Overview of the TREC 2019 Precision Medicine Track

Kirk Roberts
School of Biomedical Informatics,
The University of Texas Health Science Center, Houston, TX

Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, Bethesda, MD

Ellen M. Voorhees
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD

William R. Hersh and **Steven Bedrick**
Department of Medical Informatics & Clinical Epidemiology,
Oregon Health & Science University, Portland, OR

Alexander J. Lazar
Departments of Pathology & Genomic Medicine,
The University of Texas MD Anderson Cancer Center, Houston, TX

Shubham Pant, **Funda Meric-Bernstam**
Department of Investigational Cancer Therapeutics,
The University of Texas MD Anderson Cancer Center, Houston, TX

1 Introduction

Precision medicine is a medical paradigm in which treatments are customized entirely to the individual patient. The underlying issue that drives precision medicine is that for many complex diseases, there are no "one size fits all" solutions for patients with a particular diagnosis. The proper treatment for a patient depends upon genetic, environmental, and lifestyle choices. The ability to personalize treatment in a scientifically rigorous manner based on these factors is thus the hallmark of the emerging precision medicine paradigm. Nowhere is the potential impact of precision medicine more closely focused at the moment than in cancer, where lifesaving treatments for particular patients could prove ineffective or even deadly for other patients based entirely upon the particular genetic mutations in the patient's tumor(s). Significant effort, therefore, has been devoted to deepening the scientific research surrounding precision medicine. This includes the Precision Medicine Initiative (Collins and Varmus, 2015) launched by President Barack Obama in 2015, now known as the *All of Us* Research Program.

A fundamental difficulty with putting the findings of precision medicine into practice is that–by its very nature–precision medicine creates a very large space of treatment options (Frey et al., 2016). These can easily overwhelm clinicians attempting to stay up-to-date with the latest findings, and can easily inhibit a clinician's attempts to determine the best possible treatment for a particular patient. However, the ability to quickly locate relevant evidence is the hallmark of information retrieval (IR).

For three consecutive years the TREC Clinical Decision Support (CDS) track sought to evaluate IR

systems that provide medical evidence at the point-of-care. The TREC Precision Medicine track, then, was launched to specialize the CDS track to the needs of precision medicine so IR systems can focus on this important issue. The Precision Medicine track has focused on a single field, oncology, for a specific use case, genetic mutations of cancer. This started with the TREC 2017 Precision Medicine track, continued in 2018, and further continues with the 2019 track described here. As described above, main idea behind precision medicine is to use detailed patient information (largely genomic information in most current research) to identify the most effective treatments. Improving patient care in precision oncology then requires both (a) a mechanism to locate the latest research relevant to a patient, and (b) a fallback mechanism to locate the most relevant clinical trials when the latest techniques prove ineffective for a patient. In the first part, the track continues the previous Clinical Decision Support track (with a more focused use case), while in the second part expands the task to cover a new type of data (clinical trial descriptions). The main change between the 2017-2018 tracks and the 2019 track was to add the optional sub-task of determining the actual treatments described in literature articles (no changes were made for trials, where this data is more clearly available in semi-structured form). The idea behind this addition was to allow for an aspect-based retrieval approach, where results can be grouped by the actual treatments described for easier presentation for oncologists.

The remainder of this overview is organized as follows: Section 2 describes the historical context of medical IR evaluation that led to the Precision Medicine track; Section 3 describes the structure of the topics and the process of creating them; Section 4 outlines the retrieval tasks; Section 5 describes the evaluation method; finally, Section 6 details the results of the participant systems.

2 Background

The TREC Precision Medicine track continues a long tradition of biomedical retrieval evaluations within TREC. This started with the 2003-2007 TREC Genomics (Hersh and Voorhees, 2009) tracks, intended to connect genomics researchers to relevant biomedical literature. This was followed by the 2011 and 2012 TREC Medical Records tracks (Voorhees and Hersh, 2012), focusing on retreiving cohorts of patients from electronic health records. The 2014-2016 TREC Clinical Decision Support (CDS) track (Roberts et al., 2015, 2016a,b) targeted giving clinicians access to evidence-based literature. Then, the TREC Precision Medicine track grew from the CDS track, starting in 2017 (Roberts et al., 2017) and continuing in 2018 (Roberts et al., 2018), focusing on a more narrow problem domain (precision oncology). The 2019 Precision Medicine track continues this effort.

