Overview of the TREC 2020 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

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

1 Introduction

The precision medicine paradigm focuses on identifying treatments that are best suited to an individual patient's unique attributes. The reasoning behind this paradigm is that diseases do not uniformly manifest in people and thus "one size fits all" treatments are often not appropriate. For many diseases, such as cancer, proper selection of a treatment strategy can drastically improve results compared to the standard, frontline treatment. Generally speaking, the issues that are taken into consideration for precision medicine are the genomic, environmental, and lifestyle contexts of the patient. While precision medicine as a paradigm can be seen to broadly apply to medicine as a whole, the area where it has seen the most attention is cancer. Many cancer treatments may be lifesaving in one patient but deadly in another, primarily based on the genetic mutations of the patient's tumor. Different treatments for the same type of cancer often target the genetic pathways applicable to the specific tumor's genes. As a result, there has been a significant amount of effort devoted to identifying these genetic pathways, identifying potential drugs that could target different aspects of these pathways, and assessing the clinical efficacy of these drugs in human studies. 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.

However, the micro-targeting of patients greatly increases the space of treatment options, which results in fundamental difficulties with putting the findings of precision medicine into practice (Frey et al., 2016). The number of potential treatments and the speed at which they are developed 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. The rate of new cancer drugs being evaluated changes as quickly as once every nine months (Araya et al., 2020), so staying up to date with the latest treatment options is quite challenging for many oncologists. However, the ability to quickly locate relevant evidence is the hallmark of information retrieval (IR), so there is much potential for IR to support precision medicine in practice.

For three 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.

Disease: colorectal cancer **Gene:** ABL1 **Treatment:** Regorafenib
Disease: ovarian carcinoma **Gene:** BRCA1 **Treatment:** Carboplatin
Disease: ovarian carcinoma **Gene:** BRCA2 **Treatment:** Olaparib
Disease: non-small cell lung cancer **Gene:** ERBB2 **Treatment:** Afatinib

Table 1: Example topics from the 2020 track.

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 2019, and is wrapping up with the 2020 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.

Two main changes between the 2017-2019 tracks and the 2020 track took place. First, treatments were added to the topics in order to focus on identifying the best available evidence for a specific treatment, as opposed to identifying a list of potential treatments. Second, the clinical trials task was dropped, both because the topical change made this task less relevant and to make room for the additional judgments required to assess evidence. In order to accommodate the ability to assess evidence, the assessment process was further modified to consider not just the normal notion of retrieval relevance, but also the notion of evidence quality.

This overview is organized as follows: Section 2 provides the historical context of medical IR evaluation leading up 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 provides the results of the participant systems.

2 Background

The TREC Precision Medicine track continues the sizable tradition of biomedical retrieval evaluations within TREC. The first such track series was the 2003-2007 TREC Genomics (Hersh and Voorhees, 2009) tracks, which targeted genomics researchers seeking relevant biomedical literature. The second medical track series was the 2011-2012 TREC Medical Records tracks (Voorhees and Hersh, 2012), which focused on retreiving cohorts of patients from electronic health records (EHRs). The third series was the 2014-2016 TREC Clinical Decision Support (CDS) track (Roberts et al., 2015, 2016a,b), which targeted clinicians seeking evidence-based literature to test, diagnose, or treat patients. Finally, the 2017-2020 TREC Precision Medicine track series (Roberts et al., 2017, 2018, 2019) grew from the CDS track, narrowing the problem domain to precision oncology. The 2020 Precision Medicine track continues this effort.

