Overview of the TREC 2022 Clinical Trials 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, William R. Hersh
Department of Medical Informatics & Clinical Epidemiology,
Oregon Health and Science University, Portland, OR

Overview

Clinical trials are the primary means for new medical treatments–such as drugs, surgical procedures, and behavior interventions–to demonstrate their effectiveness in an evidence-based manner. However, clinical trials have high costs, can take years to complete, and oftentimes fail to identify a sufficient number of patients to establish clinical significance. Automated methods for improving the patient recruitment process can aid in all three areas: reducing manual and expensive chart review, more quickly identifying eligible patients, and expanding the pool of candidate patients that may be eligible.

The primary means of automating clinical trial recruitment is through the use of electronic health record (EHR) data. EHRs are responsible for documenting routine medical care, as well as being the legal and billing record. However, the re-use of EHR data for research is well-established (Hersh, 2007), commonly for observational studies but also as the source data for informatics-driven models, including machine learning (ML) and information retrieval (IR). This was the inspiration behind the TREC Medical Records track (2011-2012) (Voorhees and Tong, 2011; Voorhees and Hersh, 2012), which used short cohort descriptions as queries (e.g., "Patients treated for vascular claudication surgically") and used EHR visit records as the document collection. Unfortunately, this track was discontinued due to the the lack of an EHR dataset of sufficient size to merit a proper IR evaluation. The TREC Clinical Trials track, instead, flips the trial-to-patients paradigm to a patient-to-trials paradigm. This has enabled the building of a large test collection for clinical trial search. In this paradigm, the topic is a (synthetic) patient description and the document collection is a large set of clinical trial descriptions (which are, notably, publicly and freely available).

There are several challenges involved with task, however. The first set of challenges revolve around using clinical trial descriptions as the document collection. Clinical trial descriptions are often very long (see link to trial in Table 2). The core part of the description with regards to trial matching is the eligibility criteria, a (often long) list of inclusion criteria (the patient must meet all these requirements) and exclusion criteria (if the patient meets any of these criteria, they are ineligible and would be excluded from the trial). These criteria not only use complex medical terminology, but they are often written in a way that does not correspond directly to how patient cases are described in the EHR, making direct term mapping problematic.

The second set of challenges revolves around the patient cases. In addition to the linguistic issues of how identical clinical concepts in EHR text versus trial descriptions, patient cases contain significant amounts of extraneous information with respect to the clinical trial. That is, not all of the information in a patient case need be covered in the trial. Rather, a sufficient amount of information must be present to suggest the patient may be eligible, while also not containing information showing the patient to be excluded. This

means that many of the conditions in the patient description are irrelevant for a single clinical trial, whereas matching to a different clinical trial may involve a different subset of conditions in the patient case.

As in 2021 Roberts et al. (2021b), to ensure this task focuses on information retrieval and not supervised information extraction, we present a lengthy (5-10 sentence) patient case description as the topic that simulates an admission statement in an EHR. The evaluation is further be broken down into *Eligible*, *Excludes*, and *Not Relevant* to allow retrieval methods to distinguish between patients that do not have sufficient information to qualify for the trial (*Not Relevant*) and those that are explicitly *Excluded*. This latter category can be difficult for retrieval systems without strong semantic understanding.

Background

There is a long-established history of biomedical IR tracks within TREC. This includes the Genomics track (2003-2007) (Hersh and Bhupatiraju, 2003; Hersh et al., 2004, 2005, 2006), the Medical Records (2011-2012) (Voorhees and Tong, 2011; Voorhees and Hersh, 2012), the Clinical Decision Support track (2015-2016) (Simpson et al., 2014; Roberts et al., 2015, 2016), the Precision Medicine track (2018-2020) (Roberts et al., 2017, 2018, 2019, 2020), the Health Misinformation track (2019-2021) (Abualsaud et al., 2019; Clarke et al., 2020, 2021), and the TREC-COVID track (Roberts et al., 2021a). Of all these tracks, the Precision Medicine track is the most similar to the Clinical Trials track. Indeed, one of the tasks for the TREC Precision Medicine track was to retrieve clinical trials for synthetic patient topics. The TREC Clinical Trials track, then, has expanded this notion of clinical trial search beyond the precision medicine paradigm to all human clinical trials.

Topics

Within the context of the TREC Clinical Trials track for 2021 and 2022, a topic is a brief patient case description, such as what may be included as part of an admission note. Four of the 2022 topics are shown in Table 1. Most topics are prose paragraphs, resembling a traditional medical case description, while others have additional information in the form of a list, oftentimes lab values (which are frequently used for clinical trial eligibility). All topics tend to use language and abbreviations found in clinical notes. All the topics were derived with a specific disease in mind (e.g., the first topic from Table 1 is an *ectopic pregnancy* patient), but this disease was not provided to participants or judges. Ultimately, a patient could be eligible for trials outside the intended disease–the disease was simply used to ensure a broad distribution of topics.

