



# Interpretable Clinical Trial Search using Pubmed Citation Network

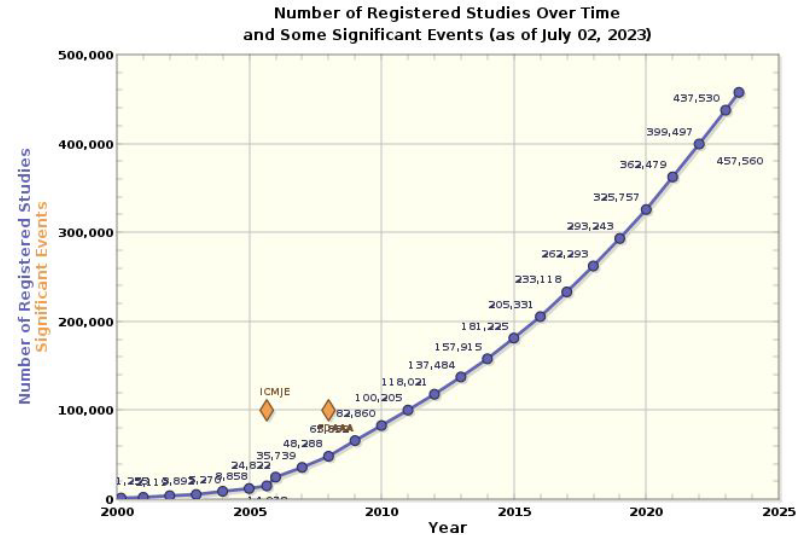
Soumyadeep Roy (IIT Kharagpur, India; L3S Research Center, Germany )

Niloy Ganguly, Shamik Sural (IIT Kharagpur, India)

Koustav Rudra (IIT (ISM) Dhanbad, India)

# Why are clinical trials important ?

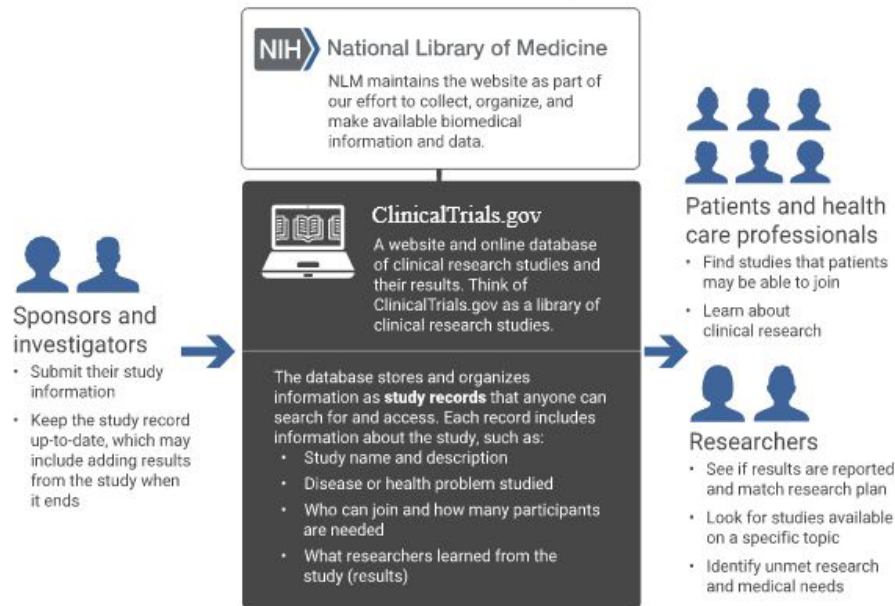
- Clinical trials provide the earliest source of information about new drugs and treatments
- Clinical trial search systems help meet the information need, given the rising volume of clinical trials and related publications



Source: <https://ClinicalTrials.gov>

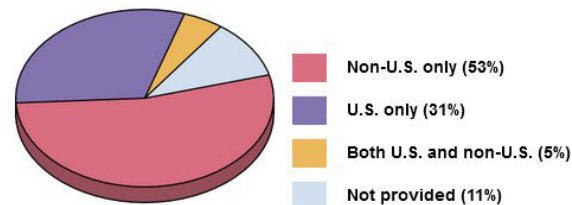
# Clinical Trial Search - Document Collection