3 Topics

The 2020 Precision Medicine track provided 40 topics for evaluation. Due to ethical and legal constraints (e.g., HIPAA), extracting real patient data from EHRs was seen as too risky (especially for genomic data). Instead, the topics were synthetically created, though often inspired by actual patients, with modification. To increase the chances of relevant literature articles, the topics were created based on FDA-approved precision oncology drugs for a given cancer/gene combination, according to an online source. This does not guarantee sufficient relevant articles/evidence exists for each topic, for a variety of reasons.

```
<topic number="1">
  <disease>colorectal cancer<disease>
  <gene>ABL1<gene>
  <treatment>Regorafenib<treatment>
<topic>
...
<topic number="6">
  <disease>ovarian carcinoma<disease>
  <gene>BRCA1<gene>
  <treatment>Carboplatin<treatment>
</topic>
...
<topic number="9">
  <disease>ovarian carcinoma<disease>
  <gene>BRCA2<gene>
  <treatment>Olaparib<treatment>
</topic>
...
<topic number="17">
  <disease>non-small cell lung cancer<disease>
  <gene>ERBB2<gene>
  <treatment>Afatinib<treatment>
</topic>
```

Figure 1: XML format for the four topics from Table 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) treatment (e.g., a chemotherapy drug), Four topics from the track are shown in Table 1. An additional four topics are shown in their corresponding XML format (i.e., what was provided to the participants) in Table 1.

4 Tasks

The task this year was focused on finding information about a specific treatment that an oncologist might consider for a patient. Specifically, the task focused on identifying critical evidence for or against the treatment in the specific population represented by the type of cancer and genetic mutation(s) in the topic. The topic structure is close to the Evidence Based Medicine (EBM) "PICO" framework, where the problem/population (P) is the cancer and its mutations, the intervention (I) is the treatment, the comparison (C) is an alternative treatment, and the outcomes (O) are the endpoints of a cancer study, such as event-free survival, quality of life or time to progression. This framework was designed to find the most relevant scientific articles for an individual patient, specifically searching the scientific literature in PubMed. The EBM notion of relevance combines topical evidence with the strength of evidence. The task emulated this notion and required strong evidence for the treatment (whether positive or negative) to be ranked over weaker evidence. There are often many treatments for a particular type of cancer and particular genetic mutations, so a useful clinical decision support tool will help oncologists narrow the treatment decision to the one most likely to help the patient. This is why strong negative evidence is important: it helps eliminate the treatment so that a more efficacious treatment can be chosen instead. The idea is to provide oncologists with the evidence that best helps them make a decision when evaluating competing alternatives. The primary literature corpus is therefore a snapshot of MEDLINE abstracts (i.e., what is searchable through the PubMed interface). The same MEDLINE-baseline snapshot that was used for the 2019 track was used this year. Of course, this collection should not be used to provide real-time decision support in 2020, but it is large enough to support the goals of this evaluation. Specifically, this corpus is composed of 29,138,916 MEDLINE abstracts.

5 Evaluation

The evaluation followed standard TREC 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 per topic. 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.

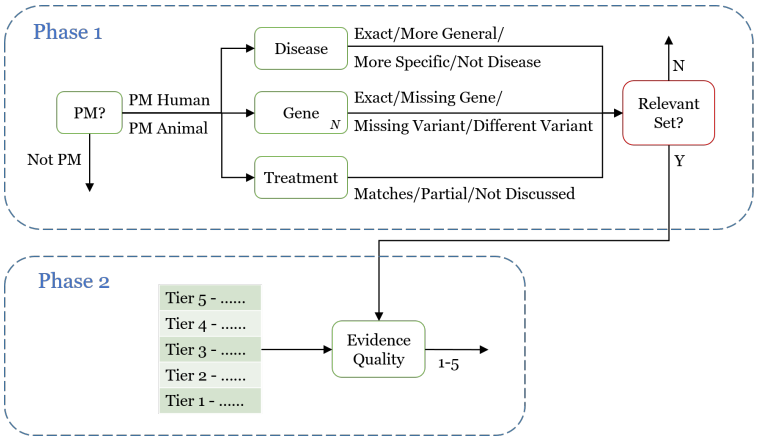


Figure 2: Two-phase assessment process. The first phase focuses on assessing the relevance of each result, whereas the second phase focuses on assessing the quality of the evidence provided by each study.

