Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-024-00159-4

RESEARCH ARTICLE

# Clinical Information Retrieval: A Literature Review

Sonish Sivarajkumar[1] · Haneef Ahamed Mohammad[2] · David Oniani[3] · Kirk Roberts[4] · William Hersh[5] · Hongfang Liu[4] · Daqing He[2] · Shyam Visweswaran[1,6,7] · Yanshan Wang[1,3,6,7]

Received: 28 March 2023 / Revised: 7 December 2023 / Accepted: 8 January 2024

## Abstract

Clinical information retrieval (IR) plays a vital role in modern healthcare by facilitating efficient access and analysis of medical literature for clinicians and researchers. This scoping review aims to offer a comprehensive overview of the current state of clinical IR research and identify gaps and potential opportunities for future studies in this field. The main objective was to assess and analyze the existing literature on clinical IR, focusing on the methods, techniques, and tools employed for effective retrieval and analysis of medical information. Adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, we conducted an extensive search across databases such as Ovid Embase, Ovid Medline, Scopus, ACM Digital Library, IEEE Xplore, and Web of Science, covering publications from January 1, 2010, to January 4, 2023. The rigorous screening process led to the inclusion of 184 papers in our review. Our findings provide a detailed analysis of the clinical IR research landscape, covering aspects like publication trends, data sources, methodologies, evaluation metrics, and applications. The review identifies key research gaps in clinical IR methods such as indexing, ranking, and query expansion, offering insights and opportunities for future studies in clinical IR, thus serving as a guiding framework for upcoming research efforts in this rapidly evolving field. The study also underscores an imperative for innovative research on advanced clinical IR systems capable of fast semantic vector search and adoption of neural IR techniques for effective retrieval of information from unstructured electronic health records (EHRs).

**Keywords** Information retrieval · Electronic health records · Natural language processing · Tools · Applications

Extended author information available on the last page of the article

Published online: 23 January 2024

## 1 Introduction

The amount of information available in electronic health records (EHRs) has grown rapidly in recent years. The clinical information in EHRs encompasses many different aspects of a patient's care, including conditions, various examination results, medical treatments, and therapeutic effects, which can be used for clinical decision support and a variety of secondary purposes [1–3]. The rapid expansion of EHRs has made it essential to have accurate and efficient access to relevant medical information contained within these documents. Despite the fact that several EHR components can be structured, 80% of EHRs are unstructured and inserted as free-text clinical notes [4]. Therefore, the ability to effectively search the clinical information embedded in the free-text clinical notes is essential for the effective utilization of patient-related information to improve medical practice and patient care, as well as to facilitate clinical research [5].

Information retrieval (IR) is a technique used by search engines to store, retrieve, and rank documents from a large collection of text documents based on users' queries [6]. It is a field of study that encompasses the design, development, and evaluation of systems and methods for the identification and retrieval of relevant information from a large corpus of documents. IR allows clinicians, medical staff, and other users to rapidly retrieve relevant information from enormous free-text EHRs, making it a very effective technique. Clinical IR is a specific type of IR that refers to the process of locating and accessing relevant medical information in various clinical textual data sources to facilitate clinical practice and research. Clinical IR research focuses on innovating the conventional IR infrastructures and methodologies to meet the information needs in clinical applications. In the clinical or biomedical domain, users may include clinicians, researchers, nurses, and other healthcare workers with varying information needs. For instance, healthcare professionals may search on disease-related keywords for retrieving patient cohorts from EHRs, or researchers may search existing literature for evidence of a rare disease.

Since unstructured medical texts predominate in EHRs, it is challenging to automatically identify critical information from unstructured EHRs for clinical practice and research. In addition, these documents may have a complex structure and contain misspelt terms and abbreviations, making retrieval difficult for typical database querying tools. Consequently, standard database querying approaches, such as structured query language (SQL), may produce inaccurate results with low recall. We require IR systems, such as ad hoc search engines, capable of handling the semantics and pragmatics of the complex text in EHRs [7]. Therefore, it is crucial to develop clinical IR systems that manage medical data to meet user requirements.

## 2 Background

IR is a scientific discipline that deals with the representation, storage, and retrieval of relevant information from a large collection of documents based on the user's information needs [8]. IR systems have existed in the field of computer science for more than 50 years. Early IR systems were mainly used by librarians to retrieve documents in their document store. In the mid-1990s, with the rise of the World Wide Web, a plethora of new forms of data—structured, semi-structured, and unstructured—proliferated over the internet. This created the need for more advanced IR systems. Libraries, legal and medical databases, desktop search engines, social media, mobile search engines, question-answering services, and chatbots are just a few examples of how IR systems have evolved over time.

