Interpretable Clinical Trial Search using Pubmed Citation Network

Soumyadeep Roy
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
soumyadeep.roy9@iitkgp.ac.in

Niloy Ganguly
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
niloy@cse.iitkgp.ac.in

Shamik Sural
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
shamik@cse.iitkgp.ac.in

Koustav Rudra
Department of Computer Science and Engineering
Indian Institute of Technology (ISM) Dhanbad
Dhanbad, India
koustav@iitism.ac.in

***Abstract*—Clinical trials are an essential source of information for practicing Evidence-Based Medicine because they help to determine the efficacy of newly developed treatments and drugs. However, most of the existing trial search systems focus on a specific disease (e.g., cancer) and utilize disease-specific knowledge bases that hinder the adaptation of such methods to new diseases. In this work, we overcome both limitations and propose a graph-based model that explores both clinical trials and the Pubmed databases to alleviate the shortage of relevant clinical trials for a query. We construct a large heterogeneous graph (750K nodes and 1.2 Million edges) made of clinical trials and Pubmed articles linked to clinical trials. As both the graph edges and nodes are labeled, we develop a novel *metapath-based similarity search* (MPSS) method to retrieve and rank clinical trials across multiple disease classes. We primarily focus on consumers and users that do not have any prior medical knowledge. As there are no multiple disease-wide trial search evaluation datasets, we contribute a high-quality, well-annotated query-relevant trial set comprising around 25 queries and, on average, approximately 95 annotated trials per query. We also perform a detailed evaluation of MPSS on the TREC Precision Medicine Benchmark Dataset, a disease-specific clinical trial search setting. We make all the codes and data publicly available at https://github.com/roysoumya/MPSS-clinical-trial-search .**

***Index Terms*—clinical trial, citation network, metapath-based similarity search, interpretability**

I. Introduction

The clinical trial search system stakeholders include medical professionals and patients or consumers; they satisfy their information needs through online trial search interfaces such as *Clinicaltrials.gov, WHO ICTRP, EmergingMed.com*, etc. Clinical trials are an important source of information for practicing Evidence-Based Medicine because they provide the earliest source of information about new drugs and treatments [1]. However, information needs significantly vary based on stakeholders. Patel et al. [2] observed that information needs for such ordinary users are related to a medical condition, location, and treatment. The primary focus of this work is ordinary users (patients or consumers).

Given a query as input from the user, the task of clinical trial search consists of two parts — (i) retrieval of relevant documents w.r.t a query from a clinical trial corpus (in our case, ClinicalTrials.gov), and (ii) ranking of the retrieved documents based on specific criteria. TREC Precision Medicine track of 2017-2020 ([3], [4]) focused on retrieval of oncology-related documents from evidence-based treatment literature and clinical trials (based on ClinicalTrials.gov data). More recently, TREC Clinical Trials tracks of 2021 and 2022 [5] focus on matching patients to relevant clinical trials; the main difference with the TREC Precision Medicine track is the nature of the query. The queries in the TREC Clinical Trials track are synthetic patient cases created by individuals with medical training in the form of an admission note. However, the TREC Precision Medicine track queries contain the disease, gene, variant, and demographic information in a fixed schema. However, in this work, we develop a trial search system that takes free-form medical queries such as *'dietary approaches for obesity treatment,' 'managing constipation in children.'* as input from the user (more details in Section III-B). The primary motivation behind making free-form queries is to make the trial search interfaces easy to use for ordinary end users [2].

There exist certain limitations of the prior works, broadly due to the following reasons — (i). **Focusing on specific diseases**: The TREC-PM task focuses on oncology-based trials; queries are quite specific in nature, requiring gene and variant

The definitive Version of Record is published in the 2023 IEEE International Conference on Digital Health (ICDH), https://doi.org/10.1109/ICDH60066.2023.00056. Soumyadeep Roy is also affiliated with the L3S Research Center, Germany.

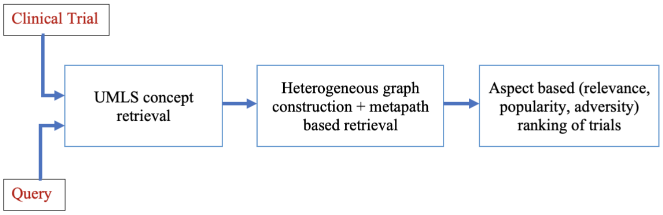
Clinical Trial
UMLS concept retrieval
Heterogeneous graph construction + metapath based retrieval
Aspect based (relevance, popularity, adversity) ranking of trials
Query

Fig. 1. **Methodological overview of MPSS.**

information within the query itself. Hence, methods proposed in this track heavily utilize cancer-specific knowledge bases like COSMIC [6] to expand queries and improve recall of retrieval systems. A drawback of such systems is that it fails to generalize to other disease classes and thus lead to the proliferation of disease-specific trial search engines like eTACTS [7] and Antidote [8]. Furthermore, their dependency on cancer-specific training data and knowledge bases hinders their adaptability to other disease classes; it disproportionately affects rarer MeSH disease classes like *Digestive System Diseases* and *Eye Diseases*. (ii). **Term mismatch between queries and trials**: Prior works [9], [10] rely on exact matching of UMLS concepts between queries and clinical trials. However, this depends on the terms specified by end users and suffers from sparsity effect and poor recall performance, i.e., many relevant clinical trials are missed due to a lack of exact matching in UMLS concepts. (iii). **Limitations in addressing multi-dimensional needs:** A stakeholder may have different information needs, i.e., some might be interested in relevant and popular trials, whereas some might be concerned about adversarial effects. However, existing trial retrieval systems primarily focus only on relevance. (iv). **Model interpretability:** Existing retrieval systems [11], [12] use the complex knowledge base and graph neural network (GNN) based methods to retrieve clinical trials. Unfortunately, such models are opaque in nature and do not provide any clues to the end-users about the ranking of the trials. Besides, post-hoc GNN explanation strategies have limitations [13], and this poses a challenge in their application to critical domains such as medicine.

To overcome the above limitations, in this paper, we propose a **M**eta**P**ath based **S**imilarity **S**earch (MPSS) method (Figure 1). We first extract the medical concepts for a given query using the QuickUMLS tool [14]. However, the retrieved trials are fewer in number. Hence, we collect the bibliographic information from the Pubmed database and develop an information network composed of two node types - *Clinical Trials* and *Pubmed* articles. A clinical trial and a linked Pubmed article are connected via reference patterns, i.e., if a Pubmed entry Y exists for a clinical trial X, we have an edge from $X \rightarrow Y$ (see Section IV-B). Finally, we perform a metapath-based search over this graph to retrieve relevant trials and rank them based on different aspects such as relevance, adversity, and popularity.

We evaluate our proposed method MPSS over 25 queries from five different disease classes. Experimental results suggest the efficacy of our proposed retrieval-ranking framework; we improve over the STM baseline model by 10.71% and 15.38% in terms of precision@5 and precision@10 metric, respectively. Further, we evaluate MPSS on the TREC 2018 Precision Medicine (PM) benchmark dataset and achieve a Precision@10 score of 0.432 as compared to the SOTA model (Team Cat_Garfield [12]) that achieves 0.626; MPSS also achieves a moderate recall performance of 0.588, highlighting the efficacy of the Pubmed-enhanced retrieval component of our proposed MPSS model. Our proposed model MPSS is more interpretable and shows the exact reason (query similarity, linked knowledge bases, or the Pubmed citation network) behind retrieving a trial for a given query. We further highlight the advantage of the unsupervised nature of MPSS and posit that it will thus allow the inclusion of rarer diseases. We thus show the generalizability of MPSS on the TREC 2018 Precision Medicine track that deals with only cancer-related queries. MPSS achieves a modest performance in an unsupervised manner as compared to the supervised SOTA models. We make the codes and data (including our constructed disease-independent evaluation dataset as described in Section III-B) publicly available at https://github.com/roysoumya/MPSS-clinical-trial-search.

