (most participants) could download files containing all 733,138 articles in the January 21, 2014 snapshot directly from the track's website. Second, for participants who were interested in utilizing additional media other than text, such as the images and videos included in the articles, track organizers published a Python script for downloading the full document content directly from the PMC Open Access FTP Service.[4] The total size of the collection with the additional media is about 2 TB. Downloading the additional media associated with the full-text articles was entirely optional for participation in the track, and none of the topics required this information. We provided this option for participants who had an interest in analyzing the medical images included in many of the articles as part of their retrieval strategies.

3 Topics

The topics for the track were medical case narratives created by expert topic developers at the U.S. National Library of Medicine that serve as idealized representations of actual medical records. The case narratives described information such as a patient's medical history, the patient's current symptoms, tests performed by a physician to diagnose the patient's condition, the patient's eventual diagnosis, and finally, the steps taken by a physician to treat the patient.

Having a set of case narratives to use as topics, there are many clinically relevant questions that can be asked of them. Ely et al. (2000) created a taxonomy of the most frequent questions posed in practice. They collected 1,396 clinical questions from 152 primary care physicians and categorized them into 64 generic

[3] `http://jats.nlm.nih.gov/archiving/versions.html`

[4] `http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/`

Table 2: Example topic descriptions

Topic	Type	Description
1	Diagnosis	A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.
11	Test	A 40-year-old woman with no past medical history presents to the ER with excruciating pain in her right arm that had started 1 hour prior to her admission. She denies trauma. On examination she is pale and in moderate discomfort, as well as tachypneic and tachycardic. Her body temperature is normal and her blood pressure is 80/60. Her right arm has no discoloration or movement limitation.
21	Treatment	A 21-year-old female is evaluated for progressive arthralgias and malaise. On examination she is found to have alopecia, a rash mainly distributed on the bridge of her nose and her cheeks, a delicate non-palpable purpura on her calves, and swelling and tenderness of her wrists and ankles. Her lab shows normocytic anemia, thrombocytopenia, a 4/4 positive ANA and anti-dsDNA. Her urine is positive for protein and RBC casts.

question types. Table 1 provides a coarse summary of some of their findings. The first column of the table indicates a broad category of clinical information need. The second column indicates the generic form of each question type. For example, the "diagnosis" type can be interpreted as posing the question: "What is the patient's diagnosis?" In the third column, we indicate which of Ely et al.'s 64 clinical question categories fit each generic form, and in the last column, we indicate how frequently questions of a given category were posed. The last row of the table indicates that clinical questions related to diagnoses, treatments, and tests account for a majority (52.72%) of the clinical questions posed by primary care physicians.

To simulate the actual information needs of physicians, our topic creators manually labeled the case narratives they constructed according to these three categories. A case narrative labeled "diagnosis," for example, requires participants of the track to retrieve PMC articles a physician would find useful for determining the diagnosis of the patient described in the report. Similarly, for a case narrative labeled "treatment," participants should retrieve articles that would suggest to a physician the best treatment plan for the condition exhibited by the patient described in the report. Finally, participants should retrieve for "test" case narratives articles that would suggest appropriate medical tests to be performed for either diagnosis or treatment of the patient. When constructing the case-based topics, the topic creators were careful to omit information related to the question type. For example, a "diagnosis" report might contain information pertaining to a patient's treatments and tests, but not the patient's diagnosis. In doing so, we hoped to more accurately mimic real clinical scenarios. The topic creators produced 10 topics for each of the 3 topic types for a total of 30 topics.

In addition to annotating the topics according to the type of clinical information required, we also provided two versions of the case narratives. The topic "descriptions" contain a complete account of the patients' visits, including details such as their vital statistics, drug dosages, etc., whereas the topic "summaries" are simplified versions of the narratives that contain less irrelevant information. A topic's description and its summary are functionally equivalent: the set of relevant documents is identical for each version. However, we provided the summary versions of the case narratives for participants who were not interested in nor equipped for processing the detailed descriptions.

Tables 2 and 3 show examples of the case-based topics. Table 2 contains descriptions for Topics 1, 11, and 21, and Table 3 contains their corresponding summaries. These particular topics are shown because they are examples of each of the 3 topic types used in the task.

To make the results of the track more meaningful, we required that participants use only all topic descriptions or only all topic summaries for any given run submission. Participants were free to submit up to five runs so that they could experiment with the different representations. The meta-data collected about a run included which version of the topics was used for the run.