Even though the applications of IR systems differ in each domain, the fundamental process of IR remains the same. Figure 1 illustrates a basic IR process diagram. In an IR system, the first step is to index the documents. Indexing is the process of structurally organizing all the data structures in a document collection, which stores the embedded information in all the documents into a single structure called an index. This process facilitates efficient storage and retrieval of data in an IR system. Inverted indexes are one of the most widely used indexing methods due to the fact that they enable quick and efficient searches of enormous document collections. It is called an "inverted" index because it saves a mapping between the words or phrases that appear in a document and the papers in which they appear, as opposed to storing a mapping between the documents and the terms they include. Each word or term in a standard inverted index is associated with a list of documents in which it appears. For instance, if the phrase "patient" appears in documents 1, 4, and 7, the inverted index may have the item "patient: [1, 4, 6]". When a user submits a query for the

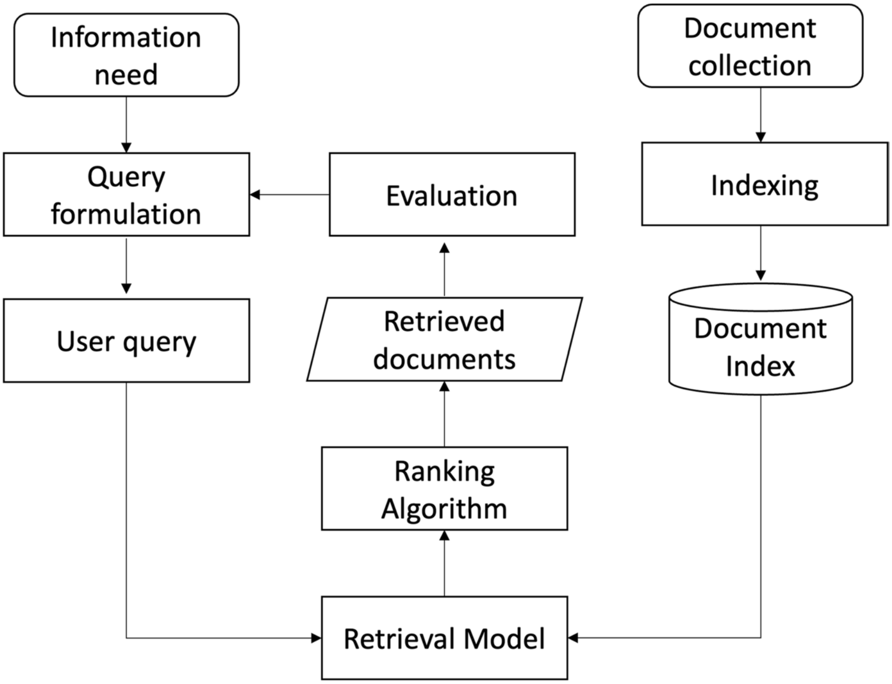

**Fig. 1** A basic IR process diagram

term “patient,” an inverted index can quickly look up the list of documents containing that term and return them to the user.

Querying refers to the process of searching relevant documents or other information in response to a particular request or query. Typically, one or more keywords or phrases are entered into a search interface or system as a search query. Then, after searching through its index or collection of documents, the IR system returns the documents that are most pertinent to the query. In addition to keyword searches, many IR systems provide advanced query types, such as Boolean queries, which enable users to specify more complicated search parameters and use logical operators**.** Query reformulation is often done to refine the query based on user feedback on the retrieved documents. The process of modifying or adding new search terms to a query in order to expand the search space is known as query expansion.

Ranking is the process of providing a relevance score to each page in a collection based on how closely it corresponds to a certain query or request. The ranking algorithm matches the user query with the document index and retrieves the relevant documents [9]. Ranking is used to establish the order in which search results are given to the user in an IR system, with the most relevant results appearing first. There are numerous ways to rank documents in an IR system, and the ranking algorithm employed can have a substantial effect on the quality and efficacy of search results. The following are examples of common ranking algorithms used in IR systems:

- Boolean models: These use Boolean logic to determine the relevance of documents to a given query. The ranking is binary, meaning that documents are either relevant or not relevant to the query, based on the presence or absence of keywords.
- Vector space models: These represent documents and queries as vectors in a high-dimensional space. The ranking is based on a similarity score between the vectors (e.g., the cosine of the angle between the vectors). Document vectors with higher similarity to the query vector indicate higher relevance of the document to the query.

  Term frequency-inverse document frequency (TF-IDF): Given a term in a query, TF-IDF is a ranking method used to measure the importance of the term in a document relative to an entire corpus of documents. It calculates the importance of terms by multiplying the frequency of the query term in a document (TF) by the inverse of the number of documents in a corpus that contain that term (IDF).