II. Related Work

Here, we provide an overview of the recent clinical trial systems and the issues consumers or ordinary users face. We then discuss our proposed model in the context of similarity search techniques on heterogeneous information networks.

A. *Clinical Trial Search Systems*

Zuccon et al. [15] explore different strategies for developing knowledge-base-based consumer health search systems. Balaneshinkordan et al. [16] develop a Markov Random Fields-based retrieval model that jointly optimizes the weights assigned to statistical and semantic unigram, bigram, and multi-phrase concepts extracted from query and document collection. Team MedIER [17] utilizes medical ontologies and performs query expansion techniques for TREC 2017, and the same team develops a system [18] based on document re-ranking and query generation in TREC 2018. Balaneshinkordan et al. [19] also used medical ontologies like the Unified Medical Language System (UMLS), the Drug-Gene Interaction Database (DGIdb), and the Catalog of Somatic Mutations in Cancer (COSMIC), for query expansion purposes.

Trial search issues faced by ordinary users: Generally, ordinary users face difficulty in formulating useful queries because it involves complicated medical terminologies. Nunzio et al. [20] investigate a combination of query formulation strategies for the clinical trial search; both expansion and reduction techniques – based on knowledge bases to increase the probability of finding relevant documents. Additionally, the content of trial search sites like ClinicalTrials.gov contains technical content and is written at a considerably higher reading grade level than the average user (or consumer), hindering the accessibility to the trial information [21]. In this work, we focus only on ordinary users from non-medical backgrounds, such as patients and consumers, and develop the design of our clinical trial search system accordingly.

B. Similarity Search in Heterogeneous Information Network

Sun and Han [22] provide an overview of working with interconnected and multi-typed data, particularly operations like similarity search and structural analysis. A metapath-based similarity measure, PathSim [23], is used to find peer objects in a network and outperform random-walk-based similarity measures. HeteSim [24] can be used to compute relevance scores between nodes (in a network) of different types. Thilakaratne et al. [25] survey the computational approaches used for *Literature-Based Discovery*, which uses the connections among different entity types like paper, author, or venue in a bibliographic network setting, to detect implicit knowledge associations. Martin et al. [26] develop a web-based interface to assist medical professionals in updating systematic reviews that collates and augments information from bibliographic databases, Clinicaltrials.gov registry, and user actions. In this work, we contribute a novel heterogeneous information network constructed by linking clinical trials to Pubmed articles through direct publications and Pubmed articles used as study references. To the best of our knowledge, this is the first usage of the Pubmed citation network to improve the retrieval performance of a clinical trial search system.

III. Dataset

In this paper, we evaluate the performance of our proposed method (MPSS) over two different kinds of clinical trial datasets— (i). General clinical trial retrieval dataset (queries and relevant trials for five different diseases), and (ii). TREC Precision Medicine Track 2018 for search over 'Neoplasms' disease class made of 50 queries. Here, we describe the document collection for the clinical trial search task, followed by the details of the above two datasets.

A. Clinical Trial Corpus

We use the dump of AACT database [27] (AACT-DB) dated May 2020. It consists of around 331, 713 clinical trials. In this study, we focus on utilizing the Pubmed bibliographic database to improve the retrieval performance of a trial search system. The *Food and Drug Administration Amendments Act* (FDAAA) mandates timely reporting of results of applicable clinical trials to ClinicalTrials.gov [28]; these results are thus made publicly available as Pubmed articles in certain cases. We select trials with at least one linked Pubmed article. Medical Subject Headings [29] (MeSH) is used for indexing the PubMed database. It also should have at least one MeSH term, and the trial is not ongoing (trial status is completed, terminated, suspended, or withdrawn). The trials are then mapped to all the 26 MeSH disease classes. To map a trial to its respective disease class, we first extract the MeSH terms corresponding to a clinical trial from the *browse conditions* table of AACT-DB, which we then match with the MeSH thesaurus (tree-like hierarchy). Table I provides an example of mapping a clinical trial to MeSH disease classes. We observe that a disease is present at the root of a tree in most cases. Thus, a clinical trial is mapped into one or more disease classes (out of 26 in total).

TABLE I
Mapping a clinical trial to MeSH disease classes

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

In this paper, we focus on the top five most frequent disease classes present in the above-mentioned clinical trial corpus (the disease classes are outlined in Table II). We avoid including rarer classes due to the paucity of ground-truth data to evaluate them. We focus on frequent classes, which demand significant human effort for annotation. It results in a total of 67237 trials. We do not consider trials that belong to the 'Neoplasms' disease class because it is already covered in the TREC Precision Medicine track [3], [4] and typically consists of more sophisticated queries which include gene and mutation information.

B. Construction of Disease-independent Evaluation Dataset

Here, we describe the construction of the query-relevant trial set for performance comparison of MPSS with baseline trial search systems. We aim to make the queries both generalizable as well as representative of real-life user queries on trial search engines. First, we select queries from five different disease classes to maintain the generalizability of the results as opposed to the TREC Precision Medicine task that focuses only on oncology-related queries. Second, we obtain semantic-based query templates identified by Patel et al. [2] based on the user (query) logs of the *TrialX* search engine; the most frequent user query template is *disease or syndrome* + *research activity*. We select the following subset of templates for preparing the queries: (i). (disease or syndrome) + (symptom or treatment) (such as *dietary approaches for obesity treatment*), (ii). disease + age group (like *managing constipation in children*), (iii). disease + safety information (like safe treatment for Alzheimer disease). The frequently-used query templates ([3], [4]) consisting of location and gene information are not considered for this work. We also consult a patient vocabulary-based lexicon called MedDRA [30]. We select five queries for each of the five disease classes (outlined in Table II) for evaluating the performance of MPSS.

We next outline the steps for determining the 25 query sets we use for evaluation purposes. First, we obtain all the diseases that fall under each disease class using the MeSH hierarchy for each of the five disease classes. Second, we apply the templates mentioned above to generate all possible queries (construction of candidate queries). We use two rules for filtering the candidate queries: (i). one query is generated based on a given disease name. (ii). a query should retrieve at least 10 trials using the model developed by Throve [10] (baseline model) in the TREC 2017 Precision Medicine Track [3] in order to avoid rare queries w.r.t the clinical trial corpus. It performs Elastic Search for clinical trial retrieval and then uses the Okapi BM25 as the ranking function.

We next divide the queries into two types based on the *safety* aspect.

1) **Type-I**: When the query mentions the safety requirement, the ranking algorithm should prioritize trials having no reported adverse events. The queries are like *constipation safe treatments, hypertension safe treatments, safe treatment for Alzheimer disease, safe treatments for asthma*. When constructing such queries, we use a manually curated lexicon set consisting of safety-related words like *safe, safety*. We use such a lexicon-based technique because it ensures high precision. Sophisticated approaches might exist to detect such terms. It will help further boost the performance of our proposed approach. However, extensive coverage of such terms is beyond the scope of this work.
2) **Type-II**: This includes all the remaining queries, not Type-I, such as *haemorrhage cure, Early Parkinson disease treatment, Treating Anemia Iron-Deficiency in CKD patients*.