## What is ClinicalTrials.gov?



## Percentage of Registered Studies by Location (as of July 02, 2023)

Total of 457,560 studies



Location	Number of Registered Studies and Percentage of Total (as of July 02, 2023)	
Non-U.S. only	244,005	(53%)
U.S. only	141,320	(31%)
Both U.S. and non-U.S.	22,390	(5%)
Not provided	49,845	(11%)
Total	457,560	(100%)

Source: <https://clinicaltrials.gov/about-site/about-ctg>

<https://classic.clinicaltrials.gov/ct2/resources/trends>

# Clinical Trial Search - Query and Ranking

**Focus Your Search**  
(all filters optional)

Condition or disease ⓘ

Other terms ⓘ

malnutrition in young children

Intervention/Treatment ⓘ

Location

Search by address, city, state, or country and select from the dropdown list

Study Status ⓘ

Looking for participants

☐ Not yet recruiting (8)
 ☐ Recruiting (20)

Clear Filters (1)

Apply Filters

Hide

<<

**Search Results**  
 Viewing 1-10 out of 203 studies

Selected (0)

Download

Manage Columns

	Study Title	NCT Number	Status	Conditions	Interventions
<input type="checkbox"/> 1	<a href="#">Integrated Research on Acute Malnutrition in Mali (IRAM-MALI)</a>	NCT04872088	Completed	<ul style="list-style-type: none"> <li>Acute <b>Malnutrition</b> in Childhood</li> <li>Wasting</li> </ul>	<ul style="list-style-type: none"> <li>Behavioral: Strengthened</li> <li>Dietary Supplement: Pre supplement</li> <li>Behavioral: Family MUAC</li> <li>4 more</li> </ul>
<input type="checkbox"/> 2	<a href="#">A Nutrition/Hygiene Education Program for the Prevention of Child Malnutrition in Rural Kenya</a>	NCT01679535	Completed	<ul style="list-style-type: none"> <li><b>Child</b> Nutrition Disorders</li> </ul>	<ul style="list-style-type: none"> <li>Behavioral: nutrition and n</li> </ul>
<input type="checkbox"/> 3	<a href="#">Community-based Follow-up of Severely Malnourished Children</a>	NCT01157741	Completed	<ul style="list-style-type: none"> <li><b>Malnourished Children</b></li> </ul>	<ul style="list-style-type: none"> <li>Behavioral: C-C</li> <li>Other: C-SF</li> <li>Other: As C-C with additi stimulation (PS) (C-PS)</li> <li>2 more</li> </ul>
<input type="checkbox"/> 4	<a href="#">Effectiveness of LNS and MNP Supplements to Prevent Malnutrition in Women and Their Children in Bangladesh</a>	NCT01715038	Completed	<ul style="list-style-type: none"> <li><b>Malnutrition</b></li> </ul>	<ul style="list-style-type: none"> <li>Dietary Supplement: LNS</li> <li>Dietary Supplement: LNS</li> <li>Dietary Supplement: MN</li> <li>1 more</li> </ul>
<input type="checkbox"/> 5	<a href="#">Clinical Study of Novel Probiotic Microbial Composite™ to Treat Undernourished Young Children</a>	NCT03150927	Unknown status	<ul style="list-style-type: none"> <li>Quality of Life</li> <li><b>Malnutrition</b></li> <li>Growth Arrest</li> </ul>	<ul style="list-style-type: none"> <li>Dietary Supplement: Pro composite</li> <li>Dietary Supplement: Plac</li> </ul>
<input type="checkbox"/> 6	<a href="#">Strategies to Increase Milk Consumption by Young Nepali Children</a>	NCT03886467	Completed	<ul style="list-style-type: none"> <li><b>Child Malnutrition</b></li> </ul>	<ul style="list-style-type: none"> <li>Other: Community-based on</li> </ul>

Source: <https://clinicaltrials.gov/search?term=malnutrition%20in%20young%20children&viewType=Table> (as of July 4, 2023)

4

## Key Takeaways

- MPSS can be used to perform clinical trial search across all disease classes
  - TREC Precision Medicine Track (2017 to 2020) focused only on cancer-related trials

