Overview of the TREC 2018 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

1 Introduction

The fundamental philosophy behind *precision medicine* is that for many complex diseases, there is 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 former 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 huge 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 and continued with the 2018 track described here. As described above, the 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). No substantial changes to the 2017 track

Disease: melanoma **Variant:** BRAF (V600E) **Demographic:** 64-year-old male
Disease: melanoma **Variant:** high serum LDH levels **Demographic:** 69-year-old female
Disease: medullary thyroid carcinoma **Variant:** RET **Demographic:** 45-year-old female
Disease: anaplastic large cell lymphoma **Variant:** ALK **Demographic:** 18-year-old male

Table 1: Example topics from the 2018 track.

were made for 2018 (with the exception of new, and an increased number of, topics). Since 2017 was the first year of the track, keeping minimal changes was an intentional choice to allow participants to develop new methods using the results of the prior year as a guide (2017 participants had no gold standard to utilize for system development).

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) (Roberts et al., 2016) track targeted giving clinicians access to evidence-based literature. Then, starting in 2017, the TREC Precision Medicine (Roberts et al., 2017) track grew from the CDS track, focusing on a more narrow problem domain (precision oncology). The 2018 Precision Medicine track continues this effort.

3 Topics

The 2018 Precision Medicine track provided 50 topics created by oncologists from and resources provided by the University of Texas MD Anderson Cancer Center. Due to the difficulty in obtaining actual patient data, the topics were synthetically created, though often inspired by actual patients, with modification.[1]

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. The first two topics are additionally shown in their corresponding XML format (i.e., what was provided to the participants) in Table 2. Note that the second example in Table 1 is actually an immunotherapy marker, not a tumor genetic variant. Six of the 50 topics for 2018 were focused on immunotherapy.

[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.

```
<topic number="1">
  <disease>melanoma<disease>
  <gene>BRAF (V600E)<gene>
  <demographic>64-year-old male<demographic>
<topic>
<topic number="2">
  <disease>anaplastic large cell lymphoma<disease>
  <gene>ALK<gene>
  <demographic>18-year-old male<demographic>
</topic>
```

Table 2: XML format for two topics from Table 1.

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. The same corpus as the 2017 track was utilized. Specifically, this corpus is composed of approximately 26.8 million MEDLINE abstracts and is supplemented with two additional sets of abstracts: (i) 37,007 abstracts from recent proceedings of the American Society of Clinical Oncology (ASCO), and (ii) 33,018 abstracts from recent proceedings of the American Association for Cancer Research (AACR). These additional datasets were added to increase the set of potentially relevant treatment information. Notably, the latest research is often presented at conferences such as ASCO and AACR prior to submission to journals (thus these proceedings may represent a more up-to-date snapshot of scientific knowledge than MEDLINE).

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 identifiers (PMIDs) 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.

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

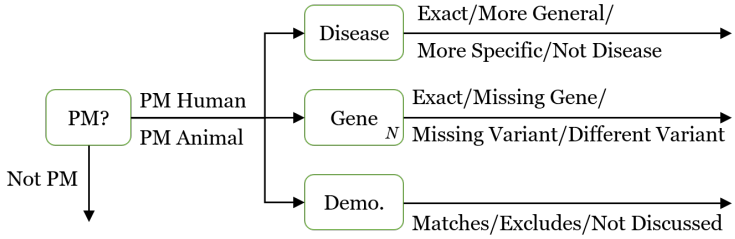

Figure 1: Two-step result assessment process

As in the 2017 Precision Medicine track, 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*).

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.

Step 1 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.

Step 2 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 because several oncology experts felt it is not a major part of precision medicine decision-making. Additionally, the same assessment tool was used in 2018 as in 2017, which included the Other field. Unfortunately, some assessors on the rare occasion marked Excludes for some results despite the lack of a criteria in the Other field to exclude. This resulted in a small number of downstream results being considered Not Relevant in the official results, but the impact of this issue was fairly minor on the overall scores.