  Best Match 25 (BM25): BM25 is a probabilistic ranking algorithm that calculates relevance scores for a document based on (similar to TF-IDF) the frequency of query terms within the document. It takes into account the document length and corpus term frequency and also incorporates user-adjustable parameters ($k1$ and $b$) for fine-tuning the relevance scores.

- Statistical language models: These use statistical techniques to model the probability of a query given a document. The ranking is based on a likelihood score, with higher scores indicating higher relevance.

- Learning-to-rank models: These use machine learning techniques to learn a ranking function from labeled data. These models can be trained on a variety of features, such as the relevance of a document, the term frequency, or the click-through rate. Deep learning–based models use deep neural networks to learn complex representations of documents and queries. These models can be trained on a variety of data and can be used for a variety of tasks, such as document retrieval or question answering.

Re-ranking is a technique used in IR systems to enhance the quality and relevance of search results by accounting for extra context or user preferences. It is the process of changing the relevance score of documents depending on new factors or information. Re-ranking can be applied in a variety of ways in an IR system. One such method, referred to as relevance feedback, involves improving the retrieval system based on user evaluation of the ranked list. The feedback could be the conventional relevance check (relevant or non-relevant) or the click through rate (CTR) for Internet webpage retrieval. The ranking algorithm is modified by learning from the retrieval errors as per user feedback. Re-ranking can also be used to incorporate additional data sources, such as external databases, or user feedback, such as ratings.

Clinical text encompasses a set of unstructured EHR documents that are distinct from general documents, medical literature, and online health resources. These documents have unique features, such as the use of medical terms, abbreviations, and context-specific phrases, all of which pose challenges for IR systems. These challenges require specialized indexing and ranking methods that consider the peculiarities of clinical text, which general IR systems would not account for.

Clinical IR uses IR methodologies to improve access to clinical information, which includes patient-specific free-text EHR documents from hospitals and providers. Thus, Clinical IR can also be defined as the process of accessing and using this clinical information in order to support clinical decision-making and improve patient care. Patient-specific information is of interest to a wide variety of users, including researchers, clinicians, and clinical trial experts. Despite increased interest in IR among clinical informatics professionals and improvements in IR techniques over the past few decades, the majority of clinical IR systems rely on conventional IR technologies.

## 3 Related Work

The primary rationale for conducting this review is the absence of concise information on the latest literature of clinical IR. IR is a crucial field that has seen significant advancements in recent years, particularly in the area of biomedical literature. A recent review by Tamine and Goeuriot [10] provides an overview of IR applications and challenges in medical texts, mainly focusing on biomedical literature. The review highlights the importance of IR in the biomedical domain and the various challenges faced while working with medical texts. Similarly, a book by Hersh [6] delves into the principles and techniques of IR as applied to the field of health and medicine. The book provides an in-depth exploration of the various techniques used in IR and how they can be applied to the field of medicine.

While the abovementioned works provide a broad picture of IR applications and challenges in the medical domain, they mainly focus on biomedical literature. In contrast, this paper aims to fill a gap in the literature by conducting a comprehensive examination of methodologies, implementations, tools, and applications of IR specifically in the clinical domain, with a focus on free-text electronic health record (EHR) data. EHRs are an essential source of patient information, and their proper management is crucial for providing efficient and effective healthcare. However, the sheer volume of data present in EHRs makes it challenging to extract relevant information. IR techniques can be used to improve the retrieval of relevant information from EHRs, making them more useful for both clinicians and researchers.

Other studies in the field of IR in healthcare and medicine include Himani and Dattani [11] who provide a survey on medical IR, Gudivada and Tabrizi [12] who review machine learning-based medical IR systems, Daei et al. [13] who examine physicians' clinical information seeking behavior, Montani and Striani [14] who survey artificial intelligence in clinical decision support, and Khattak et al. [15] who review word embeddings for clinical text. While these papers provide valuable insights into specific aspects of IR in the field of health and medicine, none of them provides a comprehensive overview of the methods used for clinical IR, specifically focusing on the use of IR techniques to improve the retrieval of relevant information from EHRs. The only paper which provides a detailed explanation of some of the retrieval methods used in unstructured EHR-based clinical IR practice is by Lopus [13]. However, this study still lacks information about patient cohort retrieval models, details about evaluation, shared tasks, and applications related to clinical IR.

The field of clinical IR has been relatively under-explored. This paper aims to fill these gaps by providing a comprehensive examination of the various techniques, tools, and methodologies used for IR on EHRs; the evaluation strategies, various shared tasks organized in clinical IR community; and various applications of IR in the clinical domain. Additionally, it aims to provide a summary of the current state-of-the-art and lay the groundwork for the next generation of systems in the field of clinical IR. Although the technologies and applications may overlap with biomedical literature, this paper provides a specific focus on the IR of clinical documents, making it a valuable resource for researchers and clinical practitioners in the field.