# Key Takeaways

- Contribute an evaluation dataset of 25 queries with trials marked as relevant or non-relevant to the given query
  - Around 95 trials per query are annotated
  - Queries range over five-most frequent MeSH disease classes in document collection:
    - Pathological Conditions, Signs & Symptoms; Cardiovascular Diseases; Nervous System Diseases; Nutritional and Metabolic Diseases; Immune System Diseases

# Key Takeaways

- MPSS focus on ordinary users (patients or consumers) and thus the nature of our queries are free-form text
  - TREC PM track had a fixed schema, with fields like gene and mutation information, which is very specific to cancer-related trials

```
<topics task="2018 TREC Precision Medicine">
  <topic number="1">
    <disease>Acute lymphoblastic leukemia</disease>
    <gene>ABL1, PTPN11</gene>
    <demographic>12-year-old male</demographic>
  </topic>
  ...
</topics>
```

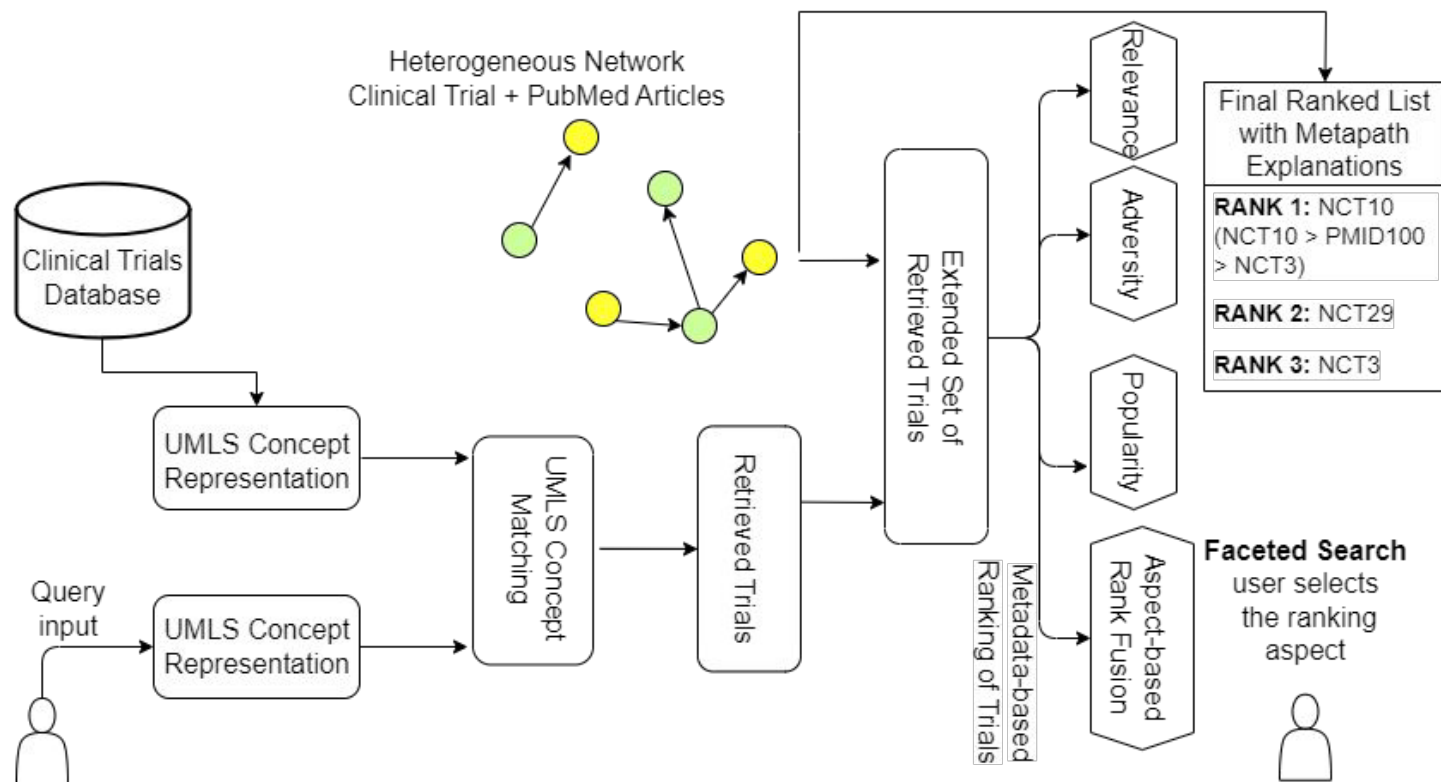
Source: <http://trec-cds.org/2018.html>

## Key Takeaways

- We construct a novel heterogeneous information network of both clinical trials and linked Pubmed articles to alleviate sparsity issue
  - Difficult to find direct links between two clinical trials
- Explore a path-based retrieval approach that becomes explainable to the end users
- Provide a combined ranking scheme based on relevance, adversity and popularity
- Develop strategy to adapt MPSS to the TREC PM task that deals with only cancer trials