We keep one Type-I query for each disease class while the remaining four are Type-II (a total of 5 queries). Three annotators label the retrieved trials for each of the 25 queries; none of them is the author of this paper and knows good English. We strictly adhere to the annotation scheme ('Definitely Relevant' category) introduced in the TREC Precision Medicine 2018 task [4]. Around 95 trials are annotated on average per query. Here, the annotators are asked to mark whether each trial of the retrieved trial set is relevant to a given query.

C. TREC Precision Medicine Benchmark Dataset

We also test the generalizability and disease-independence capability of MPSS by evaluating a benchmark dataset from the 2018 TREC-PM Track [31]. The task focuses on searching oncology disease-related clinical trials over the ClinicalTrials.gov database and provides 50 query-relevant trial sets for evaluation purposes. Oncology belongs to the MeSH [29] disease category of *Neoplasms*.

Here, the **query** follows a pre-defined schema with the following category descriptors - <disease>, <variant (gene and mutation)>, <age>, and <gender>. For each query, a set of relevant trials manually annotated by medical experts are provided, which acts as the ground truth in our case. We consider trials with query relevance marked as 'Partially Relevant' or 'Definitely Relevant' as relevant (binary) for this study. We further observe that 16% and 20% of queries contain less than 5 and 10 relevant trials in the ground truth, respectively; this further highlights the strong difficulty level of the clinical trial search task. Regarding the **document collection**, we focus only on clinical trials that belong to *Neoplasm* disease class, which leads to 7398 trials, out of which 54.2% is linked to a PubMed article.

TABLE II
QUERY LIST OVER FIVE MOST FREQUENT MESH DISEASE CLASSES

Disease Class	**Query**
Pathological Conditions, Signs & Symptoms (PAT)	constipation safe treatments, haemorrhage cure, low back pain therapy workout, postoperative delirium, managing constipation in children
Cardiovascular Diseases (CVD)	hypertension safe treatments, treating people already having hypertension, recommended antiplatelet doses for treating Coronary artery disease, out of hospital cardiac arrest, Nonvalvular atrial fibrillation
Nervous System Diseases (NER)	Dietary Therapy Epilepsies, safe treatment for Alzheimer disease, serious sleep apnea, Outcomes of cerebrovascular accident, Early Parkinson disease treatment
Nutritional and Metabolic Diseases (NMT)	dietary approaches for obesity treatment, Treating Anemia Iron-Deficiency in CKD patients, Hypercholesterolemia safe treatments, malnutrition in young children, already having Celiac Disease
Immune System Diseases (IM)	safe treatments for asthma, antiretroviral therapy first time, serious Rheumatoid arthritis, HIV infection Treatment naive, HIV infection seronegativity

IV. MPSS METHODOLOGY

The detailed methodological overview of the proposed *metapath-based similarity search* (MPSS) is presented in Figure 2. In this section, we elaborate on the *retrieval* and *ranking* components of the MPSS.

A. Clinical Trial Retrieval

Here, we describe the retrieval components of the proposed clinical trial search system, MPSS. It comprises two components - query concept extraction, followed by match-based retrieval.

1) **Query Concept Extraction:** An unsupervised, scalable medical concept extraction tool QuickUMLS [14], is used to extract UMLS medical concepts from a query. For example, the query '*malnutrition in young children*' contains two medical concepts: (i) child malnutrition (semantic class: *disease or syndrome*), and (ii) young (semantic class: *temporal concept*).
2) **Match-based Retrieval:** A set of UMLS medical concepts now represents each query. We follow the same methodology to represent a trial as the set of UMLS medical concepts extracted from its *brief title* and *brief summary* fields. We adopt a conservative approach and retrieve the clinical trials (brief title and summary fields) containing all the UMLS concept ids present in query q. We term this **S**imple **T**erm **M**atching based algorithm as STM. We keep the matching constraint tighter to take care of false-positive trials. However, this, in turn, may affect the recall part, i.e., STM can retrieve only a few trials per query. To circumvent that issue, we extend the initial retrieval set using a *metapath-based similarity*

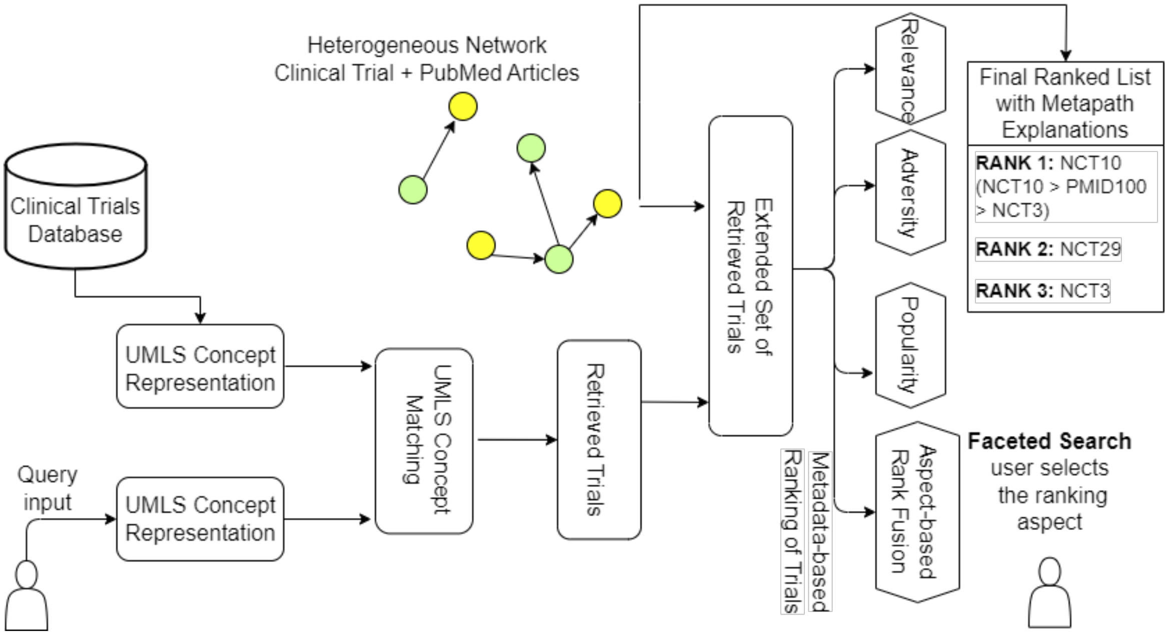
Heterogeneous Network
Clinical Trial + PubMed Articles
Clinical Trials Database
UMLS Concept Representation
Query input
UMLS Concept Representation
UMLS Concept Matching
Retrieved Trials
Extended Set of Retrieved Trials
Metadata-based Ranking of Trials
Relevance
Adversity
Popularity
Aspect-based Rank Fusion
Final Ranked List with Metapath Explanations
RANK 1: NCT10 (NCT10 > PMID100 > NCT3)
RANK 2: NCT29
RANK 3: NCT3
Faceted Search
user selects the ranking aspect

Fig. 2. **Methodological overview of MPSS. A user enters a query in free-form text, and medical concepts are extracted from the query, which is then used to retrieve relevant trials for the clinical trials registry database. Our proposed metapath-based similarity search algorithm that uses the Pubmed bibliographic network is used to improve the retrieval performance of MPSS. We then introduce three meta-data-based ranking aspects of relevance, adversity, and popularity, as well as a single ranked list combining all the aspects through aspect-based rank fusion. MPSS follows a faceted search paradigm where the user is given the option to select any one (among four) ranking aspects based on the user's information need. We present the metapath as explanations in the final ranked list in case of the additional trials retrieved using Pubmed-enhanced retrieval, which makes the trial search results more explainable**

search. We describe our approach in detail in the next section.