Finally, we will also provide insights on the current limitations and challenges faced in clinical IR and identify opportunities for improvement in the field. Our ultimate goal is to contribute to the advancement of clinical IR systems by highlighting the areas that need to be addressed and providing recommendations for future research and development.

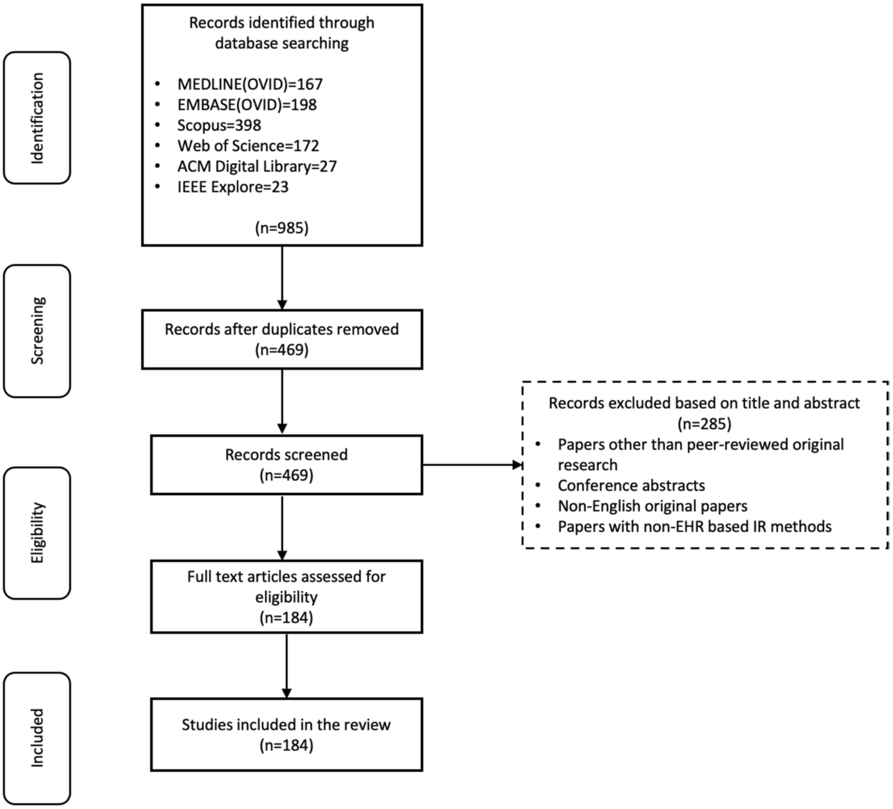
Identification
Screening
Eligibility
Included
Records identified through database searching
MEDLINE(OVID)=167
EMBASE(OVID)=198
Scopus=398
Web of Science=172
ACM Digital Library=27
IEEE Explore=23
(n=985)
Records after duplicates removed (n=469)
Records screened (n=469)
Records excluded based on title and abstract (n=285)
Papers other than peer-reviewed original research
Conference abstracts
Non-English original papers
Papers with non-EHR based IR methods
Full text articles assessed for eligibility (n=184)
Studies included in the review (n=184)

**Fig. 2** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of the article screening and identification process

# 4 Methods

## 4.1 Data Sources and Search Strategies

The review was conducted on the basis of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [16]. Figure 2 depicts the PRISMA flow diagram of the article screening and identification process. We conducted a comprehensive search of several databases from January 1, 2010, to January 4, 2023. We selected this time range because the HITECH act for EHR was passed in 2009, which provided incentives to healthcare providers to adopt and demonstrate “meaningful use” of EHRs. This led to a widespread implementation of EHR systems during 2010 and 2011. Therefore, the papers published after 2010 provide a more accurate representation of the current state of research in clinical IR as it relates to EHRs, as many of the papers before 2010 may not have had access to the same amount of EHR data and may not have been able to address the same challenges and issues.

The search strategy was designed and conducted by an experienced librarian with input from the study's principal investigator. We only included journal articles and conference proceedings that were published in English. The databases included Ovid Embase, Ovid Medline  Scopus, ACM Digital Library, IEEE Xplore, and Web of Science. The detailed search strategy listing all search terms used and how they are combined is available in the Appendix.