The 2022 track had 50 topics, which was a slight decrease from 2021 in order to ensure deeper pools could be made for evaluation.

Data

The 2022 track used the same snapshot of ClinicalTrials.gov that was used for the 2021 Clinical Trials track. Briefly, the collection is a April 27, 2021 snapshot of all the clinical trials available on ClinicalTrials.gov. U.S. policy dictates that all clinical trials conducted in the United States post their trial information to this website, which is maintained by the U.S. National Library of Medicine. The collection used for the task was hosted on the trec-cds.org website that maintains the topics and data for many biomedical TREC tracks. The data is available as XML, with this specific snapshot containing 375,581 clinical trial descriptions. Each clinical trial is assigned a National Clinical Trial (NCT) designation number (e.g., NCT00392756), which is used as the document ID for the track. These are the same IDs reported in the final publications describing the clinical trial results (a common clinical journal requirement).

Assessment

Assessment used the same interface as in 2021 (see Figure 1), which was modified from prior biomedical TREC tracks. Also as in 2021, assessors judged results with a 3-point scale:

Topic 2 A 32-year-old woman comes to the hospital with vaginal spotting. Her last menstrual period was 10 weeks ago. She has regular menses lasting for 6 days and repeating every 29 days. Medical history is significant for appendectomy and several complicated UTIs. She has multiple male partners, and she is inconsistent with using barrier contraceptives. Vital signs are normal. Serum β-hCG level is 1800 mIU/mL, and a repeat level after 2 days shows an abnormal rise to 2100 mIU/mL. Pelvic ultrasound reveals a thin endometrium with no gestational sac in the uterus.

Topic 6 A 61-year-old man comes to the clinic due to nonproductive cough and progressive dyspnea. The patient's medical conditions include hypertension, hypercholesteremia and peptic ulcer disease. He smokes 2 packs of cigarettes daily for the past 30 years. On examination, there are decreased breath sounds and percussive dullness at the base of the left lung. Other vital signs are normal. Abdomen is soft without tenderness. CT scan shows a left-sided pleural effusion and nodular thickening of the pleura. The plural fluid was bloody on thoracentesis. Biopsy shows proliferation of epithelioid-type cells with very long microvilli.

Topic 19 A 7-year-old girl is brought to the emergency department by her parents for generalized rash. The mother reports that she was playing outside wearing a skirt and felt a sharp pain in her arm while seating on a mat, plying with her doll. Her mother suspects that something had stung her. The patient's blood pressure is 75/55 mm Hg and her heart rate is 122/min. Physical examination shows erythematous, raised plaques over the trunk, extremities, and face. Lung auscultation reveals bilateral expiratory wheezes.

Topic A 15-year-old boy with mild intellectual disability is brought to the office by his parents for a routine physical examination. The boy is going to a school for students with learning disabilities. The patient was adopted, and his immunizations are up to date. Review of the patient's medical records is notable for cytogenetic studies that showed a small gap near the tip of the long arm of the X chromosome, which is consistent with fragile X syndrome, an X-linked disorder. The defect is an unstable expansion of trinucleotide repeats (CGG) in the fragile X mental retardation 1 (FMR1) gene, located on the long arm of the X chromosome. He is not using any medications and vital signs are within normal levels. His blood chemistry analysis as bellow:

Blood Chemistry Value	Normal Range	Patient Value
Glucose	90-120 mg/dl	95 mg/dl
BUN (Blood Urea Nitrogen)	7-24 mg/dl	10 mg/dl
Creatinine	0.7-1.4 mg/dl	0.8 mg/dl
Calcium	8.5-10.5 mg/dl	9 mg/dl
Sodium	134-143 mEq/L	135 mEq/L
Potassium	3.5-4.5 mEq/L	3.7 mEq/L
Chloride	95-108 mEq/L	98 mEq/L
CO2	20-30 mEq/L	25 mEq/L
Blood pH	7.38-7.42	7. 39

Table 1: Example topics from the TREC 2022 Clinical Trials track.