The assessors were instructed to spend at least 20-30 minutes investigating their topics on their own, to better understand how the genetic variant relates to the type of cancer, as well as the role the treatment plays. As in the past three years, the assessors then manually evaluated the results in a cascading manner shown in Figure 2. Unlike the past years, a second phase was added to the assessment. The first phase consisted of a manual assessment (referred to as *result assessment*) and a mapping of those manual categorizations to a simple relevance scale (referred to as *relevance assessment*). The second phase consisted of the assessors judging the strength of evidence in the study. This is all described below in further detail.

5.1 Result Assessment (Phase 1)

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

See Figure 2 for a flow chart style overview of this process. The first step is designed to save assessor time by filtering out unrelated articles, since the second step can be more time-consuming (possibly requiring a more detailed reading of the article). The assessors were free to quickly skim the abstract in order to make the initial decision. Then, if the article 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 of Phase 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.

The second step of Phase 1 is to determine the appropriate categorization for each of the three fields:

1. *Disease*:
   - **Exact**: The form of cancer in the article is identical to the one in the topic.
   - **More General**: The form of cancer in the article is more general than the one in the topic (e.g., blood cancer vs. leukemia).
   - **More Specific**: The form of cancer in the article is more specific than the one in the topic (e.g., squamous cell lung carcinoma vs. lung cancer).

- **Not Disease**: The article 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 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, as opposed to discussing related work.
   - **Missing Gene**: The article 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 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 focuses on the particular gene in the topic, but on a different variant than the one in the topic.

3. *Treatment*
   - **Matches**: The article directly evaluates the proposed treatment.
   - **Partial**: The article evaluates the proposed treatment as part of a drug combination.
   - **Not Discussed**: The article does not evaluate the proposed treatment.

5.2 Relevance Assessment (Phase 1)

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 *Treatment* is *Matches*.

2. **Partially Relevant**: Largely the same as *Definitely Relevant*, but with the exception that *Disease* can also be *More General*; *Gene* can also be *Missing Variant* or *Different Variant*; and *Treatment* can also be *Partial*.

3. **Not Relevant**: Neither of the above.

The primary evaluation metrics for Phase 1 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.

Topic 16

Tier	Description
4	RCT with > 200 patients and single drug, or Meta-analysis
3	RCT with > 50 patients and single drug, or RCT with > 200 patients and drug combination, or Systemic review
2	Any other RCT not meeting above criteria, or Single drug Phase 2 trial, or Observational study
1	Other study and/or review

Table 2: Example evidence tiers for a 2020 topic.

5.3 Evidence Assessment (Phase 2)

After completion of Phase 1, in Phase 2 (the bottom of Figure 2) the assessor developed a 4-point scale for grading the quality of evidence of a study to help identify the most important relevant studies so that these can be prioritized. (In addition to the 4 points, an N/A option was available for items that do not belong on the scale, such as articles that should have been judged as not relevant in Phase 1.) Based on the understanding the assessor gained in Phase 1, the 4-point scale was tailored for each topic. For example, the top tier (Tier 4) may only include randomized controlled trials (RCTs) for the specific drug (not in combination), while the bottom tier may only include animal studies. For a different topic, the scale could be placing RCTs on two different tiers based on the conclusiveness of the results of the study. See Table 2 for an example. An important point to note, however, is that conclusive results are considered equal, whether positive (the drug definitely worked) or negative (the drug definitely did not) and both are preferable to weaker or inconclusive results. The assessors documented the proposed tiers based on a provided template and submitted the scale to the coordinators for approval. After approval, the assessors proceeded to re-judge a sample of up to 100 abstracts that were judged minimally relevant in the first phase.