3 Topics

The 2019 Precision Medicine track provided 40 topics. Due to the difficulty in obtaining actual patient data, the topics were synthetically created, though often inspired by actual patients, with modification.[1] Out of the 40 total topics, 30 were created by oncologists from the University of Texas MD Anderson Cancer Center. The other 10 topics, unrelated to cancer, were based on the American College of Medical Genetics and Genomics (ACMG) recommendations.[2] These topics were added to assess the relative difficulty of cancer search versus other disciplines requiring precision medicine.

The topics contain three key elements in a semi-structured format to reduce the need to perform natural language processing to identify the key elements. The three key elements are: (1) disease (e.g., type of cancer), (2) genetic variants (primarily the genetic variants in the tumors themselves as opposed to the patient's DNA), and (3) demographic information (e.g., age, sex), Four topics from the track are shown in Table 1. Note that the final example in Table 1 is one of the non-cancer topics. An additional four topics are shown in their corresponding XML format (i.e., what was provided to the participants) in Table 2.

[1] Note that while clinical data is frequently de-identified for research purposes without the need for patient permission, genomic data is fundamentally difficult to de-identify. So to be safe, synthetic data was used.

[2] https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/acmg.shtml

Disease: melanoma **Variant:** BRAF (E586K) **Demographic:** 64-year-old female
Disease: gastric cancer **Variant:** ERBB2 amplification **Demographic:** 64-year-old male
Disease: gastrointestinal stromal tumor **Variant:** KIT (V654A) **Demographic:** 70-year-old male
Disease: Brugada syndrome **Variant:** SCN5A **Demographic:** 26-year-old female

Table 1: Example topics from the 2019 track.

```
<topic number="29">
  <disease>chondrosarcoma<disease>
  <gene>IDH1<gene>
  <demographic>26-year-old female<demographic>
<topic>
<topic number="30">
  <disease>endometrial cancer<disease>
  <gene>PIK3R1<gene>
  <demographic>58-year-old female<demographic>
</topic>
<topic number="31">
  <disease>aortic aneurysm<disease>
  <gene>ACTA2<gene>
  <demographic>42-year-old male<demographic>
</topic>
<topic number="32">
  <disease>Loeys-Dietz syndrome<disease>
  <gene>TGFBR2<gene>
  <demographic>42-year-old male<demographic>
</topic>
```

Table 2: XML format for two topics from the 2019 track.

4 Tasks

The two tasks in the Precision Medicine track correspond to two different corpora, each with different goals (underlined):

1. **Literature Articles**. Because precision medicine is a fast-moving field, keeping up-to-date with the latest literature can be challenging due to both the volume and velocity of scientific advances. Therefore, when treating patients, it would be helpful to present the most relevant scientific articles for an individual patient. The primary literature corpus is therefore a snapshot of MEDLINE abstracts (i.e., what is searchable through the PubMed interface). Relevant literature articles can guide precision oncologists to the best-known treatment options for the patient's condition. The treatment options are represented simply as the article abstract, participants do not need to provide a specific treatment name, simply an article describing a potential treatment. An updated snapshot was used for the 2019 track (both the 2017 and 2018 tracks used the same snapshot of 26.76 million abstracts). Specifically, this corpus is composed of 29,138,916 MEDLINE abstracts. Unlike 2017 and 2018, the ASCO and AACR abstracts were not included.

    - **Treatment Sub-Task**: Participants were given the option of also providing up to three treatments described in the article. In order to be valid, a treatment must be evaluated within the

scope of the article (e.g., a treatment that is mentioned but whose efficacy is not evaluated is not relevant). For simplicity, treatments were provided simply as strings (as opposed to, say, UMLS CUIs). The basic idea behind this sub-task is to allow search engines to organize results in a treatment-oriented manner, as opposed to a simple ranked list of articles. For instance, a search engine that returns results for two separate treatment options may be more useful than a search engine returning only articles relating to the most well-known treatment.