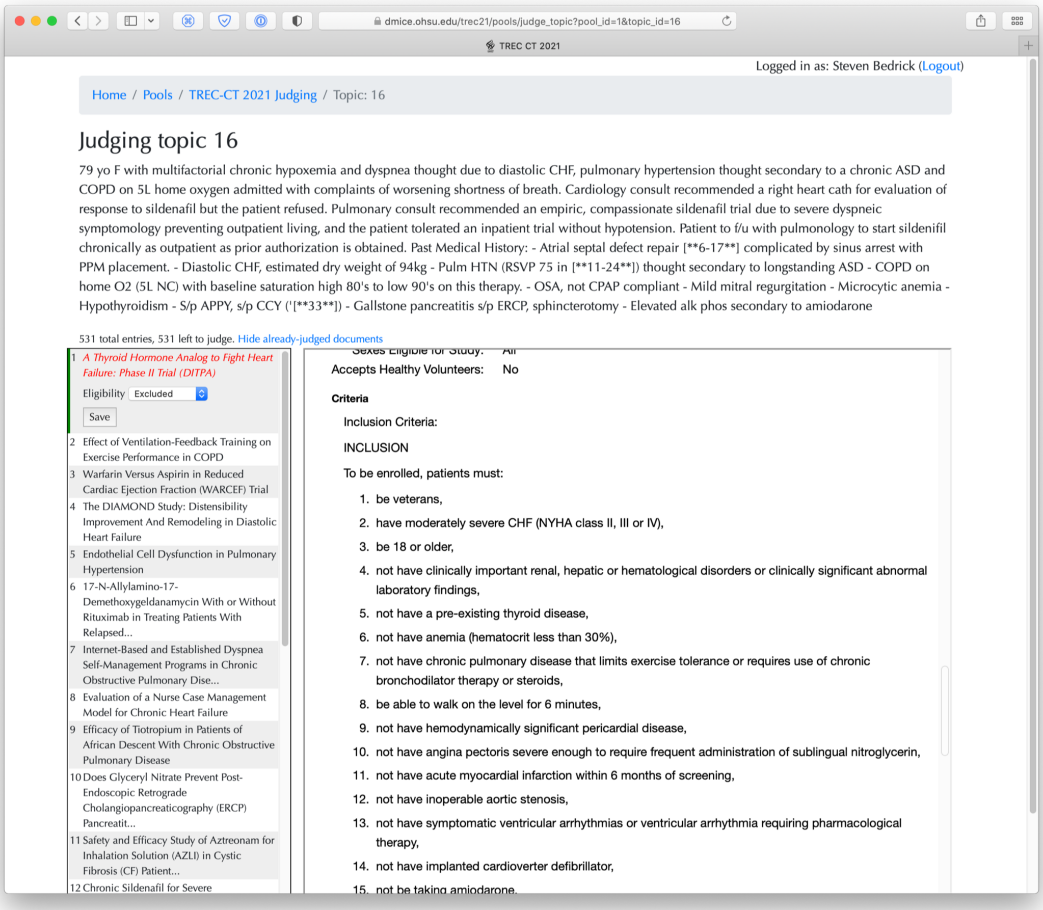

Figure 1: Screenshot of assessment platform.

1. **Not Relevant**. The patient is not relevant for the trial in any way.
2. **Excluded**. The patient has the condition that the trial is targeting, but the exclusion criteria make the patient ineligible.
3. **Eligible**. The patient is eligible to enroll in the trial.

Prior to the assessment process, the assessors were asked to spend 20-30 minutes to get any necessary background on the topic. This includes familiarizing themselves with the patient's case, any mentioned diseases, as well as the types of trials available for these diseases on ClinicalTrials.gov to get a sense for both the site itself and the ways in which the specific diseases were described in the clinical trial eligibility criteria.

Table 2 gives a sense for what clinical trial eligibility criteria (both inclusion and exclusion) look like in the trial descriptions. Two important conditions were not considered reasons for ineligibility:

- **Recruitment Status**. Clinical trials only allow recruitment during a specified window, ending either once the trial meets its recruitment goal, at a specified time point, or when the trial ia canceled. This also includes excluding recruitment prior to the start of the trial, as well as after recruitment has ended (which is still prior to study completion). Since this is not really a textual/semantic aspect relevant to information retrieval, we ignored the recruitment status. This increases the number of relevant patient/trial matches.

- **Location**. Many trials only enroll patients at certain locations. We ignored this as well, which also increases the number of relevant trials.