The evaluation metric for this phase used NDCG at rank 30 (NDCG@30). Two relevance values were used based on different weights of the assigned tiers:

- NDCG@30 using gains scores corresponding to the tiers categorized by the assessors: {0, 1, 2, 3, 4}. This is referred to as *std-gains* below.
- NDCG@30 using exponential gains scores: {0, 1, 2, 4, 8}. This intentionally emphasizes top-tier articles which are the most critical to return. This is referred to as *exp-gains* below.

6 Results

In total, there were 22,806 judgments for Phase 1 and 2,691 judgments for Phase 2. Table 3 shows the number of Definitely Relevant, Partially Relevant, and Not Relevant judgments for each topic for Phase 1. Since each result was judged only once, no inter-rater agreement is available for the judgments.

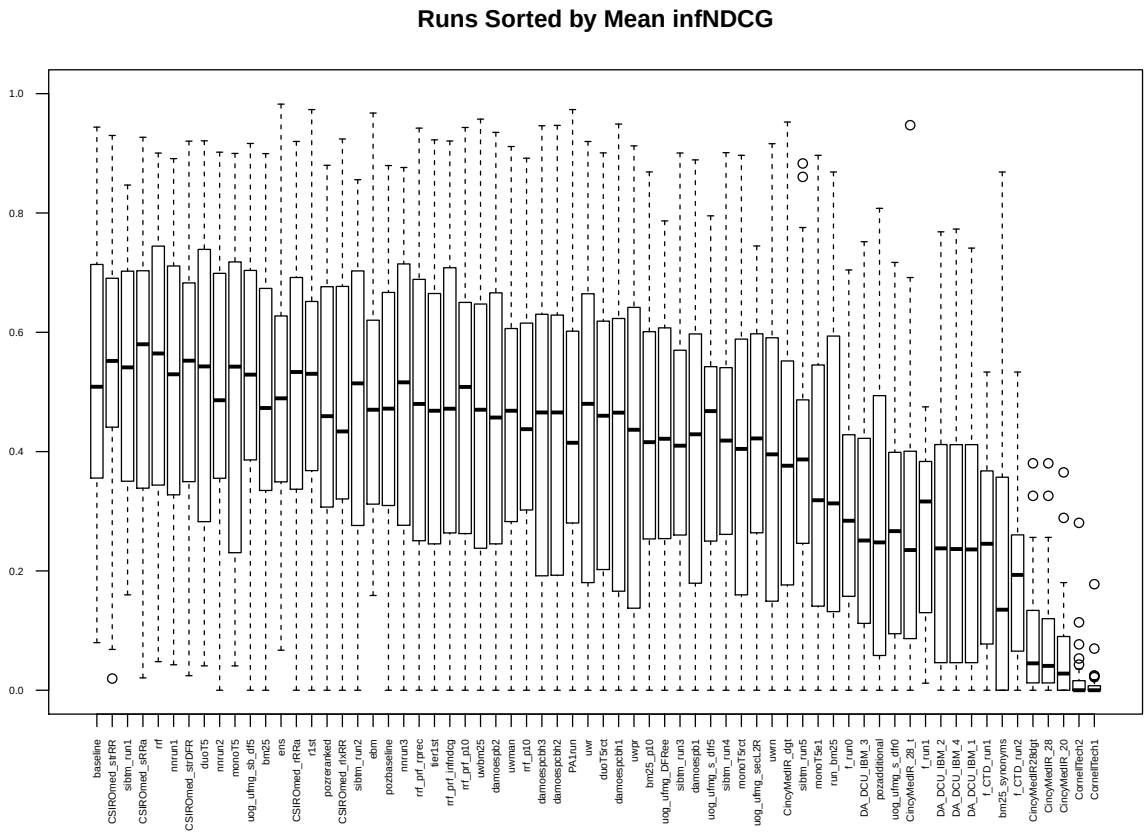
There were 16 participants in the track, submitting a total of 66 runs. See Table 4 for a list of the participants and numbers of runs. Table 5 shows the top 5 runs (top run per participant) for each metric for Phase 1, while Table 6 shows the top 5 runs for Phase 2. Figure 3 shows box-and-whisker plots for the Phase 1 scores for all the runs, both infNDCG and P@10.