# Detailed methodology of MPSS



# AACT-DB: Clinical Trials Corpus

- Consists of 331,713 clinical trials (May 2020 snapshot)
  - Made available through Clinical Trials Transformation Initiative,  
<https://aact.ctti-clinicaltrials.org/>
- Select trials with at least one linked Pubmed article
- Select trials from top five most frequent disease classes

<b>Trial ID</b>	NCT00000106	
<b>Brief Title</b>	Whole Body Hyperthermia for the Treatment of Rheumatoid Disease	
<b>MeSH Term</b>	Rheumatoid Diseases	Hyperthermia
<b>MeSH Tree</b>	C05.799	C23.888.119.455
<b>Disease class</b>	Musculoskeletal Diseases	Pathological Conditions, Signs and Symptoms

# Creation of Query-Relevant Trial Set

- 25 queries with 95 trials per query with relevance annotation
- Query Preparation: Five from each disease class
  - **Representative of real-life queries:** Templates from Patel et al. (2010) who observed it from user logs from TrialX search engine:

(2010). What do patients search for when seeking clinical trial information online?. In AMIA Annual Symposium Proceedings (Vol. 2010, p. 597). American Medical Informatics Association.

# Creation of Query-Relevant Trial Set

- 25 queries with 95 trials per query with relevance annotation
- Query Preparation: Five from each disease class
  - **Representative of real-life queries:** Templates from Patel et al. (2010) who observed it from user logs from TrialX search engine:
    - (Disease or syndrome) + (symptom or treatment): *dietary approaches for obesity treatment*
    - Disease + age group: *managing constipation in children*
    - Disease + safety information: *safe treatment for Alzheimer disease*

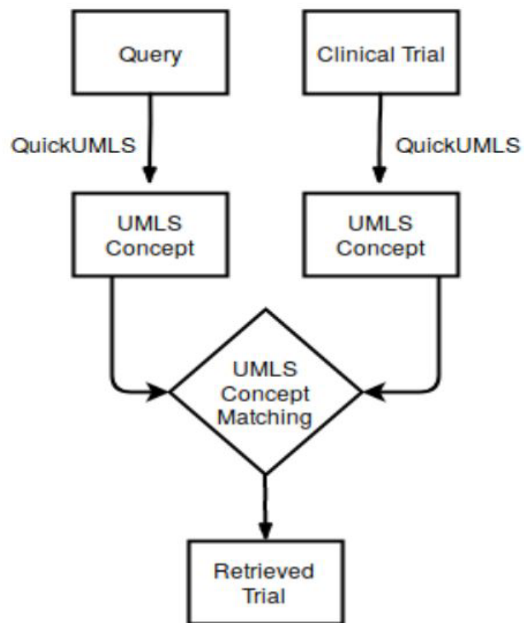
(2010). What do patients search for when seeking clinical trial information online?. In AMIA Annual Symposium Proceedings (Vol. 2010, p. 597). American Medical Informatics Association.

## Queries divided based on “safety” aspect

- **Type-1:** when query mention safety requirements, the search system should prioritize trials with no reported adverse events
  - Example: *safe treatments for asthma*
  - Develop manually curated lexicon set containing words like *safe, safety* and use exact matches to maintain high precision
- **Type-2:** all the remaining queries
  - Example: *haemorrhage cure, Early Parkinson disease treatment*

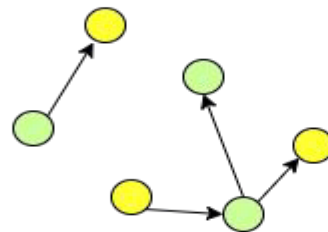
# Clinical trial retrieval: Match-based retrieval

For a query 'q', we retrieve all the trials whose (brief title + official title + brief summary) contain all the UMLS concept ids that are present in 'q'