2. **Clinical Trials**. In many oncology patients, no approved treatment is available (or, commonly, none of the available treatments have proven effective). The common recourse in this case is to determine if any potential treatments are undergoing evaluation in a clinical trial. Therefore, in such situations, it would be helpful to automatically identify the most relevant clinical trials for an individual patient. Precision oncology trials typically use a certain treatment (e.g., a form of chemotherapy or radiation) for a certain disease with a specific genetic variant (or set of variants). Such trials can have complex inclusion and/or exclusion criteria that are challenging to match with automated systems (Weng et al., 2011). The corpus is derived from ClinicalTrials.gov, a repository of past, present, and future clinical trials in the U.S. and abroad. A total of 241,006 clinical trial descriptions compose the corpus provided to participants. Note that for the purposes of this track, the state of the trial (e.g., recruiting, active, completed) and geographic location constraints are not considered.

5 Evaluation

The evaluation followed standard TREC evaluation procedures for ad hoc retrieval tasks. Participants submitted (in `trec_eval` format) a maximum of five automatic or manual runs per task, each consisting of a ranked list of up to 1,000 literature article IDs and 1,000 ClinicalTrials.gov Identifiers per topic. That is, up to 10 total runs: a maximum of 5 literature runs and 5 clinical trial runs per topic. For teams participating in the optional treatment track, these could be provided as optional strings at the end of the line.

The highest ranked articles and trials for each topic were pooled and judged by physician graduate students at OHSU and other biomedical subject matter experts.

As in the 2017 and 2018 Precision Medicine tracks, the assessment process was two-tiered: first a manual assessment was made by the human assessors based on several categories for each result (referred to here as *Result Assessment*), then a relevance score was assigned to the result based on its categorization (referred to here as *Relevance Assessment*). Treatment evaluation primarily occurred during the result assessment step, where assessors were provided with a checkbox for each treatment string provided by the participants.

5.1 Result Assessment

Result assessment can be viewed as a set of multi-class annotations. Judging an individual result, whether an article or trial, proceeds in a cascaded manner with two steps: an initial pass ensures the article/trial is broadly relevant to precision medicine, after which the assessor categorizes the article/trial according to the three fields above.

See Figure 1 for a flow chart style overview of this process. The first step is designed to save assessor time by filtering out unrelated articles/trials, since the second step can be more time-consuming (possibly requiring a more detailed reading of the article/trial). The assessors were free to quickly skim the article/trial in order to make the initial decision. Then, if the article/trial is relevant to precision medicine (by the standard outlined below), a more detailed reading may be necessary in order to accurately assess all fields.

The first step is to determine whether the article/trial is related to precision medicine. There are three options:

- **Human PM**: The article/trial (1) relates to humans, (2) involves some form of cancer, (3) focuses on treatment, prevention, or prognosis of cancer, and (4) relates in some way to at least one of the genes in the topic.
- **Animal PM**: Identical to Human PM requirements (2)-(4), except for animal research.
- **Not PM**: Everything else. This includes "basic science" that focuses on understanding underlying genomic principles (e.g., pathways), but provides no evidence for treatment.

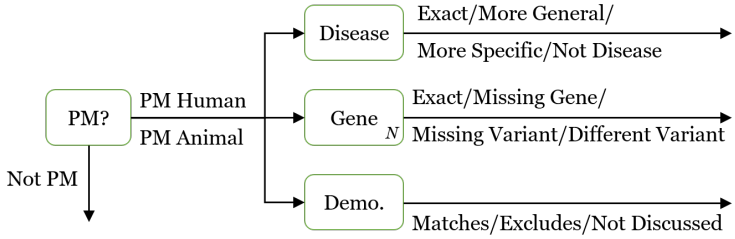

Figure 1: Two-step result assessment process

The second step is to determine the appropriate categorization for each of the three fields:

1. *Disease*:
   - **Exact**: The form of cancer in the article/trial is identical to the one in the topic.
   - **More General**: The form of cancer in the article/trial is more general than the one in the topic (e.g., blood cancer vs. leukemia).
   - **More Specific**: The form of cancer in the article/trial is more specific than the one in the topic (e.g., squamous cell lung carcinoma vs. lung cancer).
   - **Not Disease**: The article/trial is not about a disease, or is about a different disease (or type of cancer) than the one in the topic.