Topic	DR	PR	NR	Topic	DR	PR	NR
1		1	544	21	2		410
2	177	141	213	22	1	6	627
3			680	23			555
4	15	33	410	24	14	14	374
5	2	9	908	25	9	21	467
6	3	22	691	26	13	3	466
7	8	119	313	27	22	6	707
8	19	18	413	28		6	558
9	30	32	737	29	61	19	495
10	59	29	338	30			475
11	75	18	226	31			964
12	27	47	264	32	2	4	489
13	3	61	400	33	1	4	646
14	71	38	303	34			602
15	226	94	252	35	14	5	417
16	182	44	302	36	2	2	641
17	148	103	346	37		10	504
18		5	659	38			647
19	1		559	39	1	9	482
20	3	17	892	40	4	6	689

Table 3: Counts of Definitely Relevant (DR), Partially Relevant (PR), and Not Relevant (NR) results for each topic.

Team ID	Affiliation	# Runs
ALIBABA	Alibaba Group	5*
ASCFDA	Academia Sinica CFDA Lab	1
BIT.UA	Universidade de Aveiro	5
BITEM	SIB Text Mining / BiTeM group	5
CSIROmed	The Commonwealth Scientific and Industrial Research Organisation	5
CTIR	Cornell Tech	2
CincyMedIR	University of Cincinnati Department of Biomedical Informatics	5
DA_IICT	Dhirubhai Ambani Institute of Information and Communication Technology	4
MRG_UWaterloo	School of Computer Science, University of Waterloo	5*
PINGAN_NLP	Ping An Technology (Shenzhen) Co., Ltd.	5
POZNAN	Poznan University of Technology	3
READ-Biomed	Reading, Extraction, and Annotation of Documents in Biomedicine	4
UoGTr	University of Glasgow	5
h2oloo	University of Waterloo	5
ims_unipd	University of Padua	5
vohcolab	VOH.CoLAB	2
Total		66

Table 4: Participating teams and submitted runs. * Includes 3 manual runs each from ALIBABA and MRG_UWaterloo.