B. Pubmed-enhanced Retrieval using Metapath-based Similarity Search

This section consists of two major components — (i). Building a heterogeneous information network of clinical trials and PubMed articles and (ii). Developing a metapath-based similarity search algorithm to retrieve relevant clinical trials for a given query.

1) Construction of Heterogeneous Information Network: We first describe the nodes and edge details of our heterogeneous information network. It consists of two types of nodes: clinical trials (CT) and Pubmed articles (PM), and three types of edges: direct, reference, and cite. The different edge types are described as follows:

1) **Direct:** It corresponds to the Pubmed article(s) published for a clinical trial after its completion. In Pubmed, registry numbers are included in the *Secondary Source ID (SI)*[1] field. It is bidirectional, and thus both the links $\mathbf{CT} \rightarrow \mathbf{PM}$ and $\mathbf{PM} \rightarrow \mathbf{CT}$ exist. It forms a many-to-many relationship; for example, one CT is linked to multiple PMs (NCT00000542 to 30906106, 30590387) and one PM to multiple CTs (30890109 is linked to NCT00226096 and NCT00716079).

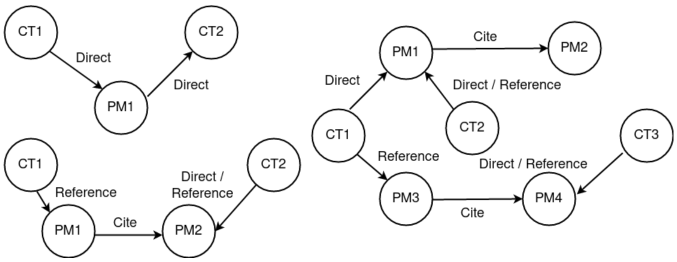

Fig. 3. Subgraph sample of heterogeneous information network. CT stands for Clinical Trials, and PM stands for Pubmed articles.

2) **Reference:** It arises when Pubmed articles act as references or references for the results of a trial. This information is directly available in *study reference* table of AACT-DB. It is unidirectional and corresponds to $\mathbf{CT} \rightarrow \mathbf{PM}$.
3) **Cite:** Pubmed articles are linked to their citations that are also Pubmed articles. We obtain the citations using Europe PMC's RESTful API[2], and only consider the top fifty most-cited papers as citation links to form edges in the heterogeneous information network being constructed. It is unidirectional and corresponds to $\mathbf{PM} \rightarrow \mathbf{PM}$.

Thus, in essence, we use the Pubmed bibliographic network that contains around 750K nodes and 1.2 million edges.

[1] https://www.nlm.nih.gov/bsd/mms/medlineelements.html#si

[2] https://europepmc.org/restfulwebservice

TABLE III
METAPATH TYPES ARRANGED IN DESCENDING ORDER OF THE DEGREE OF STRENGTH BETWEEN SOURCE AND TARGET CLINICAL TRIAL (MP1 IS STRONGEST AND MP6 IS WEAKEST)

Metapath Type	Condition
MP1	$CT \xrightarrow{direct} PM \xrightarrow{direct} CT$
MP2	$CT \xrightarrow{direct} PM \xleftarrow{reference} CT$
MP3	$CT \xrightarrow{reference} PM \xrightarrow{direct} CT$
MP4	$CT \xrightarrow{direct} PM \xrightarrow{cite} PM \xleftarrow{direct/reference} CT$
MP5	$CT \xrightarrow{reference} PM \xleftarrow{reference} CT$
MP6	$CT \xrightarrow{reference} PM \xrightarrow{cite} PM \xleftarrow{direct/reference} CT$

Essentially, the Pubmed articles act as a bridge connecting the clinical trials. Thus we utilize the Pubmed bibliographic network to form connections between two trials and assign the degree of similarity based on the sequence of edge types (elaborated in the next paragraph). These clinical trials are usually poorly connected; therefore, we observe the presence of 91.64% of Pubmed articles. The largest component comprises 93.39% of all the nodes and 97.03% of all the edges. 20559 clinical trials are isolated, i.e., they form singleton components. Fig. 3 shows possible graph connections.

2) Metapath-based Similarity Search to Improve Retrieval Performance: After developing the information network, we perform metapath based graph traversal to retrieve relevant trials. Table III shows the different metapath variants. It is arranged in decreasing order of relationship strength between the source and target clinical trials. Here, we denote the trial for which similar trials are being searched for as *source* trial, and the retrieved candidate trials with which the source trial is being matched as *target* trial.

Metapath-based Similarity Search Algorithm. Here, we develop an algorithm that computes the similarity set for each trial $SimSet(CT)$ in an offline manner, which means that the computation is carried out only once and does not change during the retrieval period. When a new query is given, we retrieve the relevant trials using UMLS concept mapping and then expand that set using precomputed $SimSet(CT)$. We formally define $SimSet(CT)$ as *the number of nodes of type 'CT' within two hops irrespective of the type of metapath linking it*. We first introduce the three types of restrictions that we use in our proposed metapath-based similarity search algorithm:

1) **Most Restricted**: For a given CT_q, we retrieve precomputed $SimSet(CT_q)$, but we do not allow any CTs that are connected to CT_q via $edge-type : citation$, i.e., MP4 and MP6 metapaths. However, if the number of retrieved trials is less than $maxSS$, we will use a moderate condition.

2) **Moderate**: Here, we maintain the same two-hop condition as before but allow MP4 and drop only trials connected via MP6. However, still, for trials for which the number of retrieved trials is less than $maxSS$, we will impose a most relaxed condition.

3) **Most Relaxed**: Here, we consider all edge types to be the same and do not discard any trials based on their $edge-type$.

We do not assign any similarity to the heterogeneous graph's isolated clinical trial nodes (whose component size is one). Through the above-mentioned three conditions, we try to balance the coverage (recall) and quality of the new retrieved trials (precision). We achieve that balance by empirically determining the maximum number of similar trials obtained for each trial (Maximum SimSet size, *maxSS*) as *five* (see section V-B for empirical results).

C. Aspect-based Ranking of Clinical Trials

After retrieving the relevant trials, the next task is to rank them based on the aspect (*Relevance, Adversity* and *Popularity*) provided by the user. We follow the approach of faceted search [32], where the user can select the ranking aspects and thus provide a form of filtering mechanism to help narrow down the search results. Users may opt for a single ranked list automatically constructed by aspect-based rank fusion combining all three ranking aspects if the users do not need any specific aspect. Next, we describe individual ranking aspects followed by two novel aspect-based rank fusions.