2. *Gene* [for each particular gene in the topic]
   - **Exact**: The article/trial focuses on the exact gene and variant as the one in the topic. If the topic does not contain a specific variant, then this holds as long as the gene is included. By "focus" this means the gene/variant needs to be part of the scientific experiment of the article/trial, as opposed to discussing related work.
   - **Missing Gene**: The article/trial does not focus the particular gene in the topic. If the gene is referenced but not part of the study, then it is considered missing.
   - **Missing Variant**: The article/trial focuses on the particular gene in the topic, but not the particular variant in the topic. If no variant is provided in the topic, this category should not be assigned.
   - **Different Variant**: The article/trial focuses on the particular gene in the topic, but on a different variant than the one in the topic.

3. *Demographic*
   - **Matches**: The article/trial demographic population matches the one in the topic.
   - **Excludes**: The article/trial demographic population specifically excludes the one in the topic.
   - **Not Discussed**: The article/trial does not discuss a particular demographic population.

Note that in the 2017 track, an "Other" field was used as well. This was dropped for 2018 and 2019 because several oncology experts felt it is not a major part of precision medicine decision-making.

5.2 Relevance Assessment

Relevance assessment is defined here as the process of mapping the multi-class result assessments described above onto a single numeric relevance scale. This allows for the computation of evaluation metrics (e.g., infNDCG, R-prec, P@10) as well as the tuning of IR systems to improve their search ranking. As already demonstrated by the need for result assessment above, for the Precision Medicine track the notion of relevance assessment becomes more complex than previous tracks.

One of the factors that makes precision medicine a difficult domain for IR is that different patient cases require different types of flexibility on the above categories. For some patients, the exact type of cancer is not relevant. Other times, the patient's demographics factors might weigh more heavily. Most notably, the very concept of precision medicine acknowledges the uniqueness of the patient, and so it is to be expected

that no perfect match is found. Not only do the topics provided to the participants not contain the necessary information to decide what factors are more/less relevant (e.g., the patient's previous treatments), in many ways it isn't realistic to assign the IR system this responsibility. Precision medicine requires a significant amount of oversight by clinicians, including the ability to consider multiple treatment options. So it might ultimately make the most sense to allow the relevance assessment to be, at least in part, designed by the clinician to allow the IR system to adjust its rankings to suit. Given the constraints of an IR shared task, however, it is necessary to define a relevance assessment process. As such, a fairly broad notion of relevance based on the above categories was used:

1. **Definitely Relevant**: The result should: be either *Human PM* or *Animal PM*; have a *Disease* assignment of *Exact* or *More Specific*; have at least one *Gene* is *Exact*; the *Demographic* is either *Exact* or *Not Discussed*.
2. **Partially Relevant**: Largely the same as *Definitely Relevant*, but with the exception that *Disease* can also be *More General* and *Gene* can also be *Missing Variant* or *Different Variant*.
3. **Not Relevant**: Neither of the above.

The primary evaluation metrics are precision at rank 10 (P@10), inferred normalized discounted cumulative gain (infNDCG), and R-precision (R-prec). For infNDCG, *Definitely Relevant* has a score of 2, *Partially Relevant* is 1, and *Not Relevant* is 0. In 2017, clinical trials were pooled using a different sampling strategy than literature articles, and therefore had different primary evaluation metrics (P@5, P@10, P@15). However, starting in the 2018 track and continuing into 2019 the same sampling strategy was used for both tasks and therefore the same primary evaluation metrics apply.

For treatment evaluation, two metrics were used: **Treatment Recall@*N*** and **Treatment F1@*N***. For both of these, the unique, valid treatments returned in the top N results were considered relative to the total number of validated treatments from all pooled runs. This intentionally penalizes results that only return a handful of treatments in the top results. This penalty is inspired by the many knowledge-based approaches to this task in the past, where one way to improve overall retrieval results is to identify a priori the standard treatments for a given situation (e.g., breast cancer patient with HER2 mutation), then focus on retrieving articles with this treatment (e.g., using query expansion). This situation is not necessarily ideal for oncologists, who often must consider multiple potential treatments. Thus, systems that present a diversity of treatment options to the user may be preferable over a standard relevance-ranked list.