Trial: NCT01160822
Title: **To Determine the Safety, Tolerability, Pharmacokinetics and Effect on Pain of a Single Intra-articular Administration of Canakinumab in Patients With Osteoarthritis in the Knee**

Inclusion Criteria:

1. Written informed consent must be obtained before any assessment is performed.
2. Male and female patients aged 40 - 80 years (inclusive).
3. Diagnosis of knee osteoarthritis
4. Radiographic evidence of tibiofemoral compartment osteoarthritis
5. Pain in the knee during the last 24 hours.The patients should also have had pain in the affected knee on most days over the last month.
6. Patients who are willing to discontinue all non-steroidal anti-inflammatory drugs (NSAIDs) or other analgesic medication taken for any condition, including their knee pain,
7. Patients who are on stable dose of opioids for at least 1 month before screening can continue to take their opioid at this stable dose throughout the study.
8. Patients must also be willing to abstain from any intra-articular or peri-articular injections to the knee or surgery during the treatment period
9. Patients who, if they are currently taking aspirin (325 mg/day or less; as anti-coagulants), are willing to remain on a stable dose one month prior to screening and throughout the study

Exclusion Criteria:

1. Subjects with known hypersensitivity to any biological or investigational drugs.
2. Patients with contraindications to knee injections
3. Patients with joint effusion
4. Patients should not have rheumatoid arthritis or any connective tissue like disease
5. Secondary osteoarthritis with history and/or any evidence of the following diseases: septic arthritis, inflammatory joint disease, gout, Paget's disease of the bone, articular fracture, major dysplasias or congenital abnormality, ochronosis, acromegaly, hemochromatosis, Wilson's disease, primary osteochondromatosis, juvenile chronic arthritis with continued activity in adulthood, heritable disorders (e.g. hypermobility). Patients with secondary osteoarthritis following menisectomy or injuries of a collateral or cruciate ligament are not excluded.
6. Presence or history of underlying metabolic, endocrine, hematologic, pulmonary, cardiac, blood, renal, hepatic, infectious, psychiatric or gastrointestinal conditions
7. Evidence of tuberculosis (TB)
8. One of the risk factors for TB such as:
   (a) Substance abuse (e.g. injection or non-injection)
   (b) Health-care workers with unprotected exposure to patients who are at high risk of TB
   (c) Patients with TB disease before the identification and correct airborne precautions of the patient
   (d) close contact (i.e. share the same air space in a household or other enclosed environment for a prolonged period (days or weeks, not minutes or hours)) with a person with active pulmonary TB disease.
9. Significant medical problems, including but not limited to the following: uncontrolled hypertension,congestive heart failure, uncontrolled diabetes type I and II
10. Subjects with evidence of hepatic or blood coagulation disorders (i.e. hemophilia, etc), anemia, idiopathic thrombocytopenic purpura, or gastrointestinal disorder: severe hepatic disease, history of alcohol and drug abuse; disease of gall bladder and pancreas; active peptic ulceration, gastrointestinal bleeding or history of severe gastro-esophageal reflux disease or severe hiatus hernia; inflammatory bowel disease.
11. Use of any therapeutic protein drug (e.g. anti-tumor necrosis factor alpha (TNFα) antibody)
12. Presence of severe renal function impairment. History of renal trauma, glomerulonephritis, patients with one kidney, or renal failure requiring regular dialysis treatment.
13. Pregnant or nursing (lactating) women, where pregnancy is defined as the state of a female after conception and until the termination of gestation, confirmed by a positive pregnancy test (serum or urine).
14. Subjects with known contra-indications to naproxen (e.g. heart or circulation problems, history of ulcer disease etc.), analgesics, antipyretics, or NSAIDs.
15. Disease of the spine or other lower extremity joints which may interfere with the assessment of the target joint.
16. Surgery on the knee within the last year. Observational arthroscopy, arthroscopic surgery or lavage of the knee within the last 6 months.
17. Use of assistive devices other than a cane (walking stick) or knee brace.
18. Subjects who have experienced, any time in the past, asthma, acute rhinitis, nasal polyps, angioneurotic edema, urticaria or other allergic-type reaction after taking acetylsalicylic acid (ASA)/ aspirin or NSAIDs.
19. Any history of prior peptic ulcer disease or prior NSAID gastrointestinal complications for the past 5 years.
20. Other protocol defined inclusion/exclusion criteria may apply.

URL: `https://www.clinicaltrials.gov/ct2/show/NCT01160822`

Table 2: Example clinical trial inclusion/exclusion criteria.

In other words, time and space are not considered when assessing eligibility, only the medical inclusion/exclusion criteria for the trial itself.

Evaluation

The 2022 Clinical Trials track evaluation followed standard TREC evaluation procedures for ad hoc retrieval tasks, as was the case for 2021. Participants submitted results in the `trec_eval` format, with a maximum of five runs (automatic or manual) per task. Each run consisted of a ranked list of up to 1,000 clinical trial IDs (ClinicalTrials.gov Identifiers) per topic. The highest ranked trials for each topic were pooled and judged by physician graduate students at OHSU, indexers at the U.S. National Libary of Medicine, and other biomedical subject matter experts.