1) Defining Individual Ranking Aspects: The different ranking aspects are defined as follows:

1) **Relevance**: An undirected graph *G(V, E)* is first constructed using the clinical trials retrieved for a given query as vertices. The edge weights between (V_i, V_j) vertices are computed as the Szymkiewicz–Simpson coefficient [33] between clinical trials in terms of UMLS concepts extracted from *brief title* and *brief summary* fields of a clinical trial. We then apply PageRank [34] algorithm on graph G (comprises around 100 nodes on average since for each trial from the retrieved trial set, 5 more similar trials based on $SimSet(CT)$ are retrieved on average. As reported in Section III-B, we annotate all the retrieved trials for each query, leading to the annotation of 95 trials per query. With Pubmed-enhanced retrieval, five additional trials on average are retrieved. This leads to a total of 100 trials retrieved per query, and we perform PageRank over these trials and determine the importance of a given trial based on the PageRank score. We now compute the total count of the presence of all the terms (present in a synset) in the brief summary and brief title fields of a trial and use it as a measure to determine the trial relevance w.r.t a query. We first rank the trials in the decreasing order of the brief summary count, followed by decreasing order of the brief title count and PageRank score for tie-breaking.

2) **Adversity**: After retrieving the relevant trials, the trials are first ranked in non-decreasing order in terms of the number of subjects affected (using the *Subjects Affected* field). Trials with no reported adverse events (i.e., *number of subjects affected* is equal to zero) are placed at the top of the ranked list.

3) **Popularity**: We map each clinical trial to the linked Pubmed articles (one-to-many relationship) and finally determine its corresponding citation count; we sum the citations in case a trial is mapped to multiple Pubmed articles. We use the REST API [35] service provided by *NCBI E-utilities* for this task. The retrieved clinical trials are first ranked in decreasing order of citation count (popularity value) and use the *Relevance* ranking aspect for tie-breaking.

2) Aspect-based Rank Fusion to obtain a Single Ranked list: Although multiple ranked lists may be optimal for some users, it may be helpful also to provide a single ranked list that combines all the aspects in case a user does not have a specific ranking aspect in mind. In the single aspect rankings, we do not focus on the query assignment, i.e., whether the user is interested in the trial safety (Type-I queries). Based on this requirement, we propose three different variations of MPSS.

1) **METARRF**: Here, the user does not provide the query type (i.e., Type-I or Type-II); hence, all the ranking aspects are given equal weights using an unsupervised rank fusion method called *Reciprocal Rank Fusion* (RRF) [36] to produce a single ranked list. Formally, given a set of D documents to be ranked and a set of rankings R, we determine the combined 'RRF-based relevance score' for each document $d \in D$ as:

$$RRF - Score(d) = \sum_{r \in R} \frac{1}{k + r(d)} \qquad (1)$$

where $r(d)$ is the rank of document d in ranking r, and k is a parameter intended to reduce the impact of low ranks on the score (in our experiments, we used k as equal to the minimum value between the total number of retrieved trials and 60 (the value recommended by authors of the original paper [36]).

2) **METAADV**: Here, the adversity aspect is given the most importance, whereas relevance and popularity are given the same weight. We first rank the retrieved trials based on the adversity aspect. Then, for the ones that do not have any adverse events reported, we rank them in descending order of their respective 'RRF-based relevance score.'

3) **METACOMB**: We combine the power of both METARRF and METAADV. For Type-I queries, we deploy METARRF over the clinical trials, and METAADV is applied for the rest of the queries.

D. Adaptation of MPSS *for TREC 2018 Precision Medicine Track*

As mentioned in Section I, our objective is to design a retrieval framework independent of any specific disease class. However, the precision medicine track (PM) contains cancer-related trials that cover other information such as genes, mutation, etc. Section III-C further explains the task objective, query, and problem setup in a detailed manner. Hence, we modify the retrieval and ranking component of the MPSS approach to make it applicable to TREC PM dataset.

TABLE IV
AN EXAMPLE OF EXTRACTING GENE SYNONYMS AND INTERACTING DRUGS, GIVEN *gene name* FROM QUERY

Query	Disease: *melanoma*, Gene (*Variant*): *BRAF (V600E)*, Demographic: *64-year-old male*
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, ...

1) Pubmed-enhanced Retrieval.: Table IV shows the query processing steps utilizing various biomedical databases for retrieving the trials relevant to a given query. The following steps are executed to retrieve the trials:

1) We filter the trials from the entire corpus based on *age* (numerical value) and *gender* fields. Instead of free-flowing textual queries, TREC 2018 queries are structured into three parts such as disease, gene, and demographic, as depicted by an example in Table IV.
2) We perform concept-based matching (i.e., overlap in terms of UMLS medical concepts) on *disease* field with *conditions* field of a clinical trial. The concept-based matching is performed with QuickUMLS [14].
3) We then use regular expression-based matching to separate the gene and mutation information from the 'variant' field. The *Gene* field of a query mainly contains genetic variant or mutation information.
4) We further improve the retrieval performance (in terms of recall) using a metapath-based similarity search, which leads to the addition of trials from the "Neoplasms" MeSH disease class (4013 out of a total of 7398 'Neoplasms' trials are linked to a Pubmed article) to our previously constructed heterogeneous information network (see Section IV-B1).
5) We also introduce a new metapath type where we connect two trials with a common Pubmed article as a reference or result reference. This helps to mitigate the sparsity issue, which is further heightened as the task focuses on a single disease class of 'Neoplasms.'
6) We utilize the Drug-Gene Interaction Database [37] (DGIdb) for identifying Pubmed articles that report drug interactions with a specific gene (identified by Entrez Gene Id [38]). We then use the constructed heterogeneous information network to extract clinical trials having a *Direct*-type link with such Pubmed articles (based on drug-gene interaction). This further improves the retrieval performance of MPSS, particularly recall.

2) Ranking Based on Relevance.: We first rank the trials in non-increasing order of Gene Relevance (based on similarity to the 'Gene' field of a query), followed by individual term frequency terms for Mutation, Gene, and Gene synonyms (see Table IV), and finally by PageRank score (based on the importance of a trial among the retrieved set in terms of trial metadata).

1) **Computing Gene Relevance:** We assign a gene relevance score as **two** when both gene and variant information match and **one** when only the gene matches without the variant information. To perform a **gene match**, we first perform UMLS concept-based matching using QuickUMLS between the 'Gene' field of a query and trial metadata (*brief title, brief summary, detailed description,* and *eligibility criteria*). A gene match is positive if at least one overlapping UMLS concept exists. We obtain gene synonyms if the gene match is negative using the NCBI [39] Gene API. We then perform an exact term match of the gene (including gene synonyms) with the trial metadata. If there is a positive gene match, we then perform a **gene variant match** which simply involves an exact term match with the 'Variant' part of the 'Gene' field with the same trial metadata as used during the gene match. Since most gene and mutation information are single tokens, we also compute term frequency terms for a gene, its gene synonyms, and mutation.
2) **Computing PageRank score:** Given the set of retrieved trials, we use them to construct an undirected graph and then apply the PageRank algorithm. We follow exactly the same strategy as previously defined in Section IV-C1. Here, a key difference is that the retrieval is based on quite general fields of a query, such as *Disease* and *Demographics*. Thus, on average, the retrieved trial set's size is significantly larger than our constructed dataset (around 100 trials per query). Therefore, we add a higher similarity threshold (Szymkiewicz–Simpson or Overlap coefficient > 0.4) while forming edges in the PageRank graph.

V. Experiments and Results

In this section, we first describe our experimental setup, i.e., baselines and evaluation setup. After that, we perform the evaluation of retrieval and relevance-based ranking framework. Finally, we also discuss the performance of MPSS on the TREC 2018 Precision Medicine benchmark dataset.