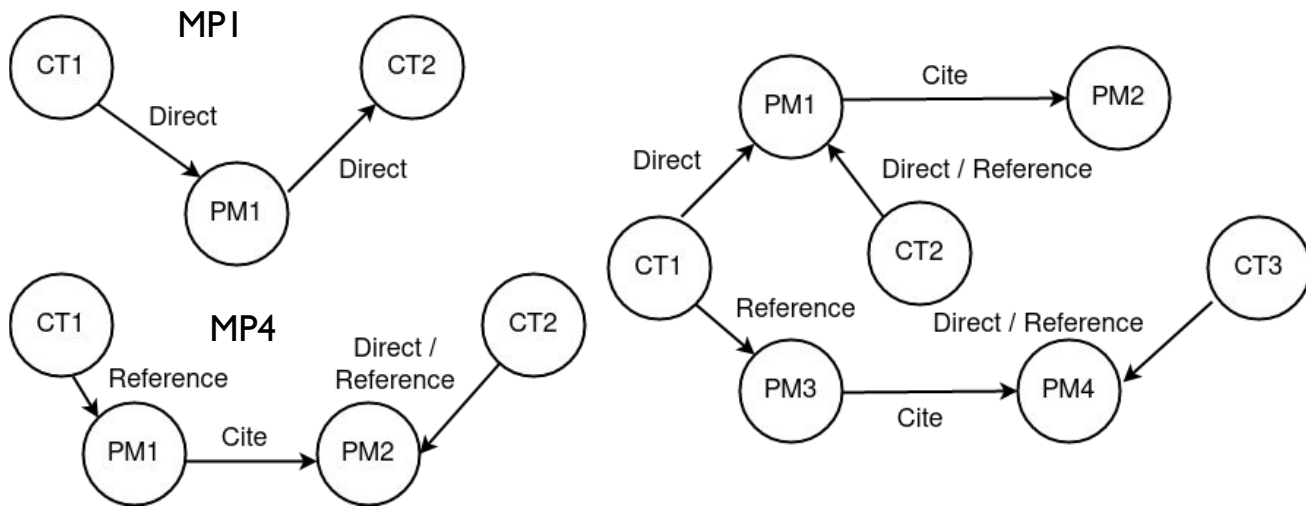


# Heterogeneous Information Network Construction

- Two node types: Clinical Trial (CT) and Pubmed article (PM)
- Three edge types:
  - **Direct:** Pubmed article published for a CT after its completion
  - **Reference:** PM acts as reference or result references of a CT
  - **Cite:** PM is linked to their citations to other PM's



# Heterogeneous Information Network Construction



Subgraph samples. Comprises of **750K nodes** and **1.2 million edges**.  
Clinical trial nodes are sparsely connected (8.36% nodes are of clinical trials, rest are Pubmed articles)



# Proposed Metapath-based Similarity Search

- One-time computation that computes the similarity set  $\text{SimSet}(\text{CT})$  for each clinical trial in the corpus
- Empirically determine maximum SimSet Size by balancing the coverage (recall) and quality of new retrieved trials (precision)

Restriction Type	MP1	MP2	MP3	MP4	MP5	MP6
Most restricted	✓	✓	✓		✓	
Moderate	✓	✓	✓	✓	✓	
Most Relaxed	✓	✓	✓	✓	✓	✓

# Multiple stakeholders = Multiple ranking aspects

- Explore the “faceted search paradigm”
- Individual ranking aspects: **Relevance, Adversity, Popularity**
- Aspect-based rank fusion to obtain a single ranked list (Variations)
  - **MetaRRF**: All ranking aspects are given same weightage
  - **MetaADV**: More weightage to adversity aspect
  - **MetaCOMB** (proposed): Use MetaADV for Type-1 queries and MetaRRF for Type-2 queries

# Adaptation to TREC 2018 Precision Medicine Track

Query	Disease: <i>melanoma</i> , Gene (Variant): <i>BRAF (V600E)</i> , Demographic: <i>64-year-old male</i>
Gene Name	BRAF (Entrez Gene Id: 673)
Variant	V600E
Gene Description	B-Raf proto-oncogene, serine / threonine kinase
Gene Synonyms from NCBI Gene	NS7, B-raf, BRAF1, RAFB1, B-RAF1
Interacting Drugs from DGIdb	Pictilisib bismesylate, panobinostat, binimetinib, oxaliplatin, fostamatinib, ...