6 Results

In total, there were 22,429 judgments for the literature articles and 14,188 judgments for the clinical trials. Table 3 shows basic statistics of the results and relevance assessments. Table 4 shows the number of Definitely Relevant, Partially Relevant, and Not Relevant judgments for each topic. Since each result was judged only once, no inter-rater agreement is available for the judgments.

There were a total of 15 participants in the track. For the literature articles, 14 participants submitted 62 runs. For the clinical trials, 12 participants submitted 53 runs. Only 3 teams participated in the treatment sub-task. See Table 5 for a list of the participants and numbers of runs. Table 6 shows the top 10 runs (top run per participant) for each metric on each corpus. Figures 2 and 3 show box-and-whisker plots for the top 10 runs. Table 7 shows the treatment sub-task results. Finally, Table 8 shows an aggregate view of the performance on the cancer versus non-cancer topics.

7 Conclusion

The goal of the Precison Medicine track is to inform the creation of information retrieval systems to support clinicians working in precision medicine (specifically oncologists in this track) in making better treatment decisions for individual patients. Participants were provided with synthetic patient data consisting of a type of cancer, one or more genetic variants, and patient demographics. Given this, participants were challenged with retrieving relevant treatments (in the form of literature articles) and relevant trials (in the form of clinical trial descriptions) for the specific patient.

		Literature Articles					Clinical Trials				
Type	Class	Total	Mean	Median	Min	Max	Total	Mean	Median	Min	Max
PM	Human PM	8,775	219	203	11	491	5,713	143	133	2	403
	Animal PM	440	11	7	0	57	1	0	0	0	1
	Not PM	9,101	228	234	24	486	8,109	203	204	24	460
Disease	Exact	5,171	129	82	6	425	1,451	36	12	0	234
	More Specific	786	20	10	0	127	636	16	3	0	200
	More General	1,326	33	14	0	366	1,403	35	18	0	240
	Not Disease	1,932	48	26	2	186	2,224	56	49	0	181
1st Gene	Exact	4,186	105	70	0	353	1,605	40	14	0	187
	Missing Variant	2,100	53	1	0	307	1,408	35	1	0	225
	Different Variant	716	18	4	0	157	339	8	0	0	128
	Missing Gene	2,213	55	22	0	278	2,362	59	38	0	297
2nd Gene	Exact	68	2	0	0	64	7	0	0	0	7
	Missing Variant	529	13	0	0	431	340	9	0	0	268
	Different Variant	34	1	0	0	20	6	0	0	0	5
	Missing Gene	85	2	0	0	67	103	3	0	0	102
Demographics	Matches	591	15	6	0	123	4,983	125	87	1	398
	Not Discussed	8,124	203	167	10	489	389	10	1	0	237
	Excludes	500	13	9	0	58	342	9	5	0	75
Relevance	Definitely Relevant	2,571	64	37	0	279	485	12	4	0	125
	Partially Relevant	2,973	74	34	1	385	1,700	43	43	0	318
	Not Relevant	12,772	319	308	101	548	11,637	291	291	82	462

Table 3: Descriptive statistics (per-topic) of manual judgments (both results assessment and relevance assessment) for both literature articles and clinical trials. Note: only 3 topics had a 2nd Gene, but means are still provided across 50 topics.

Acknowledgments

The organizers would like to thank Kate Fultz Hollis for managing the assessment process. KR is supported by the National Institutes of Health (NIH) grant 2R00LM012104-02 and the Cancer Prevention and Research Institute of Texas (CPRIT) grant RP170668. DDF is supported by the Intramural Research Program of the U.S. National Library of Medicine, NIH. Finally, the organizers are grateful to the National Institute of Standards and Technology (NIST) for funding the assessment process.