A. Experimental Setup

This section describes the baselines and metrics used for evaluation.

1) Baseline Models: We only consider unsupervised approaches due to the lack of training data. We explore the state-of-the-art search systems for 'TREC-PM 2017 Task-B' that focus on clinical trials. We observe that they either do not publish well-documented codebases or use cancer-specific ontologies like COSMIC [40], making it unsuitable for developing disease-independent trial search systems. We consider the following baselines:

1) **BAS:** Proposed by Throve [10] in the TREC 2017 Precision Medicine Track [3]. It uses ElasticSearch API for the retrieval task where the following fields of clinical trials - brief title, brief summary, and then use the Okapi BM25 as the ranking function.
2) **STM:** UMLS concept based approach proposed by [9].
3) **METASTM:** We update the STM method by adding the metapath based search results.

To the best of our knowledge, these systems do not use any disease-specific knowledge bases (particularly oncology-based trials) and have made their codebase publicly available.

2) Evaluation Setup: We measure the precision score at 5, 10, 15, and 20 for the entire query set. Since the complete ground truth set of clinical trials for each query is not available, it is not possible to measure *recall*. For the TREC 2018 Precision Medicine task, we use the same test set of 50 queries and report the same official evaluation metrics of Precision@10, R-precision, and infNDCG, when compared with the SOTA models [12], [20], [41] for a fair comparison.

B. Performance Evaluation of Retrieval Stage

We evaluate the proposed metapath-based similarity search formulation where each clinical trial is strongly associated with a list of the five most similar trials. We use the proposed clinical trial similarity search method to retrieve more trials. We now **investigate whether this similarity search improves the retrieval performance and mitigates the sparsity issues introduced by the clinical trials graph**. We mainly target the queries where the number of trials retrieved is less than 21. On manual inspection, we observe that the additional trials added to the original retrieved set of trials using the metapath-based similarity search are relevant to the query. This attests to the maintenance of the 'quality of search aspect.' We achieve 35% coverage improvement of METASTM over STM. Introducing a heterogeneous information network improves recall by 10% on the TREC-PM task. The improvements at maxSS=3, 5, 7, and 9 are 42%, 59%, 86%, and 92% respectively, as shown in Figure 4.

We incorporate the '*quality of search*' aspect while developing the retrieval component of a clinical trial, empirically determining the maximum number of similar trials obtained for each trial (maxSS) as *five*. The precision at ranks 5, 10 of MPSS with threshold value seven drastically drops by 33.4% and 37% respectively as compared to the performance of MPSS with threshold $five$. Therefore, maxSS=5 helps to maintain a balance between recall and precision.

C. Performance Evaluation of Rankings based on Relevance

We compare the rankings of METARRF, METAADV, METACOMB in terms of mean precision values computed based on relevance (we do not have ground truth to evaluate the other ranking aspects) at different ranks of 5, 10, 15 and 20, as shown in Table VI. We incorporate the query type information through the final model MPSS, where we identify queries that highlight the safety aspect of a trial. **METACOMB outperforms all the competing baseline models in mean precision values at all ranks (5, 10, 15, and 20)**. We further observe that the performance improvement between METASTM and METACOMB is evident based on

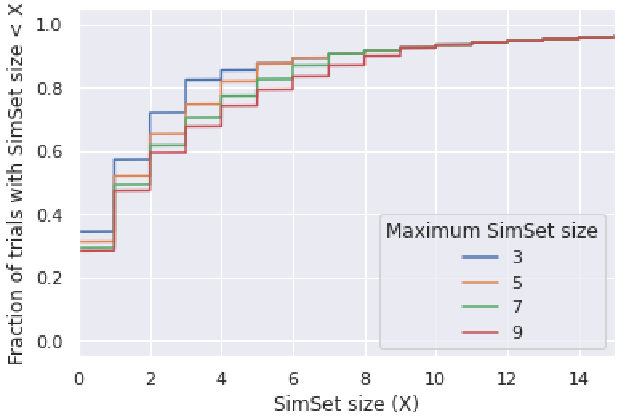

Fig. 4. CDF plot of SimSet(CT) distribution for different Maximum SimSet size (maxSS) values

TABLE V
RETRIEVAL PERFORMANCE COMPARISON FOR DIFFERENT MAXIMUM SIMILARITY SET THRESHOLD VALUES (K) OF MPSS. $k = 0$ IS SAME AS STM (WITHOUT METAPATH) MODEL.

Query	**Maximum Similarity Set Size**				
	0	3	5	7	9
already having Celiac Disease	19	23	24	24	24
constipation safe treatments	13	18	25	25	25
Dietary Therapy Epilepsies	9	15	20	23	23
HIV infection Treatment naive	12	23	25	37	43
Hypercholesterolemia safe treatments	17	21	24	30	30
safe treatment for Alzheimer disease	16	27	27	30	30
Treating Anemia, Iron-Deficiency in CKD patients	15	16	16	19	19

their absolute value, but they are not statistically significant. This is due to the limited ground truth data comprising only 25 points and thus requires a larger performance difference to achieve statistical significance. This result is interesting because METACOMB (that gives weight to aspects other than relevance) performs comparably with a relevance-only model (METASTM) and strongly indicates that the aspect-based rank fusion approach does not reduce the quality of search.

METASTM outperforms STM (its non-metapath version) at all ranks; the difference increases as we move from rank 5 to rank 20. We also observe that the absolute value of performance improvement in terms of mean precision value is least in the case of Precision@5 and maximum for Precision@20. This implies that the previous methods work well when the retrieved trials are highly similar to the query but fail to accommodate the borderline cases where the trials are not directly similar but still relevant to the query. We observe that METARRF performs poorly compared to STM. This indicates that the simple strategy of assigning equal weights to the different aspects fails. This indicates that: (i) some aspects are more important than others and hence should be given more weight, (ii) it depends on the type of query, or (iii) both are possible. Intuitively, we incorporate the 'adversity' aspect as the primary aspect during a clinical trial search and make a trade-off with the relevance of a trial. This may consequently harm search performance. However, through the METAADV model, we perform comparably with STM and additionally incorporate the adversity aspect for ranking the trials.

TABLE VI
PERFORMANCE COMPARISON OF RELEVANCE RANKING MODELS. WE REPORT PAIRED T-TESTS TO CHECK FOR STATISTICAL SIGNIFICANCE. *** INDICATES P-VALUE < 0.001, ** FOR P-VALUE < 0.01, * FOR $0.01 <$ P-VALUE < 0.05

Query Type	**Model**	**P@5**	**P@10**	**P@15**	**P@20**
Type-I	METARRF	0.6	0.46	0.48	0.5
	METAADV	**0.96**	**0.92**	bf 0.89	**0.77**
Type-II	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	$0.52^{0.06}$	0.47*	0.46**
	METASTM	0.59	0.56	0.54	0.52
	METARRF	$0.54^{0.07}$	0.51*	0.47*	$0.48^{0.07}$
	METAADV	0.54	0.55	0.53	0.51
	METACOMB	**0.62**	**0.60**	**0.55**	**0.54**

Finally, we perform a query-wise analysis of precision and nDCG@20 values. We observe that STM shows significant improvement for 10 such queries. However, STM achieves a precision@10 value of less than 0.31 for 28% cases because of the limitation during the retrieval stage. For 3 out of 25 queries, BAS retrieves at least five trials and reports the precision scores as 1.0 (outperforming STM in all cases). The trials retrieved by BAS will always be relevant because it performs exact lexical matching between a query and the brief title of a clinical trial. We prefer dealing with UMLS concepts because the query has many variations, making direct matching quite difficult. In the case of STM, one-fifth of the trials (five in total) have poor retrieval performance and are unable to retrieve at least 20 trials for a given query, whereas, in the case of METASTM, for every query, we have at least twenty trials except for one query only. We observe that METAADV achieves very high precision@10 values of around 0.9 and significantly higher than METARRF for only Type-I queries. In contrast, for only Type-II queries, METARRF performs much better than METAADV. This shows that the **extended retrieval stage can address the sparsity of clinical trial search**.