### 4.2 Article Selection

A total of 985 articles were retrieved from five libraries, of which 469 articles remained after deduplication. To filter out articles that did not actually focus on the EHR-based clinical IR process, the articles were manually screened based on the title, abstract, and method sections. Papers that did not mention the details of the clinical IR method used and that did not include EHR-based IR methods were eliminated. This helped to ensure the quality and reliability of the papers included in the review. Articles without full text or methodology description were excluded as well. Following this screening process, 184 articles remained to be comprehensively reviewed by the study team. The papers were categorized into the following broad types during the full-text review: (1) Methodology, (2) Application, and (3) Review. A paper was categorized as "methodology" if it focused on the development and evaluation of new methods or techniques for clinical IR. A paper was categorized as "application" if it described the use of existing clinical IR methods in real-world scenarios. Those categorized as "review" provide an overview of a particular area of clinical IR research. Those papers that could potentially fall into more than one category were carefully evaluated and categorized based on the primary focus of the paper. This categorization allows for a separate and comprehensive review of the methodology and application sections.

## 5 Results

This section presents an in-depth analysis of the publication sources and venues of all 184 papers that were selected for the review. We begin by presenting a summary of the year-wise distribution of papers and publication venues in the field of clinical IR research, over the 13-year time frame of the review. Next, we analyze the content and type of the articles published in clinical IR research, providing insights into the areas of research being conducted in the field. We then present the existing tools available for clinical IR as presented in the reviewed literature, summarizing the methods used in clinical IR research and providing a detailed insight into the available algorithms and frameworks for clinical IR. Additionally, we consolidate the evaluation methods and metrics used in these papers, providing a comprehensive overview of the various metrics used to evaluate the performance of clinical IR systems. Then we provide a brief summary of clinical IR shared tasks, giving readers

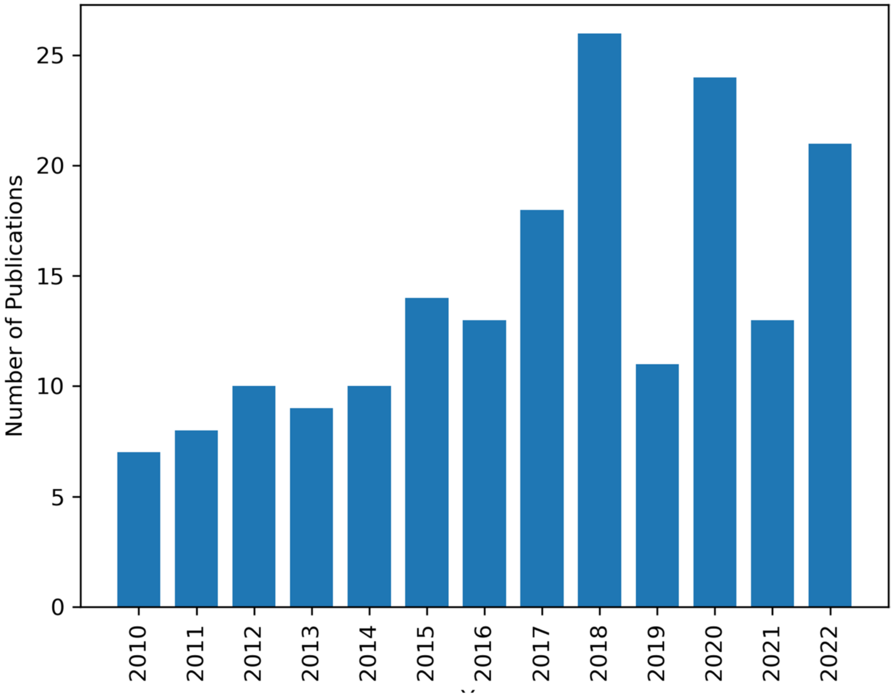

**Fig. 3** Distribution of clinical IR publications per year

an overview of the publicly available datasets for clinical IR research. Finally, the last part of the section describes the practical IR applications in the clinical domain, including patient cohort detection, chart review, and others.

As shown in Fig. 3, there is a clear upward trend in the papers published on clinical IR from 2010 ($n=7$) through 2018 ($n=26$). This trend can be directly correlated to the increasing number of EHR-related publications from 2009 to 2015 [17]. However, subsequent years witnessed a downtrend and stagnation in the number of publications, with a shift in focus towards clinical IR applications. The downtrend may be attributed to several factors such as a shift in the TREC clinical shared tasks from EHR-based retrieval systems to other applications and a lack of new annotated clinical IR datasets being released during this time frame.