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., P@10, infNDCG) 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 demographic 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 fixed relevance assessment process. As such, a fairly broad notion of relevance based on the above categories was used:

		Literature Articles					Clinical Trials				
Type	Class	Total	Mean	Median	Min	Max	Total	Mean	Median	Min	Max
PM	Human PM	8,634	173	151	45	492	5,809	116	114	8	446
	Animal PM	590	12	7	0	102	5	0	0	0	1
	Not PM	13,205	264	269	38	546	8,374	167	152	0	437
Disease	Exact	5,158	103	75	10	338	2,189	44	20	0	226
	More Specific	1,915	38	26	0	192	1,233	25	9	0	131
	More General	686	14	7	1	80	896	18	9	0	127
	Not Disease	1,455	29	21	0	108	1,496	30	21	0	106
1st Gene	Exact	4,927	99	85	1	370	2,150	43	28	0	173
	Missing Variant	1,278	26	0	0	165	595	12	0	0	103
	Different Variant	917	18	2	0	197	524	10	1	0	110
	Missing Gene	2,102	42	20	0	230	2,545	51	21	0	354
2nd Gene	Exact	125	3	0	0	89	24	0	0	0	204
	Missing Variant	88	2	0	0	87	39	1	0	0	38
	Different Variant	16	0	0	0	14	0	0	0	0	0
	Missing Gene	482	10	0	0	252	357	7	0	0	204
Demographics	Matches	855	17	5	0	104	5,142	103	91	0	437
	Not Discussed	7,691	154	119	19	428	296	6	0	0	111
	Excludes	678	14	8	0	63	376	8	1	0	114
Relevance	Definitely Relevant	3,442	69	45	1	305	873	17	12	1	131
	Partially Relevant	2,146	43	20	1	277	1,174	23	14	1	118
	Not Relevant	16,841	337	341	119	566	12,141	243	219	73	441

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.

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, for the 2018 track the same sampling strategy was used for both tasks and therefore the same primary evaluation metrics apply.

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 27 participants in the track. For the literature articles, 24 participants submitted 103 runs. For the clinical trials, 21 participants submitted 90 runs. 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.