The topics were provided in XML format. Topic numbers were specified using the `number` attribute of each `<topic>` element and topic types (i.e., diagnosis, test, and treatment) were specified with the `type`

Table 3: Example topic summaries

Topic	Type	Summary
1	Diagnosis	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.
11	Test	40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.
21	Treatment	21-year-old female with progressive arthralgias, fatigue, and butterfly-shaped facial rash. Labs are significant for positive ANA and anti-double-stranded DNA, as well as proteinuria and RBC casts.

attribute. Topic descriptions were given in `<description>` elements and topic summaries were given in `<summary>` elements. Below is an example of the format.

```
<topics>
 <topic number="1" type="diagnosis">
  <description>
   Description of topic 1
  </description>
  <summary>
   Summary of topic 1
  </summary>
 </topic>
 ...
</topics>
```

4 Judgments

The retrieval task in the track was an ad hoc task. Participants were permitted to submit in `trec_eval` format a maximum of five automatic or manual runs, each run consisting of a ranked list of up to one thousand PMCIDs per topic. As shown in Table 4, the track had a total of 26 participants that together submitted a total of 102 retrieval runs. We were encouraged by the broad participation in a medically-oriented track, especially given that 2014 was the inaugural year of the Clinical Decision Support track.

All of the 102 runs contributed to the judgment sets, which were constructed to be compatible with computing inferred retrieval measures (Yilmaz et al., 2008). Inferred measures are used as a means of getting more accurate estimates of a run's quality than is likely possible with traditional measures when judging a relatively small number of documents. The runs were sampled following an effective sampling strategy (Voorhees, 2014) for computing inferred measures. In particular, judgment sets were created using two strata: all documents retrieved in ranks 1–20 by any run in union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 21–100 by some run. Documents in the judgment set were judged on a three-point scale of 0: "not relevant," 1: "possibly relevant," and 2: "definitely relevant." For the evaluation reported here, the measures were computed by conflating the possibly relevant and definitely relevant sets into a single relevant set, except for the infNDCG measure, which makes use of the different relevance grades. A total of 34,949 documents were judged across the topics, with a mean of 1265.0 [min: 908, max: 1669] documents judged per topic.

The assessment was performed by physicians, most of whom are biomedical informatics students in the Department of Medical Informatics and Clinical Epidemiology at Oregon Health & Science University. (A few are physicians from other sites. The topic creators did not perform assessment.) For a document to be judged definitely relevant to a given topic, it had to provide information of the specified type (i.e., diagnosis, test, and treatment) and provide information relevant to the particular patient described in the topic. The assessors were encouraged to not view a retrieved article as providing a "correct answer" to the generic clinical question posed by the topic, but were instead instructed to judge a document relevant if there was a reasonable chance a physician might find the article useful having seen the patient described in

Table 4: Participating groups and submitted runs

Group		Affiliation	No. Runs
1.	BiTeM_SIBtex	University of Applied Sciences, Geneva	5
2.	BigPig	University of Michigan	3
3.	CSEIITV	Indian Institute of Technology (BHU), Varanasi	2
4.	DA_IICT	Dhirubhai Ambani Institute of Information and Communication Technology	5
5.	DawitAfshin	Dawit Girmay and Afshin Deroie, York University	4
6.	ECNUCS	East China Normal University	4
7.	Georgetown	Georgetown University	5
8.	HENRI_TUDOR_LUX	CRP Henri Tudor	5
8.	IKMLAB	Institute of Medical Informatics, NCKU	1
10.	KISTI	Korea Institute of Science and Technology Information	5
11.	LIMSI	LIMSI-CNRS	5
12.	Merck_DA	Merck KGaA	3
13.	NovaSearch	Universidade Nova Lisboa	5
14.	OHSU	Oregon Health & Science University	4
15.	Philips	Philips Research North America	1
16.	SNUMedinfo	Medical Informatics Laboratory	5
17.	TUW	Vienna University Of Technology	5
18.	UCLA_MII	Medical Imaging Informatics, University of California, Los Angeles	5
19.	UTDHLTRI	University of Texas at Dallas	4
20.	atigeo	Atigeo	5
21.	cuhk_sls	The Chinese University of Hong Kong	5
22.	georgetown_ir	Georgetown University IR Lab	5
23.	hltcoe	Johns Hopkins University Human Language Technology Center of Excellence	4
24.	ir.cs.sfsu	Computer Science Department, San Francisco State University	1
25.	super_kxlab	bupt-kxlab	1
26.	udel_fang	InfoLab Group at the University of Delaware	5
Total submitted runs			102

the topic. Documents were judged not relevant if they either did not provide information of the specified type or they were not topical to the patient. Finally an article was judged possibly relevant if an assessor believed it was not immediately informative on its own, but that it may be relevant in the context of a broader literature review.