	literature articles			clinical trials		
Topic	DR	PR	NR	DR	PR	NR
1	1	352	218	0	138	132
2	10	344	127	0	214	103
3	16	2	467	8	2	401
4	70	153	295	14	75	352
5	0	112	425	0	52	393
6	0	385	349	1	318	158
7	230	192	101	125	167	82
8	9	61	265	0	23	317
9	131	111	221	60	9	232
10	4	79	293	0	35	296
11	40	100	247	6	39	199
12	3	30	332	0	9	445
13	2	71	366	0	40	271
14	49	3	548	12	1	356
15	158	188	204	17	166	149
16	179	94	126	6	36	238
17	41	37	295	6	12	241
18	2	112	429	0	66	314
19	59	10	506	11	24	450
20	38	11	349	6	14	274

	literature articles			clinical trials		
Topic	DR	PR	NR	DR	PR	NR
21	121	55	353	68	10	174
22	109	27	305	26	34	251
23	23	2	414	6	2	453
24	1	38	503	0	32	269
25	233	20	148	20	13	245
26	35	23	286	4	3	306
27	22	160	267	3	89	201
28	279	41	245	57	52	152
29	13	21	376	1	7	263
30	26	4	512	8	7	362
31	8	8	395	1	1	347
32	20	5	299	0	0	462
33	3	2	431	0	0	364
34	65	6	311	1	1	349
35	142	68	221	7	4	225
36	85	13	347	5	1	309
37	130	20	223	4	1	312
38	63	1	318	0	1	424
39	144	11	292	1	1	349
40	7	1	363	0	2	417

Table 4: Counts of Definitely Relevant (DR), Partially Relevant (PR), and Not Relevant (NR) results for each topic.

References

Collins, F. S. and Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372:793–795.

Frey, L. J., Bernstam, E. V., and Denny, J. C. (2016). Precision medicine informatics. *Journal of the American Medical Informatics Association*, 23:668–670.

Hersh, W. and Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval*, 12:1–15.

Roberts, K., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016a). Overview of the TREC 2016 Clinical Decision Support Track. In *Proceedings of the Twenty-Fifth Text Retrieval Conference*.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., and Lazar, A. (2018). Overview of the TREC 2018 Precision Medicine Track. In *Proceedings of the Twenty-Seventh Text Retrieval Conference*.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A., and Pant, S. (2017). Overview of the TREC 2017 Precision Medicine Track. In *Proceedings of the Twenty-Sixth Text Retrieval Conference*.

Roberts, K., Simpson, M. S., Demner-Fushman, D., Voorhees, E., and Hersh, W. R. (2016b). State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS Track. *Information Retrieval*, 19(1).

Roberts, K., Simpson, M. S., Voorhees, E., and Hersh, W. (2015). Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of the Twenty-Fourth Text Retrieval Conference*.

Voorhees, E. M. and Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. In *Proceedings of the Twenty-First Text REtrieval Conference*.

Weng, C., Wu, X., Luo, Z., Boland, M. R., Theodoratos, D., and Johnson, S. B. (2011). EliXR: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association*, 18(Suppl 1):i116–i124.

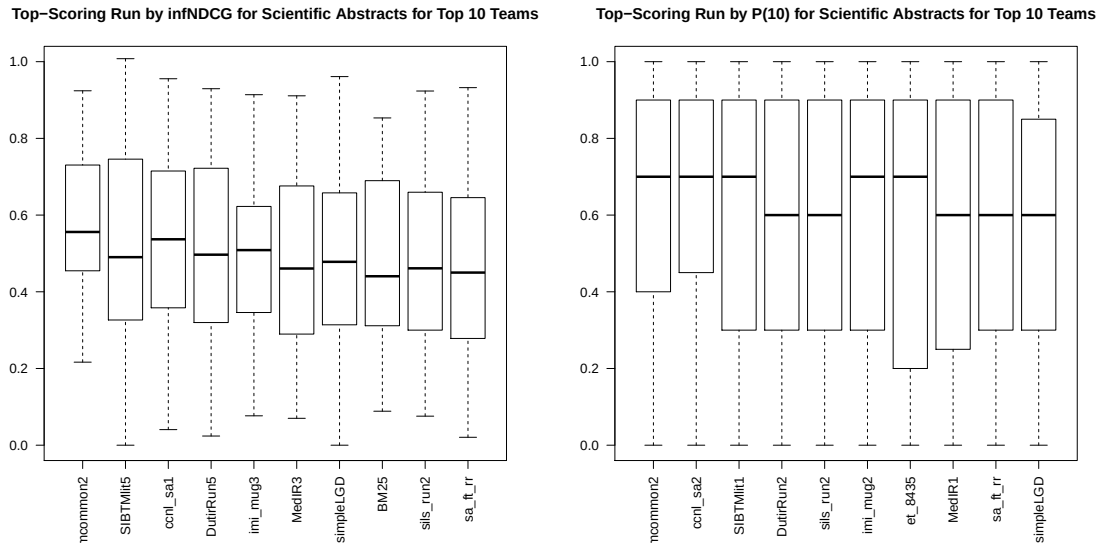

Figure 2: Top-performing runs (showing only best run per participant) on literature articles.