D. Performance Evaluation on TREC Precision Medicine Benchmark Dataset

We show the performance comparison between MPSS and the state-of-the-art (SOTA) models of the TREC 2018

TABLE VII
PERFORMANCE EVALUATION OF MPSS ON TREC 2018 PRECISION MEDICINE BENCHMARK DATASET

Model	**Prec@10**	**R-Prec**	**infNDCG**
Cat_Garfield [12]	**0.626**	**0.4294**	**0.5504**
ims_unipd [20]	0.566	0.413	0.0.540
UTDHLTRI [41]	0.538	0.3675	0.4794
MPSS (ours)	0.432	0.303	0.281

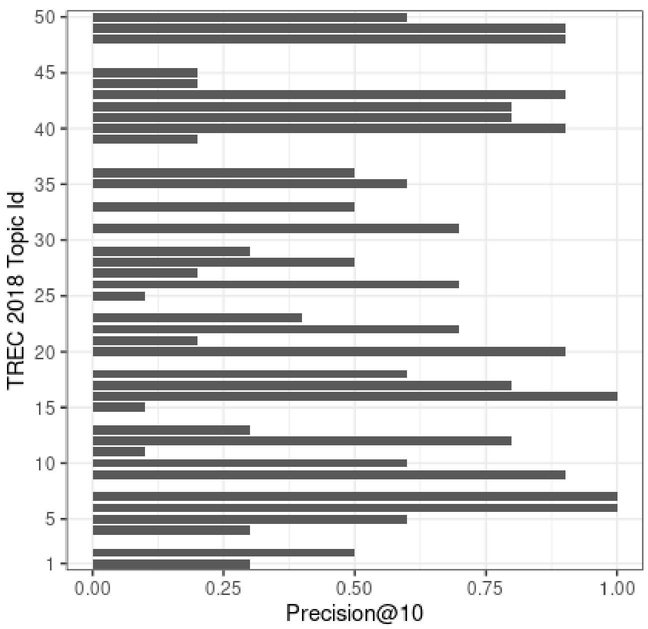

Fig. 5. Query-wise Precision@10 scores achieved by MPSS on the TREC 2018 Precision Medicine task. Each query is identified by a unique topic id.

Precision Medicine task in Table VII. Figure 5 shows the query-wise precision@10 scores for our MPSS model. MPSS achieves a modest Precision@10 score of 0.432 as compared to the TREC-PM 2018 state-of-the-art model (Team Cat_Garfield) score of 0.626 (as reported in Roberts et al. [31]). MPSS achieves precision at ranks one, two, and five of 0.5, 0.42, and 0.46, respectively, and a decent recall performance of 0.588. **MPSS is more interpretable than the black box SOTA models** because one can identify the connection between the query and the final retrieved set of trials. For example, However, the performance gap between MPSS and the SOTA models is quite expected due to the following reasons: (i). The SOTA models utilize the ground-truth data from the 2017 TREC Precision Medicine task to perform supervised learning. In contrast, MPSS is purely unsupervised and does not utilize any training data. (ii). Unlike SOTA models, MPSS do not utilize disease-specific knowledge sources such as COSMIC [40] because of its disease-independence nature.

We observe that the **addition of the drug interactions and gene-drug linked publications data** (INT), followed by the addition of the new metapath based on Pubmed references, improves over MPSS without Pubmed-enhanced retrieval model by 9.75% (0.359 to 0.394) in terms of Precision@10 scores. Next, with the **addition of term frequency** of the gene, gene synonym, and mutation information to the relevance ranking function, the precision at rank ten (Prec@10) further improved from 0.394 to 0.432 (+9.64%). Thus, there is a scope to improve the relevance scoring function. We believe a cancer-specific knowledge base like COSMIC [40] may further improve gene relevance computation.

VI. Conclusion

In this paper, we propose a *metapath-based similarity search* approach, MPSS, for clinical trial search across multiple disease classes. The primary challenges in trial retrieval are the sparsity, term mismatch between query and documents, and explainability of the retrieved trials. We construct a novel heterogeneous information network of both clinical trials and linked Pubmed articles to alleviate the sparsity issue. Further, we explore the path-based retrieval approach that becomes explainable to the end users. Finally, we provide a combined ranked list based on relevance, adversity, and popularity. We contribute an annotated (query-relevant trial) retrieval set for 25 queries (95 trials are annotated per trial on average) across five disease classes. We also evaluate MPSS in a zero-shot setting (without any task-specific training) on the benchmark dataset of the TREC 2018 Precision Medicine Track. 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.

Limitations. Our developed dataset is not exhaustive, i.e., we do not capture all acronyms and 'micro-text' variations of a query. We fail to perform when queries contain medical history information such as *already having celiac disease, antiretroviral therapy first time*, etc. We could consider inclusion and exclusion criteria to address these issues and further extend our dataset.

Future Work. We plan to extend our dataset by incorporating more disease classes and richer queries. The queries could be formulated based on the '*exemplar query*' formulation [42]. We will improve the trial coverage of the heterogeneous information network by including trials that do not have a linked publication, making MPSS more practically useful. Subsequently, we can improve our heterogeneous information network by incorporating more edge types and populating them through a continuous learning strategy. In the future, we would like to deploy an online system to collect user feedback and update MPSS. We can use the large (constructed) heterogeneous information network to learn optimal node representations ([43], [44]), specifically clinical trial nodes. This may help improve the performance of downstream tasks such as similarity search [44].

Acknowledgment

Soumyadeep Roy is supported by the Institute Ph.D. Fellowship at the Indian Institute of Technology Kharagpur. This research was funded in part by the Federal Ministry of Education and Research (BMBF), Germany, under the project LeibnizKILabor with grant No. 01DD20003, and the IMPRINT-funded project titled "Development of a Remote Healthcare Delivery System: Early Diagnosis, Therapy, Follow-up and Preventive Care for Non-communicable Diseases (Cardio-pulmonary)". Furthermore, this work is supported in part by the Science and Engineering Research

Board, Department of Science and Technology, Government of India, under Project SRG/2022/001548. Koustav Rudra is a recipient of the DST-INSPIRE Faculty Fellowship [DST/INSPIRE/04/2021/003055] in the year 2021 under Engineering Sciences.

References

[1] S. Y. Jiang and C. Weng, "Cross-system evaluation of clinical trial search engines," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2014, p. 223—229, 2014.

[2] C. O. Patel, V. Garg, and S. A. Khan, "What do patients search for when seeking clinical trial information online?" in *AMIA Annual Symposium Proceedings*, vol. 2010, 2010, p. 597.