Due to the nature of the task, one can judge results according to both traditional relevance (*Eligible* and *Excluded*) as well as eligibility (*Eligible*). The focus of our evaluations for this track was on the latter, as this is more desirable from an application perspective. Explicitly, for NDCG, an *Eligible* trial was given a score of 2, an *Excluded* trial was given a score of 1, and a *Not Relevant* trial was given a score of 0. For all other metrics, *Eligible* is treated as relevant and *Excluded* is combined with *Not Relevant*.

Results

A total of 41 runs were submitted by 11 teams. Of these, there were 6 manual runs and there were 35 automatic runs.

All 41 runs were pooled to depth 40. This resulted in a total of 35,394 runs for judging. The judged results include 3,949 *Eligible* trials (11%) and 3,047 *Excluded* trials (9%), with the remaining 28,481 trials judged *Not Relevant*. Table 3 shows the per-topic counts of relevant, partially relevant, and total judged trials.

Table 4 shows the participant results across the four primary metrics for the track (NDCG@10, P@10, RPrec, and MRR). Figures 2 shows the distribution of scores (only the best run per team) for each of the participants.

The per-topic perspective across all 41 runs is shown in Figures 3 and 4. These figures demonstrate a wide range of scores, with many topics being "easy" for most participants, many topics being "hard" for most participants, and many other topics with a wide range of scores. When comparing these figures with Table 3, there is a reasonable alignment between topics with few relevant results and those where the mean automatic runs were poor.

References

Abualsaud, M., Smucker, M. D., Lioma, C., Maistro, M., and Zuccon, G. (2019). Overview of the TREC 2019 Decision Track. In *Proceedings of the Twenty-Eighth Text Retrieval Conference*.

Clarke, C. L. A., Maistro, M., and Smucker, M. D. (2021). Overview of the TREC 2021 Health Misinformation Track. In *Proceedings of the Thirtieth Text Retrieval Conference*.

Clarke, C. L. A., Smucker, M. D., Maistro, M., and Zuccon, G. (2020). Overview of the TREC 2020 Health Misinformation Track. In *Proceedings of the Twenty-Ninth Text Retrieval Conference*.

Hersh, W. and Bhupatiraju, R. T. (2003). TREC Genomics Track Overview. In *Proceedings of the Twelfth Text REtrieval Conference*.

Hersh, W., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., and Kraemer, D. F. (2004). TREC 2004 Genomics Track Overview. In *Proceedings of the Thirteenth Text REtrieval Conference*.

Hersh, W., Cohen, A., Yang, J., Bhupatiraju, R. T., Roberts, P., and Hearst, M. (2005). TREC 2005 Genomics Track Overview. In *Proceedings of the Fourteenth Text REtrieval Conference*.