Once the initial judgments were obtained, eight topics were independently rejudged by a different assessor. Table 5 shows the agreement between the two assessors for these topics when the two relevance levels were conflated into a single relevant category. The middle columns in the table give the counts of the number of documents that fall into the cells of the contingency table; for example, column "RN" gives the counts of the number of documents for which the first assessor judged it relevant and the second judged it not relevant. The final column gives the overlap of the relevant sets of the two assessors where overlap is defined as the size of the intersection of the relevant sets divided by the size of the union of the relevant sets.

Mean overlap across the eight topics with multiple judgment sets is somewhat on the low side as compared to other studies of relevance judgment agreement (Voorhees, 2000), but is not inconsistent with those studies, especially given the small sample set size of just eight topics.

5 Results

Figure 1 shows the distribution of evaluation scores per topic computed across the set of 102 submitted runs. The graph on the left of the figure shows infNDCG scores computed at a cut-off of 100 and graph on the right shows Precision(10) scores. The graphs are box-and-whisker plots in which the horizontal line in a box is the median value, the lower and upper limits of the box are the first and third quartile values, the whiskers extend to values that are within 1.5 × interquartile-distance, and circles plot individual outliers. Figure 2 shows the topic type and topic text of topics that had interesting behavior as measured by infNDCG.

Recall that topics 1–10 are of type "diagnosis", topics 11–20 are of type "test", and topics 21–30 are of

Table 5: Agreement between assessors for dual-judged topics

Topic	NN	NR	RR	RN	Overlap
1	1349	32	35	47	0.3070
5	1360	1	14	119	0.1045
12	838	17	114	508	0.1784
17	1040	53	13	6	0.1806
19	977	25	70	134	0.3057
25	1351	70	28	6	0.2692
27	437	17	296	158	0.6285
28	1070	10	35	17	0.5645
Mean					0.3173

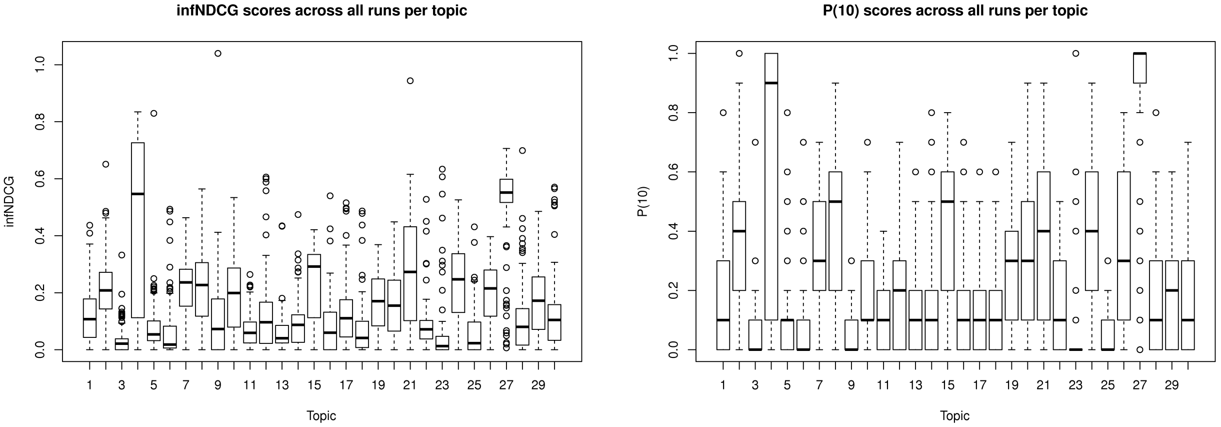

Figure 1: Per-topic scores computed over entire set of 102 runs for infNDCG (left) and P(10) (right)

Easiest: best median and best best infNDCG score

4: [diagnosis] *4-year-old boy with fever, conjunctivitis, strawberry tongue, desquamation of the fingers and toes*

9: [diagnosis] *soft, flesh-colored, pedunculated lesions on neck*

Hardest: worst median and worst best infNDCG score

23: [treatment] *heavy smoker with productive cough, shortness of breath, tachypnea, and oxygen requirement*

11: [test] *severe right arm pain and hypotension*

Large difference between best and median infNDCG scores

5: [diagnosis] *shortness of breath 3 weeks after surgical mastectomy*

21: [treatment] *progressive arthralgias, fatigue, and butterfly-shaped facial rash*