### 5.1 Publication Venue

After careful analysis of the 184 papers, we segmented the papers into the type of publication—journal article or part of conference proceedings. We observed that 114 articles were published in journals, while 70 papers were published in conference proceedings. From 2010 to 2014, clinical IR-related papers were predominantly published in conference proceedings, as illustrated in Fig. 4. This is partly due to the clinical IR shared tasks in conferences like the Text REtrieval Conference (TREC)

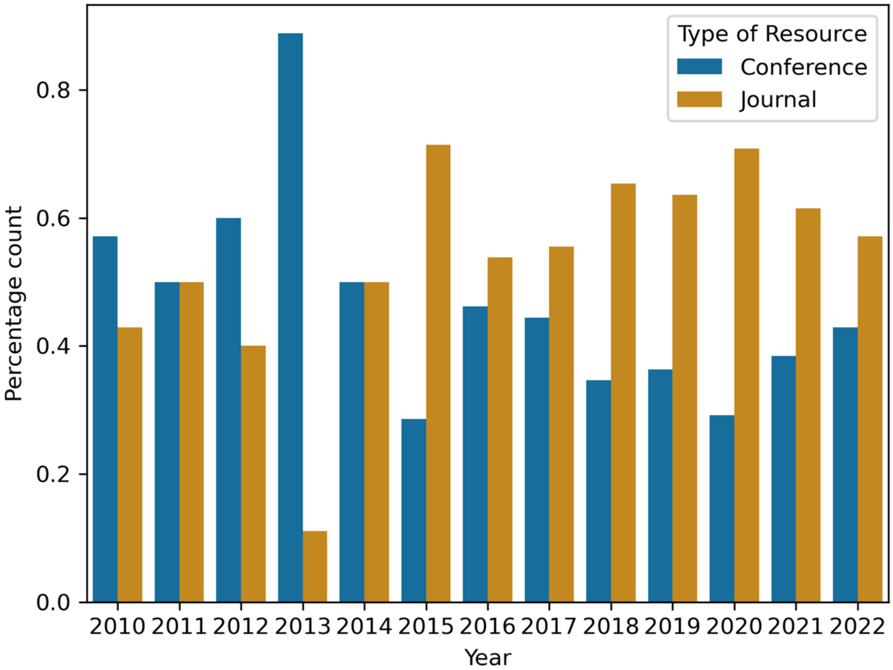

**Fig. 4** Categorization of publication types

in early 2010s. These shared tasks involve making annotated data available to participants who compete to develop algorithms for specific IR tasks. With the availability of more annotated data and a standardized evaluation framework provided by the shared tasks, researchers are able to develop and compare different methods more effectively, which may have led to the rise in published clinical IR articles in conference proceedings.

Since 2015, there has been a rapid increase in the number of clinical IR articles published in journals. This increase in publications can potentially be attributed to the greater adoption of EHRs within healthcare systems, which has led clinicians and healthcare professionals to identify the necessity for more sophisticated search engines. Therefore, the growing demand for advanced clinical IR systems and their potential applications in healthcare may have contributed to the observed increase in the number of articles published in clinical and informatics journals.

We observed that the 184 papers were published in 107 unique venues, of which 51 are conferences and 56 are journals. Overall, the publication venues with three or more papers are (1) "Journal of Biomedical Informatics" ($n=8$), (2) "BMC Medical Informatics & Decision Making" ($n=7$), (3) "JMIR Medical Informatics" ($n=7$), (4) "IEEE International Conference on Healthcare Informatics" ($n=6$), (5) "Journal of the American Medical Informatics Association" ($n=6$), (6) "Medical Informatics in Europe Conference (MIE) ($n=5$)", (7) "JAMIA Open" ($n=4$), (8) "AMIA Annual Symposium Proceedings" ($n=4$), (9) "MEDINFO" ($n=4$), and (10) "IEEE International Conference on Bioinformatics" ($n=3$). "Applied Clinical Informatics,"