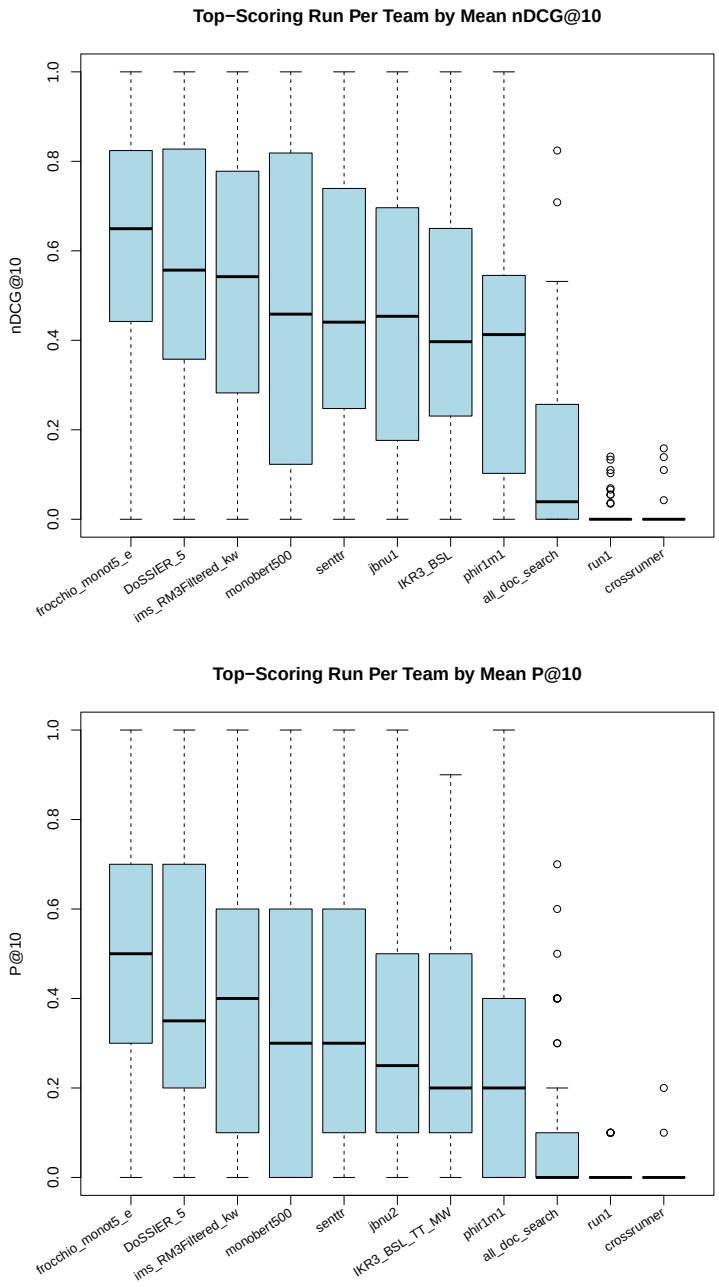

Figure 2: Distribution of scores across topics for each team.