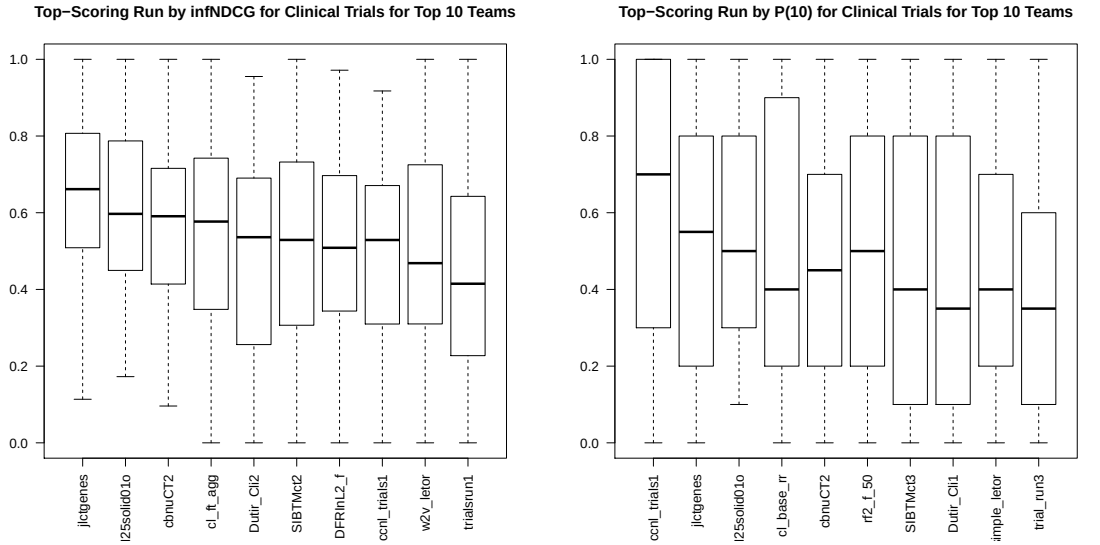

Figure 3: Top-performing runs (showing only best run per participant) on clinical trials.

Team ID	Affiliation	# Runs	
		Articles	Trials
BITEM_PM	BITEM SIB Text Mining Group	5*	5
Brown	Brown University	-	5
cbnu	Chonbuk National University	4	4
CCNL	Communication & Computer Network Lab of Guangdong	5	2
CincyMedIR	University of Cincinnati	5	-
CSIROmed	Commonwealth Science and Industrial Research Organisation	4	5
DUTIR	Dalian University of Technology	5	2
ECNU-ICA	East China Normal University - ICA	4	5
imi_mug	Medical University of Graz	5*	-
ims_unipd	University of Padua	5	5
julie-mug	Friedrich-Schiller-University Jena	5*	5
POZNAN	Poznan University of Technology	4	5
UNC_SILS	University of North Carolina at Chapel Hill	4	-
UNIVAQ	University of L'Aquila	2	5
WCMC	Weill Cornell Medicine	5	5
Total		62	53

Table 5: Participating teams and submitted runs. *Teams participating in treatment sub-task.