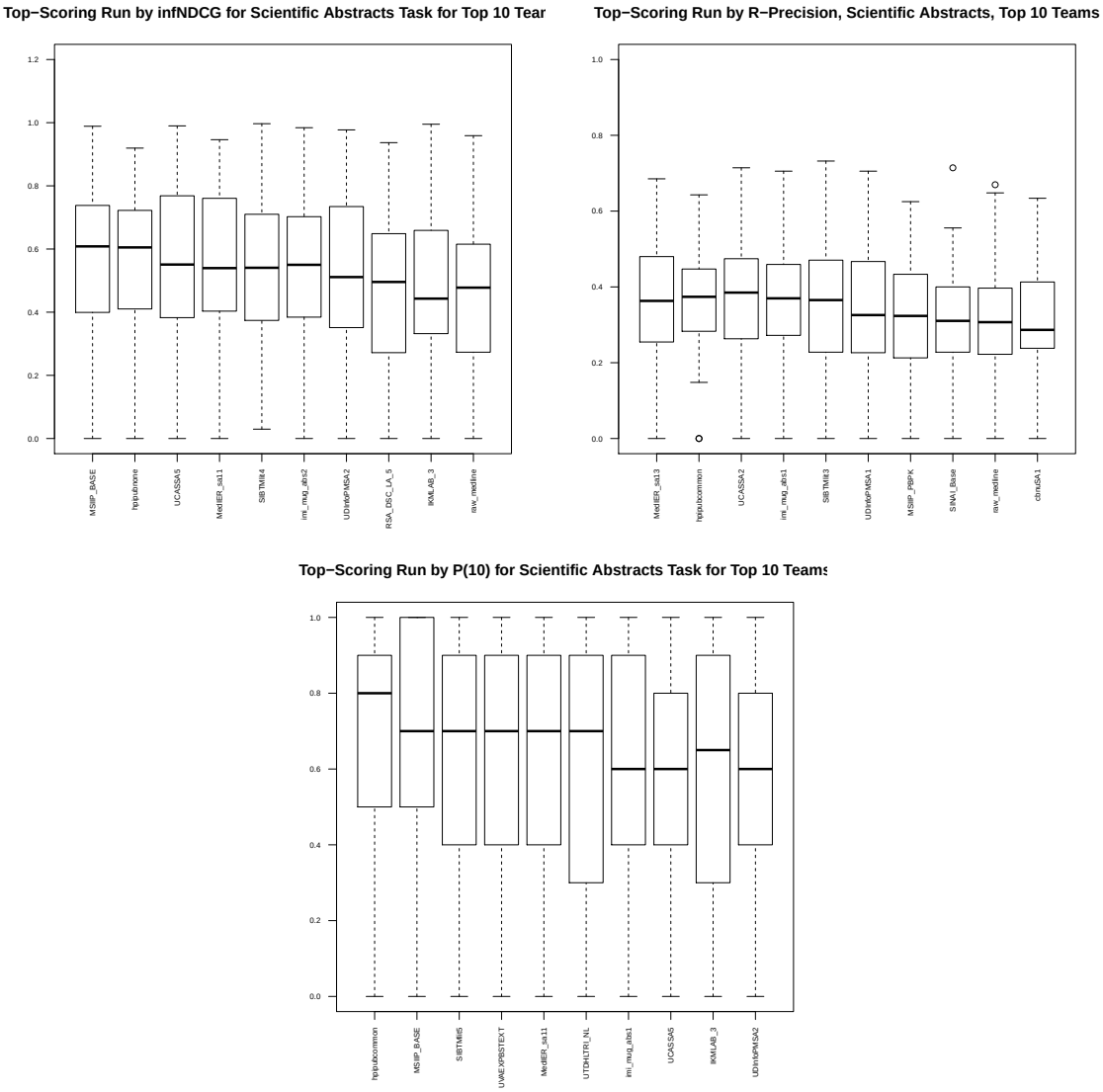

Figure 2: Top-performing runs (showing only best run per participant) on literature articles.