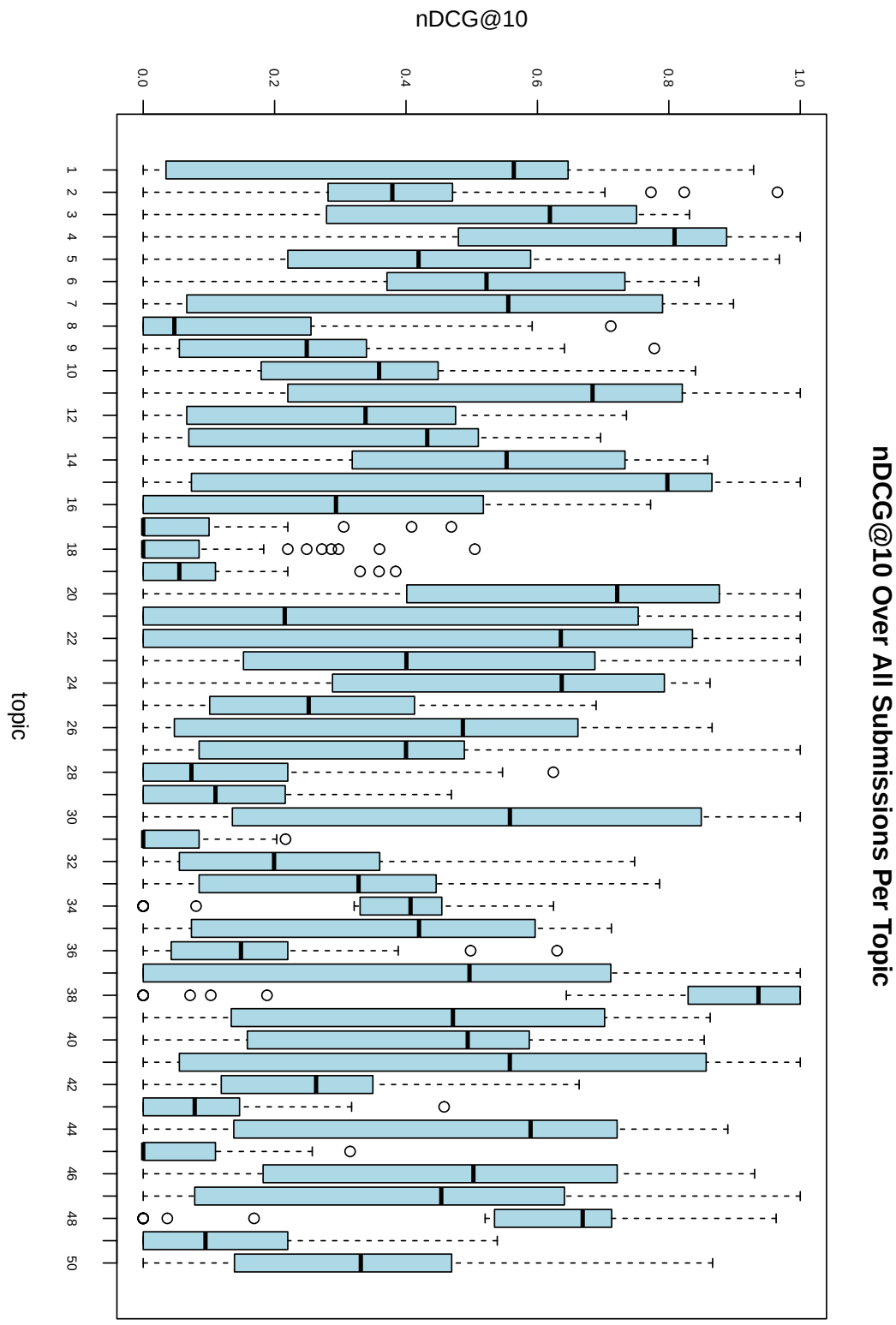

Figure 3: Distribution of run results across all 75 topics for NDCG@10

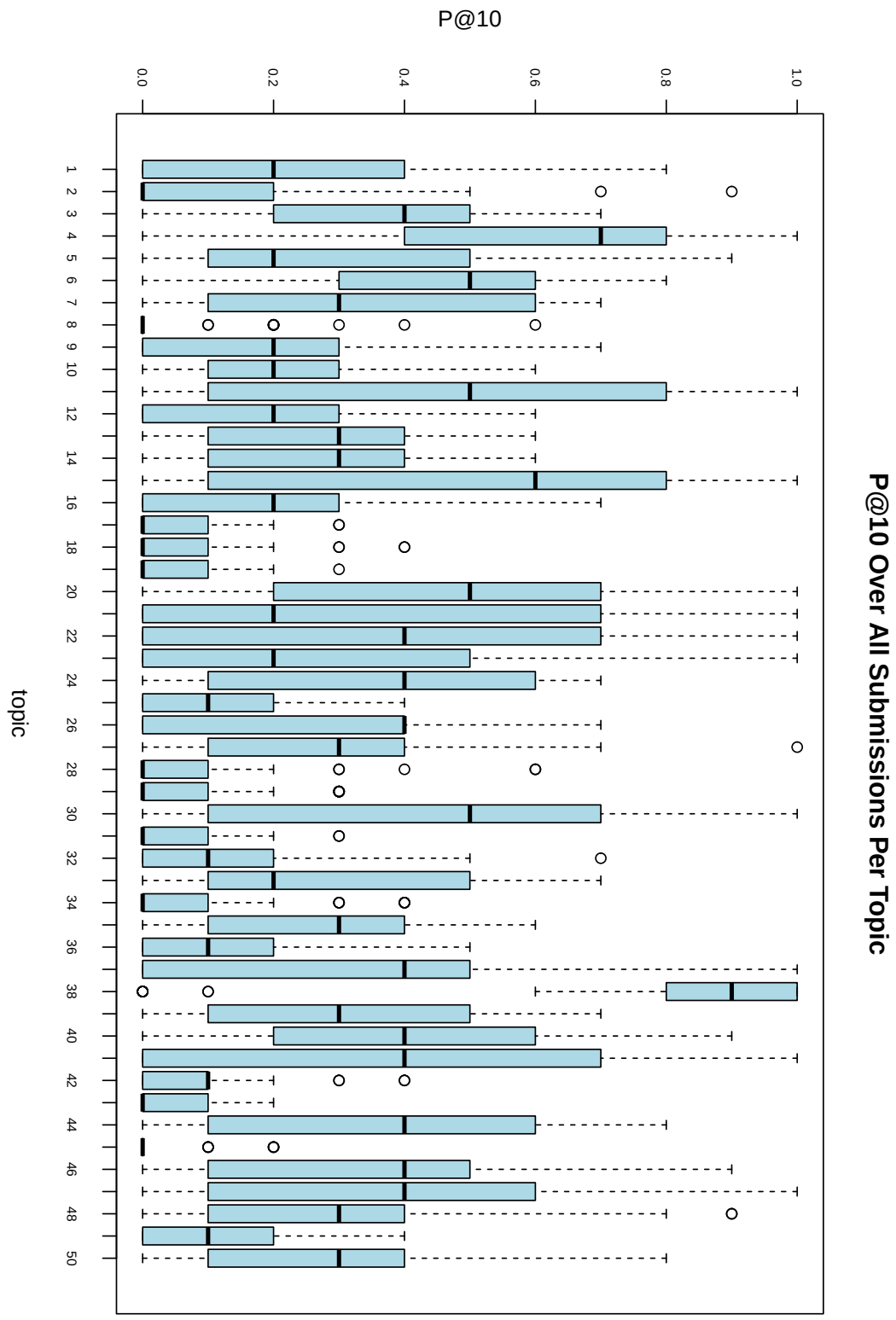

Figure 4: Distribution of run results across all 75 topics for P@10

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006). TREC 2006 Genomics Track Overview. In *Proceedings of the Fifteenth Text REtrieval Conference*.

Hersh, W. R. (2007). Adding Value to the Electronic Health Record Through Secondary Use of Data for Quality Assurance, Research, and Surveillance. *The American Journal of Managed Care*, 13(6):277–278.

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L. L., and Hersh, W. R. (2021a). Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *Journal of Biomedical Informatics*, 121:103865.