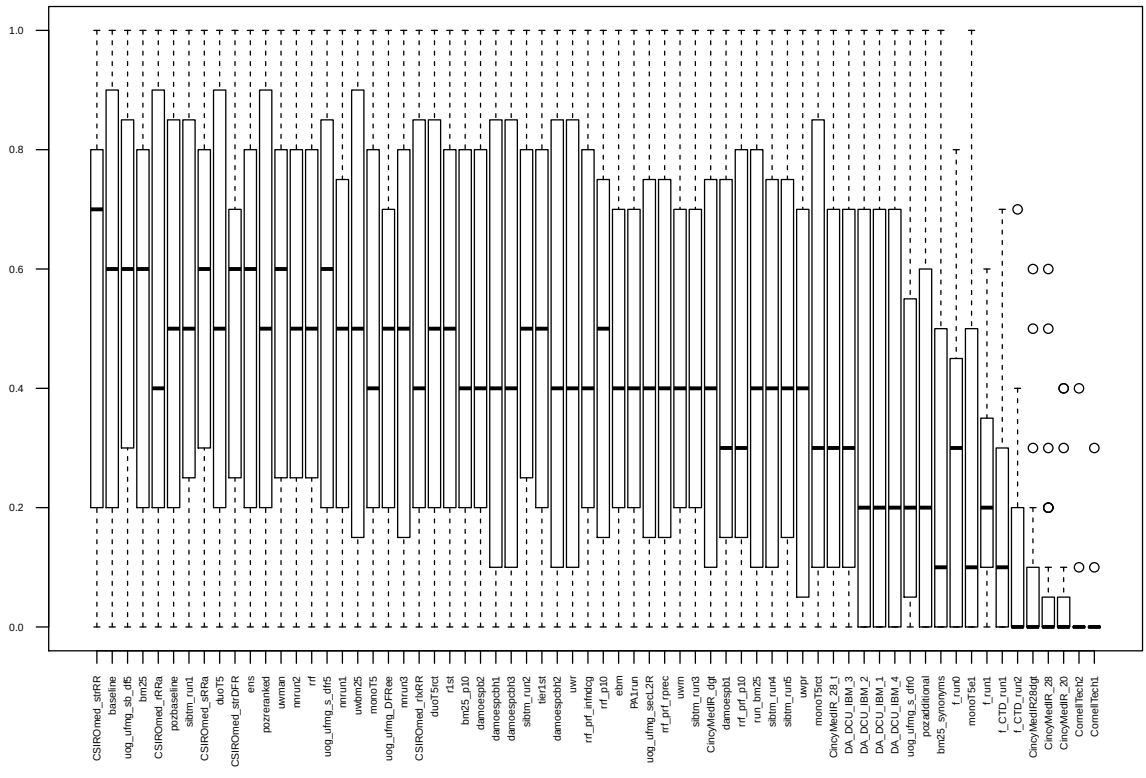


Figure 3: Phase 1 score distributions for all runs.

Team	**Run**	**infNDCG**
BIT.UA	baseline	0.5325
CSIROmed	CSIROmed_strRR	0.5303
BITEM	sibtm_run1	0.5276
h2oloo	duoT5	0.5116
UoGTr	uog_ufmg_sb_df5	0.4979
		R-prec
CSIROmed	CSIROmed_strRR	0.4358
BIT.UA	baseline	0.4207
PINGAN_NLP	r1st	0.4176
BITEM	sibtm_run1	0.4020
h2oloo	monoT5	0.4018
		P@10
CSIROmed	CSIROmed_strRR	0.5645
BIT.UA	baseline	0.5516
UoGTr	uog_ufmg_bg_df5	0.5484
ASCFDA	bm25	0.5355
POZNAN	pozbaseline	0.5323

Table 5: Top overall systems in Phase 1 (best run per team).

Team	**Run**	NDCG@30 std-gains
ALIBABA	damoespcbh3	0.4780
h2oloo	monoT5rct	0.4238
BIT.UA	rrf	0.3717
UoGTr	uog_ufmg_sb_df5	0.3682
MRG_UWaterloo	uwman	0.3562
		NDCG@30 exp-gains
ALIBABA	damoespcbh3	0.4519
h2oloo	monoT5rct	0.4193
UoGTr	uog_ufmg_sb_df5	0.3290
BIT.UA	rrf	0.3289
ims_unipd	rrf_prf_rprec	0.3157

Table 6: Top overall systems in Phase 2 (best run per team).

7 Conclusion

For four years, the Precision Medicine track has sought to inform the creation of information retrieval systems that support precision medicine generally, and precision oncology most specifically. The focus of these tracks was IR systems that aided in the treatment decision process, be it providing evidence-based literature articles for evaluated treatment or relevant clinical trials for ongoing treatment evaluations. For its first three years (2017-2019), the track used topics seeking specific treatments. For its final year (2020), the track focused instead on identifying the highest-quality evidence for a specific treatment. Participants were provided with 40 topics representing synthetic patients and ranked according to multiple scales (relevance and evidence).

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 R21EB029575-01 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

Araya, A., Zeng, J., Johnson, A., Shufean, M. A., Rodon, J., Meric-Bernstam, F., and Bernstam, E. V. (2020). Rate of change in investigational treatment options: An analysis of reports from a large precision oncology decision support effort. *International Journal of Medical Informatics*, 143:104261.

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., and Lazar, A. (2019). Overview of the TREC 2019 Precision Medicine Track. In *Proceedings of the Twenty-Eighth 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*.