Figure 2: Notable topics based on infNDCG scores over all submissions

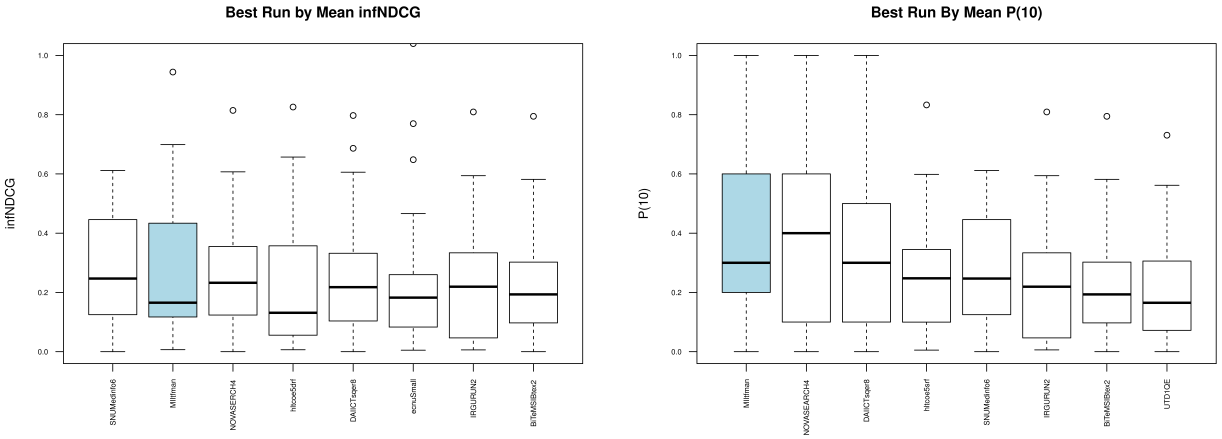

Figure 3: infNDCG (left) and P(10) (right) scores for the most effective run from the top 8 participants

type "treatment." There is little apparent difference in performance across topic types when considering all runs. Some participants did do topic-type-specific processing in their runs, for example by emphasizing particular MeSH terms related to the topic type when those terms were found in documents. However, participants found it difficult to improve retrieval effectiveness using such processing, largely because relevant documents for a topic do not necessarily have a focus on that type. That is, an article useful for diagnosing a case frequently is not an article focused on the process of diagnosis.

Figure 3 shows the distribution of evaluation scores across topics for individual runs. As in Figure 1, the figure shows box-and-whisker plots with infNDCG scores on the left and P(10) scores on the right. The runs included in the graph are the most effective runs by mean value of the respective measure for each of the top eight participants, and runs are ordered by decreasing mean. Runs plotted with blue shading are manual runs.

Generally, mean retrieval scores were relatively poor. This is perhaps reflective of the difficulty of the retrieval task, the utility of the document collection in providing relevant information for the types of generic clinical questions posed in this task, or challenges involved in assessing the retrieved documents.

The topic statements developed for the track contained both a longer description field and a shorter summary field, each field representing the same fundamental information need. The motivation for including both fields was the research question of whether systems can recognize and successfully down-weight the non-essential information included in the descriptions. Many participants included a direct comparison of otherwise identical summary- and description-based runs among their submissions. The summary-based runs were more effective than the description-based runs in these tests. However, it should be noted that the submission with the best mean infNDCG score was an automatic run that used the description field.

6 Conclusion

TREC 2014 was the inaugural year of the Clinical Decision Support Track. The broad goal of the track is to inform the creation of robust clinical decision support systems, and in doing so, help improve patient care. In this first year, we focused on linking idealized case reports to published biomedical literature and attempted to address common generic clinical information needs including inquiries pertaining to diagnoses, treatments and tests. Twenty-six groups participated in the track and together they submitted a total of 102 runs. Retrieval results were generally lower than expected reflecting the difficulty of the retrieval task, the utility of the document collection in addressing clinical needs, or challenges involved in assessing the retrieved documents. We hope to further investigate these issues in future evaluations.

Acknowledgements

We would like to thank Swapna Abhyankar, Amos Cahan, Dina Demner-Fushman, and Laritza Rodriguez for their help in preparing the topics; Suchet Chachra for his help in preparing the document collection; and the Department of Medical Informatics and Clinical Epidemiology at Oregon Health & Science University for performing the assessment.

Matthew Simpson was partially supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

Ely, J.W., Osheroff, J.A., Gorman, P.N., Ebell, M.H., Chambliss, M.L., Pifer, E.A., et al. A taxonomy of generic clinical questions: classification study. BMJ 2000;321(7258):429–432.

Yilmaz, E., Kanoulas, E., Aslam, J.A.. A simple and efficient sampling method for estimating AP and NDCG. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008, p. 603–610.

Voorhees, E.M.. The effect of sampling strategy on inferred measures. In: Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2014, p. 1119–1122.

Voorhees, E.M.. Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management 2000;36:697–716.