[3] K. Roberts, D. Demner-Fushman, E. M. Voorhees *et al.*, "Overview of the trec 2017 precision medicine track," *TREC*, pp. 500–324, 2017.

[4] K. Roberts *et al.*, "Trec precision medicine 2018 track," in *TREC*, 2018. [Online]. Available: http://www.trec-cds.org/2018.html

[5] TREC, "Clinical trials track," 2022. [Online]. Available: http://www.trec-cds.org/2022.html

[6] S. A. Forbes, N. Bindal, S. Bamford *et al.*, "Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic Acids Research*, vol. 39, no. suppl-1, pp. D945–D950, 2010.

[7] R. Miotto, S. Jiang, and C. Weng, "etacts: A method for dynamically filtering clinical trial search results," *Journal of Biomedical Informatics*, vol. 46, no. 6, pp. 1060 – 1067, 2013.

[8] Antidote, "Antidote trial search engine," 2019. [Online]. Available: https://www.antidote.me/

[9] S. Roy, K. Rudra, N. Agrawal *et al.*, "Towards an aspect-based ranking model for clinical trial search," *Computational Data and Social Networks. CSoNet 2019*, vol. 11917, pp. 209–222, 2019.

[10] A. Thorve, "Team ajinkyathrove at trec 2017 precision medicine track," 2017. [Online]. Available: https://github.com/ajinkyathorve/TREC-2017-PM-CDS-Track

[11] T. R. Goodwin, M. A. Skinner, and S. M. Harabagiu, "Utd hltri at trec 2017: Precision medicine track," in *National Institute of Standards and Technology (NIST)*, 2017. [Online]. Available: https://trec.nist.gov/pubs/trec26/papers/UTDHLTRI-PM.pdf

[12] X. Zhou, X. Chen, J. Song *et al.*, "Team cat-garfield at trec 2018 precision medicine track," in *National Institute of Standards and Technology (NIST)*, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec27/papers/Cat_Garfield-PM.pdf

[13] A. Longa, S. Azzolin, G. Santin *et al.*, "Explaining the explainers in graph neural networks: a comparative study," *arXiv preprint arXiv:2210.15304*, 2022.

[14] L. Soldaini and N. Goharian, "Quickumls: a fast, unsupervised approach for medical concept extraction," in *MedIR workshop, SIGIR*, 2016.

[15] G. Zuccon and B. Koopman, "Choices in knowledge-base retrieval for consumer health search," in *European Conference on Information Retrieval*, 2018, pp. 72–85.

[16] S. Balaneshin-kordan, A. Kotov, and R. Xisto, "Wsu-ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources," 2015. [Online]. Available: https://trec.nist.gov/pubs/trec24/papers/wsu_ir-CL.pdf

[17] V. V. Tong Yin, Danny Wu, "Retrieving documents based on gene name variations: Medier at trec 2017 precision medicine track," in *National Institute of Standards and Technology (NIST)*, 2017. [Online]. Available: https://trec.nist.gov/pubs/trec26/papers/MedIER-PM.pdf

[18] J. L. et al, "Retrieving scientific abstracts iteratively: Medier at trec 2018 precision medicine track," in *National Institute of Standards and Technology (NIST)*, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec27/papers/MedIER-PM.pdf

[19] S. Balaneshinkordan and A. Kotov, "Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine," *Journal of Biomedical Informatics*, vol. 98, p. 103238, 2019.

[20] G. Maria, D. Nunzio, S. Marchesin *et al.*, "Exploring how to Combine Query Reformulations for Precision Medicine," in *TREC*, no. i, 2019, pp. 1–14. [Online]. Available: https://trec.nist.gov/pubs/trec28/papers/ims_unipd.PM.pdf

[21] N. L. Atkinson, S. L. Saperstein, H. A. Massett *et al.*, "Using the internet to search for cancer clinical trials: A comparative audit of clinical trial search tools," *Contemporary Clinical Trials*, vol. 29, no. 4, pp. 555–564, 2008.

[22] Y. Sun and J. Han, "Mining heterogeneous information networks: a structural analysis approach," in *SIGKDD Explorations*, vol. 14, no. 2, 2012, pp. 20–28.

[23] Y. Sun, J. Han, X. Yan *et al.*, "Pathsim: meta path-based top-k similarity search in heterogeneous information networks," *Proc. VLDB Endow.*, vol. 4, no. 11, p. 992–1003, Aug. 2011.

[24] C. Shi, X. Kong, P. S. Yu *et al.*, "Relevance search in heterogeneous networks," in *Proceedings of the 15th International Conference on Extending Database Technology*, ser. EDBT '12, 2012, p. 180–191.

[25] M. Thilakaratne, K. Falkner, and T. Atapattu, "A systematic review on literature-based discovery: general overview, methodology, & statistical analysis," *ACM Computing Surveys*, vol. 52, no. 6, 2019.

[26] P. Martin, D. Surian, R. Bashir *et al.*, "Trial2rev: Combining machine learning and crowd-sourcing to create a shared space for updating systematic reviews," *JAMIA Open*, vol. 2, no. 1, pp. 15–22, 01 2019.

[27] "Improving public access to aggregate content of clinicaltrials.gov by the clinical trials transformation initiative," 2016. [Online]. Available: https://aact.ctti-clinicaltrials.org/

[28] M. L. Anderson, K. Chiswell, E. D. Peterson *et al.*, "Compliance with results reporting at ClinicalTrials.gov," *N. Engl. J. Med.*, vol. 372, no. 11, pp. 1031–1039, Mar. 2015.

[29] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[30] M. D. for Regulatory Activities, "Meddra patient-friendly terms list," 2019. [Online]. Available: https://www.meddra.org/patient-friendly-term-list

[31] K. Roberts, D. Demner-Fushman, E. M. Voorhes *et al.*, "Overview of the TREC 2018 precision medicine track," *Text REtrieval Conference (TREC)*, pp. 1–6, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec27/papers/Overview-PM.pdf

[32] P. Vora, "Searching and filtering," in *Web Application Design Patterns*, 2009, pp. 143–180.

[33] Wikipedia, "Overlap coefficient," 2019. [Online]. Available: https://en.wikipedia.org/wiki/Overlap_coefficient

[34] L. Page, S. Brin *et al.*, "The pagerank citation ranking: bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999.

[35] E. Sayers, "A general introduction to the e-utilities." 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25497/

[36] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, p. 758–759.

[37] S. L. Freshour, S. Kiwala, K. C. Cotto *et al.*, "Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1144–D1151, 11 2020.

[38] D. Maglott, J. Ostell, K. D. Pruitt *et al.*, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D54–58, Jan 2005.

[39] R. Agarwala, T. Barrett, J. Beck *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res*, vol. 46, no. D1, pp. D8–D13, 01 2018.

[40] W. S. Institute, "Cosmic, catalogue of somatic mutations of cancer," 2019. [Online]. Available: https://cancer.sanger.ac.uk/cosmic

[41] S. J. Taylor, S. M. Harabagiu, and T. R. Goodwin, "Utd hltri at trec 2018: Precision medicine track." in *TREC*, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec27/papers/UTDHLTRI-PM.pdf

[42] D. Mottin, M. Lissandrini, Y. Velegrakis *et al.*, "Exemplar queries : a new way of searching," *The VLDB Journal*, vol. 25, no. 6, pp. 741–765, 2016.

[43] T.-y. Fu, W.-C. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1797–1806.

[44] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 135–144.