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., Bedrick, S., and Hersh, W. R. (2021b). Overview of the TREC 2021 Clinical Trials Track. In *Proceedings of the Thirtieth Text Retrieval Conference*.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., and Bedrick, S. (2020). Overview of the TREC 2020 Precision Medicine Track. In *Proceedings of the Twenty-Ninth 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., Voorhees, E., and Hersh, W. (2015). Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of the Twenty-Fourth Text Retrieval Conference*.

Simpson, M. S., Voorhees, E., and Hersh, W. (2014). Overview of the TREC 2014 Clinical Decision Support Track. In *Proceedings of the Twenty-Third 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*.

Voorhees, E. M. and Tong, R. M. (2011). Overview of the TREC 2011 Medical Records Track. In *Proceedings of the 10th Text REtrieval Conference*.

Topic	Relevant	Partial	Total Judged
1	76 (11%)	117 (16%)	717
2	59 (8%)	339 (43%)	781
3	25 (4%)	54 (8%)	655
4	268 (39%)	75 (11%)	690
5	59 (8%)	51 (7%)	742
6	118 (18%)	20 (3%)	647
7	15 (2%)	47 (7%)	653
8	14 (2%)	26 (3%)	779
9	77 (10%)	53 (7%)	792
10	113 (15%)	105 (14%)	759
11	235 (29%)	69 (8%)	822
12	51 (7%)	15 (2%)	714
13	14 (2%)	7 (1%)	647
14	33 (6%)	89 (17%)	523
15	122 (22%)	49 (9%)	555
16	36 (5%)	12 (2%)	778
17	24 (3%)	6 (1%)	752
18	36 (4%)	13 (1%)	901
19	10 (1%)	4 (0%)	944
20	173 (30%)	61 (11%)	572
21	160 (23%)	13 (2%)	697
22	77 (11%)	48 (7%)	731
23	121 (15%)	67 (8%)	824
24	13 (2%)	18 (2%)	736
25	86 (13%)	95 (15%)	641
26	30 (5%)	26 (5%)	566
27	25 (5%)	4 (1%)	457
28	49 (6%)	19 (2%)	866
29	14 (2%)	19 (2%)	795
30	306 (40%)	76 (10%)	761
31	25 (4%)	14 (2%)	642
32	84 (10%)	135 (15%)	883
33	60 (8%)	12 (2%)	706
34	12 (2%)	68 (12%)	564
35	139 (20%)	121 (17%)	703
36	65 (10%)	12 (2%)	658
37	39 (6%)	39 (6%)	707
38	139 (37%)	20 (5%)	380
39	144 (22%)	120 (18%)	658
40	106 (15%)	57 (8%)	713
41	54 (7%)	16 (2%)	774
42	45 (6%)	134 (19%)	707
43	13 (1%)	71 (8%)	904
44	108 (16%)	158 (23%)	689
45	11 (2%)	10 (1%)	699
46	141 (20%)	173 (25%)	701
47	90 (15%)	6 (1%)	604
48	88 (17%)	241 (45%)	532
49	16 (2%)	2 (0%)	950
50	121 (17%)	30 (4%)	723
TOTAL	3939 (11%)	3036 (9%)	35394

Table 3: Per-topic counts of relevant and partially-results.

Run Name	Team	NDCG@10
frocchio_monot5_e	h2oloo	0.6125
DoSSIER_5	DOSSIER	0.5565
ims_RM3Filtered_kw*	iiia-unipd	0.5051
monobert500	CSIROmed	0.4912
senttr	els_dshs	0.4758
jbnu1	jbnu	0.4530

Run Name	Team	P@10
frocchio_monot5_e	h2oloo	0.5080
DoSSIER_5	DOSSIER	0.4560
ims_RM3Filtered_kw*	iiia-unipd	0.3980
monobert500	CSIROmed	0.3620
senttr	els_dshs	0.3540
jbnu2	jbnu	0.3220

Run Name	Team	RPrec
frocchio_monot5_e	h2oloo	0.3297
DoSSIER_3	DOSSIER	0.2810
ims_RM3Filtered_kw*	iiia-unipd	0.2790
jbnu1	jbnu	0.2233
monobert500	CSIROmed	0.2136
senttr	els_dshs	0.2128

Run Name	Team	MRR
frocchio_monot5_e	h2oloo	0.7262
DoSSIER_2	DOSSIER	0.6607
zs_bert_500	CSIROmed	0.6117
ims_BM25Filtered_kw*	iiia-unipd	0.6085
jbnu2	jbnu	0.5543
phir1m1	phi_lab	0.5516

Table 4: Top 6 runs (best run for each team) for all four metrics (manual runs marked with *).