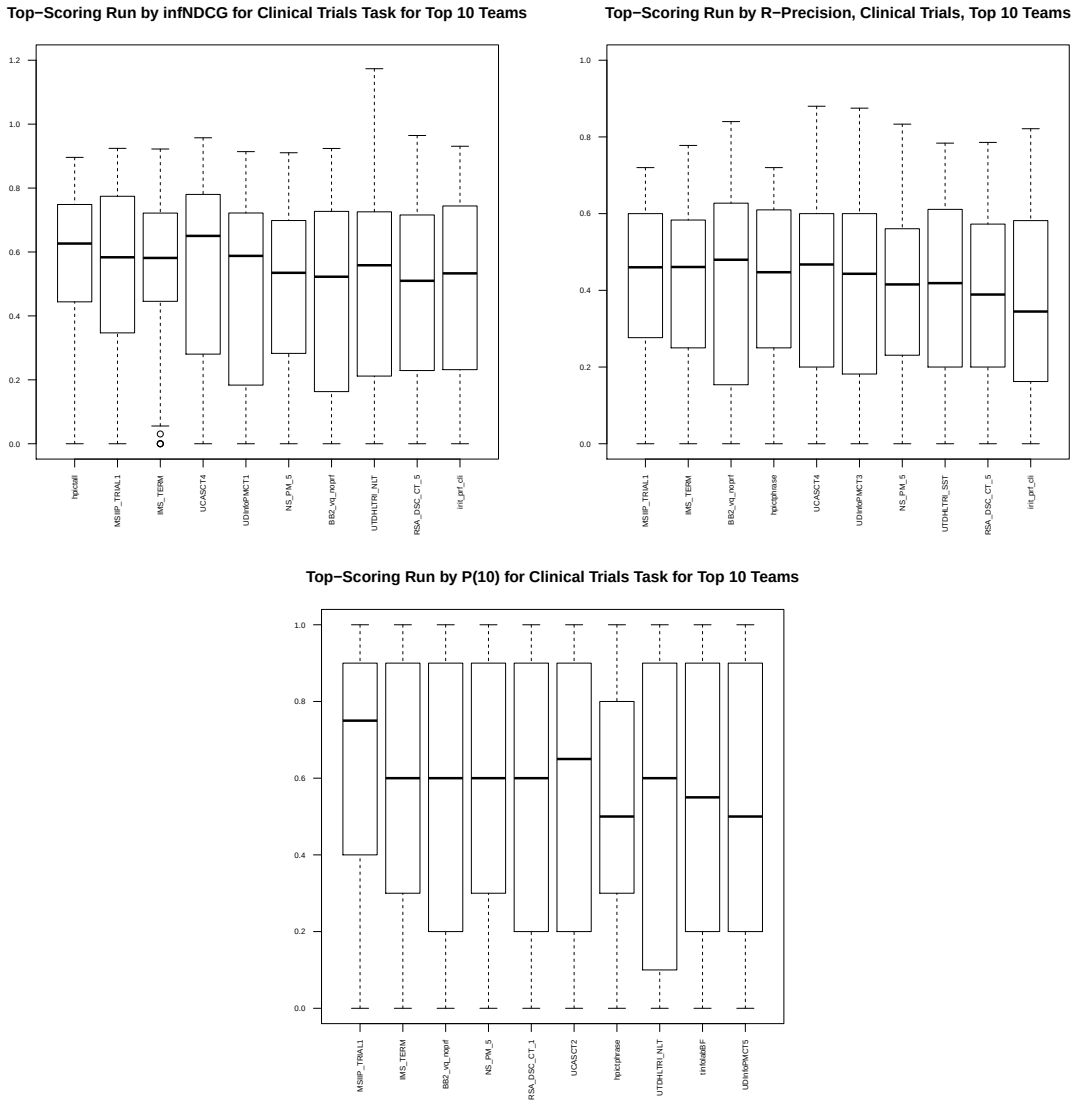

Figure 3: Top-performing runs (showing only best run per participant) on clinical trials.

Topic	literature articles			clinical trials		
	DR	PR	NR	DR	PR	NR
1	108	61	252	50	60	97
2	75	180	119	19	107	73
3	41	168	188	45	78	95
4	6	277	153	1	46	189
5	216	25	192	96	24	104
6	163	83	127	46	73	86
7	10	134	358	0	118	118
8	43	89	227	14	13	160
9	10	99	259	2	14	175
10	9	82	240	0	25	206
11	23	31	245	5	18	187
12	2	61	300	0	24	201
13	0	96	277	2	25	207
14	30	101	250	9	21	201
15	5	22	374	0	15	299
16	3	3	433	1	0	288
17	177	9	341	10	16	364
18	7	40	489	0	33	368
19	94	15	410	27	5	373
20	14	30	538	0	4	441
21	128	138	326	2	64	285
22	21	45	566	16	27	323
23	9	113	266	0	26	294
24	1	1	565	0	1	385
25	19	7	461	2	2	351

Topic	literature articles			clinical trials		
	DR	PR	NR	DR	PR	NR
26	31	2	366	37	0	183
27	163	11	296	12	13	129
28	111	1	227	6	3	188
29	88	23	353	40	51	252
30	98	3	247	19	3	214
31	61	2	450	7	4	358
32	155	3	406	48	7	288
33	151	25	421	38	36	294
34	61	8	378	3	2	281
35	20	9	492	2	2	368
36	57	5	482	27	10	331
37	119	16	438	51	14	381
38	33	21	291	2	9	162
39	83	12	337	6	11	171
40	305	6	317	131	6	173
41	6	0	372	2	12	203
42	15	15	348	4	1	375
43	74	1	325	8	2	241
44	24	39	355	0	106	224
45						
46	103	3	398	12	11	217
47	226	1	259	32	15	219
48	58	20	292	4	12	208
49	32	2	316	2	0	286
50	109	7	368	20	0	214

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

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. The 2018 track was the second year for the track. This year saw continued high participation numbers, as well as enabling participants to build on systems and results from 2017.

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.

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. (2016). 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., Lazar, A., and Pant, S. (2017). Overview of the TREC 2017 Precision Medicine Track. In *Proceedings of the Twenty-Sixth 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.

Team ID	Affiliation	# Runs	
		Articles	Trials
ASU_Biomedical	Arizona State University	3	0
Brown	Brown University	5	5
Cat_Garfield	Tsinghua-iFlytek Joint Laboratory	5	5
cbnu	Chonbuk National University	3	3
CSIROmed	Commonwealth Scientific and Industrial Research Organisation	3	3
ECNUica	East China Normal University	5	5
FDUDMIIP	School of Computer Science, Fudan University	5	5
hpi-dhc	Hasso Plattner Institute	5	5
IKMLAB	Institute of Medical Informatics of National Cheng Kung Univ.	5	5
imi_mug	Medical University of Graz	5	5
ims_unipd	Information Management Systems (IMS) Group	0	3
InfoLabPM	InfoLab, Faculty of Engineering, University of Porto	4	3
IRIT	Institut de Recherche en Informatique de Toulouse	0	1
KlickLabs	Klick Inc.	4	5
MayoNLPTeam	Mayo Clinic	4	3
MedIER	University of Michigan	5	0
NOVASearch	Universidade NOVA Lisboa	0	5
PM_IBI	Integrative Biomedical Informatics Group, Barcelona	3	0
Poznan	Poznan University of Technology	1	5
RSA_DSC	Research Studios Austria / Studio Data Science	5	5
SIBTextMining	SIB Text Mining Group (HES-SO)	5	4
SINAI	Universidad de Jaen	3	0
UCAS	University of Chinese Academy of Sciences	5	5
udel_fang	InfoLab at University of Delaware	5	5
UNTIIA	University of North Texas	5	0
UTDHLTRI	The University of Texas at Dallas	5	5
UVA_ART	University of Virginia Medical Center	5	0
Total		103	90