# Adapting Relevance Ranking for TREC-PM task

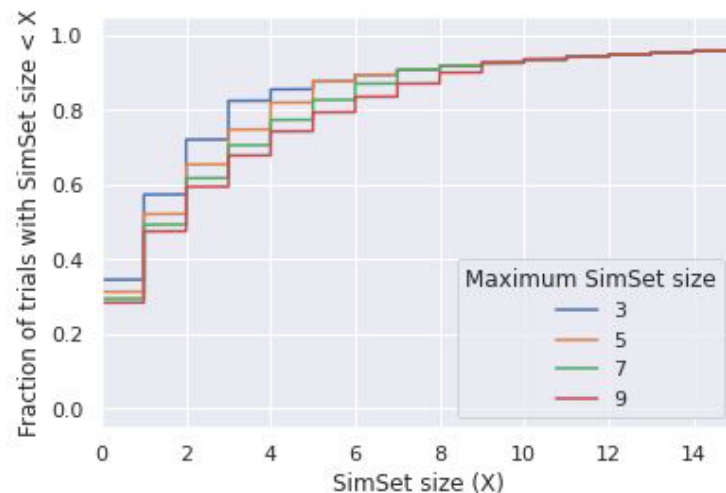
- Sort the trials in non-increasing order of “Gene relevance”, and continue with stable sorting in the following order
  - Term frequency of Mutation, Gene, Gene synonyms
  - PageRank score among retrieved set of trials

# Evaluation Setup

- Evaluation metric: Precision at ranks 5, 10, 15 and 20
  - Precision and nDCG score
  - Cannot measure recall due to lack of complete retrieved set of clinical trials for each query
- Evaluation datasets - MPSS is trained in unsupervised manner
  - Disease-independent Evaluation Dataset (25 queries)
  - TREC 2018 Precision Medicine Benchmark Dataset (50 queries)

# Performance Evaluation of Retrieval Stage

- Higher values of maximum SimSet size (maxSS) proportionately **improves recall**
- Precision@5 and @10 first increases with increase in maxSS, **peaks at maxSS=5**, and then **drops by 33.4% and 37% respectively at maxSS=7**



Query Type	Model	P@5	P@10	P@15	P@20
Type-1	MetaRRF	0.6	0.46	0.48	0.5
	MetaADV	0.96	0.92	0.89	0.77
Type-2	MetaADV	0.43	0.46	0.44	0.44
	MetaRRF	0.52	0.53	0.46	0.47
All	BAS	0.12	0.08	0.08	0.08
	STM	0.56	0.52	0.47	0.46
	MetaSTM	0.59	0.56	0.54	0.52
	MetaRRF	0.54	0.51	0.47	0.48
	MetaADV	0.54	0.55	0.53	0.51
	MetaCOMB	0.62	0.60	0.55	0.54

MetaCOMB outperforms the baselines models in terms of P@5, P@10, P@15 and P@20

# Performance Evaluation of Ranking Stage

- Aspect-based rank fusion does not reduce the quality of search
  - MetaCOMB performs comparably with MetaSTM (relevance only)
- Metapath-based similarity search helps improve clinical trial search performance
  - MetaSTM outperforms STM (its non-metapath version) at all ranks; the performance increases as we move from rank 5 to rank 20
- Query representation by UMLS concepts is more flexible than lexical matches and accommodates query variations



# Performance Evaluation of Ranking Stage

- Giving equal weightage to all ranking aspects irrespective of the type of queries fails
- MetaADV (more weightage to adversity aspect) achieves very high precision@10  $\sim 0.9$  for **Type-1 queries**, outperforming MetaRRF
- MetaRRF performs better than MetaADV for Type-2 queries