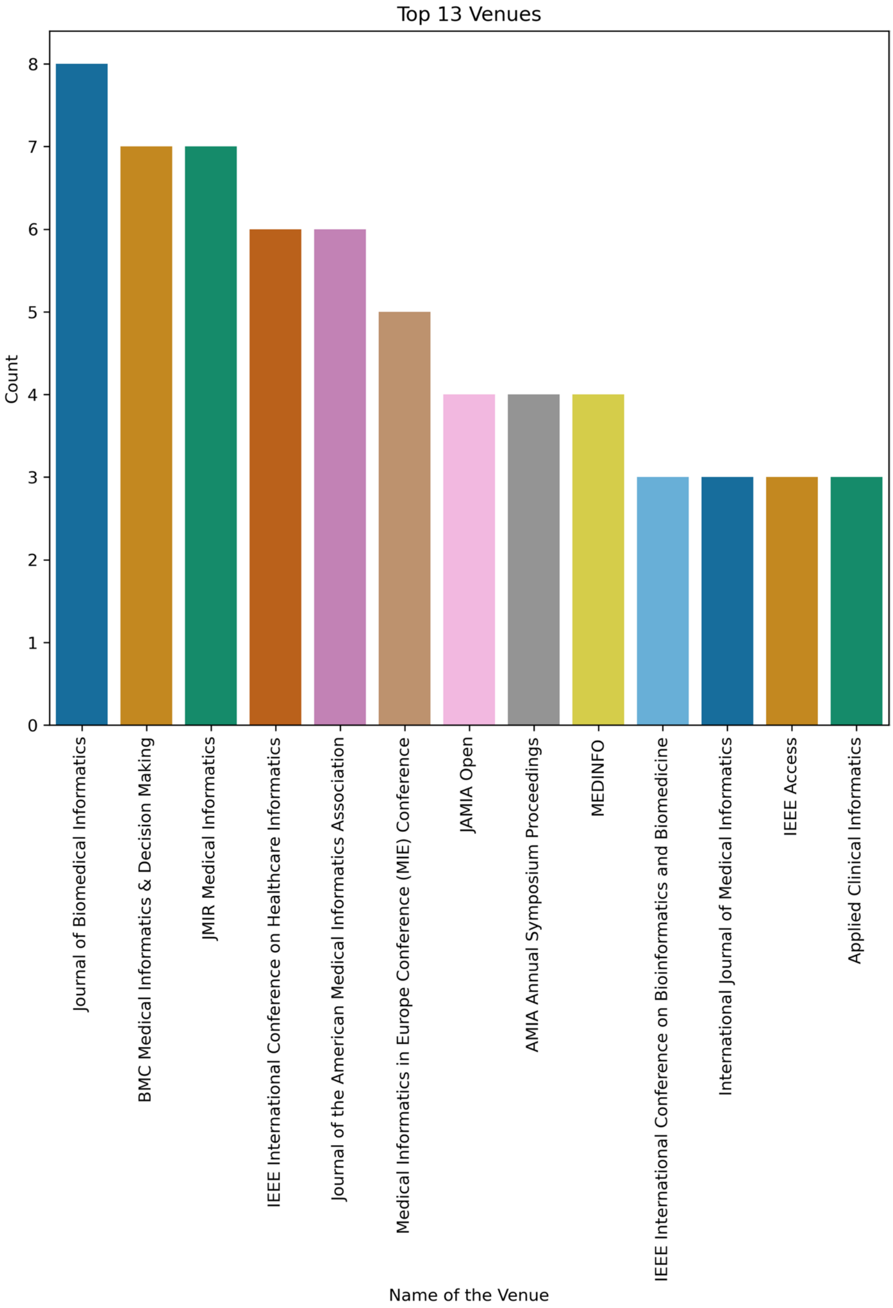

**Fig. 5** Top 13 publication venues for clinical IR articles

International Journal of Medical Informatics," and "Journal of Biomedical Informatics and IEEE Access" also have 3 publications on clinical IR.

Figure 5 shows that out of the top 13 publication venues, 8 are journals. This trend is likely influenced by the practical applications of clinical IR in areas such as treatment and diagnostics, as compared to the computer science field, which may place a greater emphasis on theoretical study and tend to publish more in conferences. The emphasis on practical applications in clinical IR may have encouraged researchers to prioritize the development and evaluation of methodologies and their applications rather than theoretical study. Additionally, the medical domain has a historical preference for publishing in journals over conferences, which may also contribute to the preponderance of journal publications in the clinical IR research community.

We also observed a tail in the distribution count of publication venues, where a large number of journal and conference proceedings venues have just one clinical IR publication. The dispersed distribution of papers across different venues indicates that studies on clinical IR are highly segmented, which makes it essential for a scoping review to gather the findings and trends together in one central location. Our paper aims to fulfill this need by providing a comprehensive overview of the field, consolidating the dispersed research in one place.