Table 5: Participating teams and submitted runs.

Literature Articles

infNDCG

Team	Run	Score
Cat_Garfield	MSIIP_BASE	0.5621
hpi-dhc	hpipubnone	0.5605
UCAS	UCASSA5	0.5580
MedIER	MedIER_sa13	0.5515
SIBTextMining	SIBTMlit4	0.5410
imi_mug	imi_mug_abs2	0.5391
udel_fang	UDInfoPMSA2	0.5081
RSA_DSC	RSA_DSC_LA_5	0.4855
UTDHLTRI	UTDHLTRI_NL	0.4797
IKMLAB	IKMLAB_3	0.4710

R-prec

Team	Run	Score
MedIER	MedIER_sa13	0.3684
hpi-dhc	hpipubcommon	0.3658
UCAS	UCASSA2	0.3654
imi_mug	imi_mug_abs1	0.3630
SIBTextMining	SIBTMlit3	0.3574
udel_fang	UDInfoPMSA1	0.3289
Cat_Garfield	MSIIP_PBPK	0.3257
SINAI	SINAI_Base	0.3082
FDUDMIIP	raw_medline	0.3072
cbnu	cbnuSA1	0.2992

P @ 10

Team	Run	Score
hpi-dhc	hpipubnone	0.7060
Cat_Garfield	MSIIP_BASE	0.6680
SIBTextMining	SIBTMlit5	0.6320
UVA_ART	UVAEXPBSTEXT	0.6260
MedIER	MedIER_sa11	0.6220
UTDHLTRI	UTDHLTRI_NL	0.6160
imi_mug	imi_mug_abs2	0.6000
UCAS	UCASSA5	0.5980
IKMLAB	IKMLAB_3	0.5960
udel_fang	UDInfoPMSA2	0.5800

Clinical Trials

infNDCG

Team	Run	Score
hpi-dhc	hpictall	0.5545
Cat_Garfield	MSIIP_TRIAL1	0.5503
ims_unipd	IMS_TERM	0.5395
UCAS	UCASCT4	0.5347
udel_fang	UDInfoPMCT1	0.5057
NOVASearch	NS_PM_5	0.4992
Poznan	BB2_vq_noprf	0.4894
UTDHLTRI	UTDHLTRI_NLT	0.4794
RSA_DSC	RSA_DSC_CT_5	0.4743
IRIT	irit_prf_cli	0.4736

R-prec

Team	Run	Score
Cat_Garfield	MSIIP_TRIAL1	0.4294
ims_unipd	IMS_TERM	0.4128
Poznan	BB2_vq_noprf	0.4101
hpi-dhc	hpictphrase	0.4081
UCAS	UCASCT4	0.4005
udel_fang	UDInfoPMCT3	0.3967
NOVASearch	NS_PM_5	0.3931
UTDHLTRI	UTDHLTRI_SST	0.3920
RSA_DSC	RSA_DSC_CT_5	0.3721
IRIT	irit_prf_cli	0.3658

P @ 10

Team	Run	Score
Cat_Garfield	MSIIP_TRIAL1	0.6260
ims_unipd	IMS_TERM	0.5660
Poznan	BB2_vq_noprf	0.5580
NOVASearch	NS_PM_5	0.5520
RSA_DSC	RSA_DSC_CT_3	0.5480
UCAS	UCASCT1	0.5460
hpi-dhc	hpictphrase	0.5400
UTDHLTRI	UTDHLTRI_NLT	0.5380
udel_fang	UDInfoPMCT5	0.5240
InfoLabPM	tinfolabBF	0.5240

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