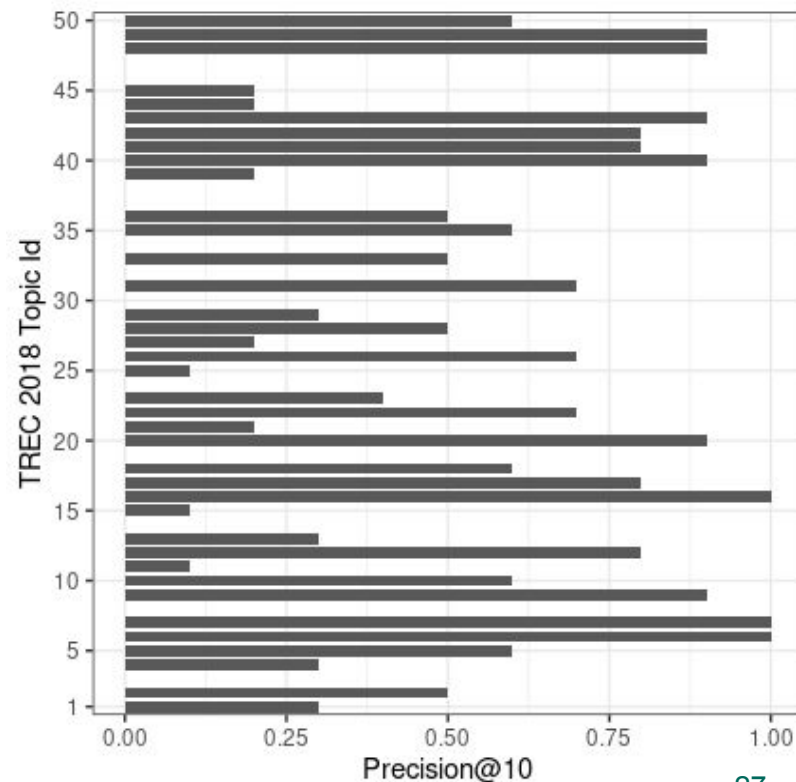
# TREC 2018 Adaptation Results

- MPSS achieves precision at ranks one, two and five of 0.5, 0.42, and 0.46 respectively
- MPSS is purely unsupervised and do not use 2017 TREC-PM training data
- MPSS does not utilize do not utilize disease-specific knowledge bases like COSMIC

Model	P@10	R-Prec	infNDCG
Cat_Garfield	<b>0.626</b>	<b>0.429</b>	<b>0.550</b>
ims_unipd	0.566	0.413	0.540
UTDHLTRI	0.538	0.368	0.479
MPSS	0.432	0.303	0.281

# TREC 2018 Adaptation Ablation Results

- Addition of drug interactions and gene-drug linked publications data improves Precision@10 by 9.75%
- Addition of adapted gene relevance further improved Precision@10 by 9.64%, thus achieving 0.432



# Conclusion

- Develop a metapath-based similarity search approach, MPSS, for clinical trial search across multiple disease classes
- Construct a heterogeneous information network of both clinical trials and linked Pubmed articles to alleviate the sparsity issue
- Explore the path-based retrieval approach that becomes explainable to the end users

# Conclusion

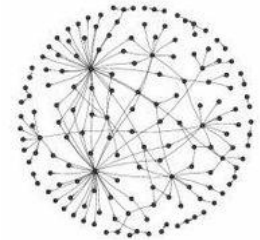
- Provide a combined ranked list based on relevance, adversity, and popularity
- Contribute an annotated (query-relevant trial) retrieval set for 25 queries (95 trials are annotated per trial on average) across five disease classes
- Evaluate MPSS in a zero-shot setting (without any task-specific training) on TREC 2018 Precision Medicine Track

## Code and Data Availability

- We make all the codes and data publicly available at <https://github.com/roysoumya/MPSS-clinical-trial-search>
- Specifically, we contribute a disease-independent evaluation dataset for clinical trial search systems that may encourage more research in this critical domain.

# Acknowledgements

- Soumyadeep Roy is supported by the Institute Ph.D. Fellowship at the Indian Institute of Technology Kharagpur
- Complex Networks Research Group, Department of Computer Science and Engineering, IIT Kharagpur  
(<https://cnerg-iitkgp.github.io/>)



# Acknowledgements

- Leibniz AI Lab, L3S Research Center, Leibniz University Hannover, Germany: For funding the conference author registration fees, travel and accommodation expenses (<https://leibniz-ai-lab.de/>)





Thanks for listening

Any Questions?

Please feel free to reach me at: [soumyadeep.roy9@iitkgp.ac.in](mailto:soumyadeep.roy9@iitkgp.ac.in)

To know about my current and past projects, please visit [datanalytics101.com/](https://datanalytics101.com/)