Literature Articles

infNDCG

Team	Run	Score
julie-mug	jlpmcommon2	0.5783
BITEM_PM	SIBTMlit5	0.5339
CCNL	ccnl_sa1	0.5309
DUTIR	DutirRun5	0.5108
imi_mug	imi_mug3	0.4812
CincyMedIR	MedIR3	0.4801
POZNAN	SAsimpleLGD	0.4755
ims_unipd	BM25	0.4747
UNC_SILS	sils_run2	0.4692
ECNU-ICA	sa_ft_rr	0.4672

R-prec

Team	Run	Score
julie-mug	jlpmcommon2	0.3572
DUTIR	DutirRun2	0.3273
BITEM_PM	SIBTMlit2	0.3166
imi_mug	imi_mug2	0.3122
CincyMedIR	MedIR3	0.3111
POZNAN	SAsimpleLGD	0.3092
CCNL	ccnl_sa2	0.3066
CSIROmed	bm25_6801	0.3029
ims_unipd	BM25neopgngm	0.2999
UNC_SILS	sils_run1	0.2858

P @ 10

Team	Run	Score
julie-mug	jlpmcommon2	0.6525
CCNL	ccnl_sa2	0.6500
BITEM_PM	SIBTMlit1	0.6275
DUTIR	DutirRun2	0.5975
UNC_SILS	sils_run2	0.5925
imi_mug	imi_mug2	0.5750
CSIROmed	et_8435	0.5725
CincyMedIR	MedIR1	0.5675
ECNU-ICA	sa_ft_rr	0.5675
POZNAN	SAsimpleLGD	0.5400

Clinical Trials

infNDCG

Team	Run	Score
julie-mug	jlctgenes	0.6451
ims_unipd	BM25solid01o	0.6239
cbnu	cbnuCT2	0.5568
ECNU-ICA	cl_ft_agg	0.5355
DUTIR	Dutir_Cli2	0.5038
BITEM_PM	SIBTMct2	0.4963
CSIROmed	DFRInL2_f	0.4930
CCNL	ccnl_trials1	0.4862
POZNAN	w2v_letor	0.4810
WCMC	trialsrun1	0.4320

R-prec

Team	Run	Score
julie-mug	jlctprec	0.4820
ims_unipd	BM25solid01o	0.4386
cbnu	cbnuCT2	0.4121
ECNU-ICA	cl_base_rr	0.4001
BITEM_PM	SIBTMct4	0.3698
CSIROmed	bm25_ct_f_61	0.3586
POZNAN	simple_letor	0.3503
DUTIR	Dutir_Cli1	0.3453
CCNL	ccnl_trials2	0.3440
WCMC	trialsrun1	0.3230

P @ 10

Team	Run	Score
CCNL	ccnl_trials1	0.5947
julie-mug	jlctphrase	0.5474
ims_unipd	BM25solid01o	0.5368
ECNU-ICA	cl_base_rr	0.5053
CSIROmed	rf2_f_50	0.4921
cbnu	cbnuCT2	0.4921
BITEM_PM	SIBTMct3	0.4711
DUTIR	Dutir_Cli1	0.4579
POZNAN	w2v_noletor	0.4421
WCMC	trial_run3	0.3658

Table 6: Top overall systems (best run per participant).

Team ID	Run ID	Recall @ 10	F1 @ 10	Recall @ 25	F1 @ 25
julie-mug	jlpmtrboost	0.2857	0.3118	0.4603	0.3793
julie-mug	jlpmtrcommon	0.2698	0.3019	0.4469	0.3716
BITEM_PM	SIBTMlit3	0.2412	0.2815	0.3696	0.3193
BITEM_PM	SIBTMlit2	0.1533	0.1985	0.2329	0.2632
BITEM_PM	SIBTMlit4	0.1533	0.1985	0.2329	0.2632
BITEM_PM	SIBTMlit5	0.1504	0.1911	0.2462	0.2638
imi_mug	imi_mug3_t	0.1382	0.1610	0.2330	0.2275
imi_mug	imi_mug2_t	0.1267	0.1472	0.2261	0.2194
BITEM_PM	SIBTMlit1	0.0980	0.1312	0.1517	0.1914

Table 7: Treatment results for the 9 participating runs on literature articles.

		cancer	non-cancer
infNDCG	median	0.4417	0.3869
	mean	0.4458	0.4253
R-prec	median	0.2667	0.2500
	mean	0.2863	0.2850
P@10	median	0.6000	0.2000
	mean	0.5333	0.2783

Table 8: Comparison of average per-topic, per-run scores between cancer topics (#1-30) and non-cancer topics (#31-40).