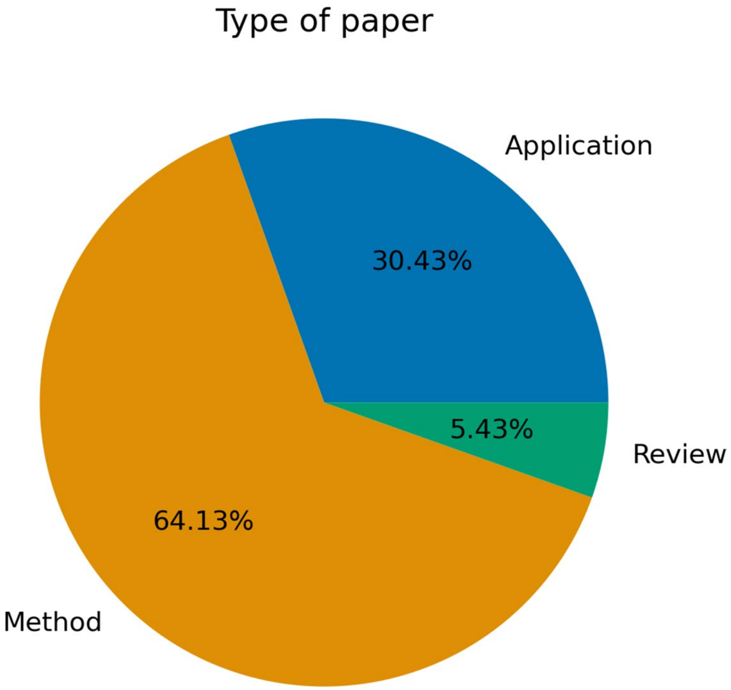

**Fig. 6** Categorization of clinical IR article types

## 5.2 Article Type

We further analyzed the type of the clinical IR publication and segmented the 164 papers into one of the following categories: (1) Application, (2) Method, or (3) Review, as shown in Fig. 6. The majority of the papers ($n=118$, 64.13%) are Method papers, which detail novel approaches to clinical IR system design, including novel algorithms, frameworks, and procedures. The next most common type was Application studies ($n=56$, 30.43%); articles in this category discuss the use and implementation of clinical IR. The remaining publications ($n=10$, 5.43%), Reviews, include surveys and reviews of current technologies for query expansion and semantic search.

## 5.3 Implementations

In this section, we present an overview of the different IR methodologies present in the literature, dividing the discussion into two subsections: clinical IR tools and methodologies. The clinical IR tools subsection focuses on IR tools and systems that have been developed or implemented for the clinical IR. The methodologies subsection discusses querying, indexing, and ranking methodologies that have been proposed and evaluated in the field of clinical IR.

### 5.3.1 Clinical Information Retrieval Tools

Traditionally, SQL-based searching or querying systems were used to build clinical IR systems, but these were not effective in searching the highly unstructured free-text EHR data [18]. Consequently, advanced clinical IR tools are now being developed using more modern search engine techniques.

**IR Tools** Lucene is a Java-based IR tool that provides a set of APIs for building full-text search on documents [19]. It includes tools for indexing, searching, and ranking documents, as well as support for various query types, such as Boolean query searches. Lucene is widely used as the foundational tool for building custom search applications and is also used as the core search engine in many commercial products.

Solr is an open-source enterprise search platform built on top of Lucene [20]. It provides a standalone server that can be used to index and search large collections of documents, as well as a rich set of features for managing and scaling search applications, including support for distributed search and faceted navigation. Solr is commonly used to build search applications for websites, intranets, and other large-scale systems.

Elasticsearch is an open-source full-text search engine, which provides a distributed indexing system on the top of Lucene. Many clinical IR systems have been developed leveraging Elasticsearch, some of which are as follows. Researchers from Mayo clinic developed a distributed infrastructure with two Hadoop clusters to process the HL7 messages into an Elasticsearch index. This Elasticsearch index could provide high-speed text searching (0.2 s per query) on an index containing a dataset of 25 million HL7-derived JSON documents [21]. SigSaude is another platform that integrated patient information from student-run clinics of the Federal University of Rio Grande do Norte. The platform was built on top of an Elasticsearch index, and the data views were created using Kibana [22].

Lemur is a research project focused on developing IR and natural language processing techniques for use in large-scale search applications [23]. It includes tools for indexing, searching, and evaluating the performance of IR systems, as well as support for a variety of advanced features such as query expansion and language modeling. Lemur is primarily used as a research platform and is not as widely used as Lucene, Solr, or Elasticsearch in commercial applications.

### 5.3.2 IR Systems

Essie is a concept-based search engine developed by NIH, with concept-based query expansion and probabilistic relevancy ranking [24, 25]. Lucene-based search engines have long been used for clinical IR and patient cohort detection [26, 27]. Yadav et al. proposed a modified Apache Lucene ranking algorithm based system which has an feedback system based on the number clicks and likes-dislikes for the search results [28].

EMERSE (Electronic Medical Record Search Engine), launched in 2005, is one of the earliest non-commercial EMR search engines. EMERSE supports free-text queries and has been used by many hospital systems. Researchers from the University of Michigan documented how EMERSE has been used in their hospital system, enabling the retrieval of information for clinicians, administrators, and clinical or translational researchers [29]. EMERSE uses clinical narratives and may not be the best search engine if queries involve structured electronic health record data such as demographic information or lab tests. EMERSE has been successfully used in screening clinical notes to identify patient cohorts, such as to identify glaucoma patients with poor medication compliance [30].

CogStack is an IR system which was built to integrate document retrieval and information extraction for a large UK NHS Trust [31]. The CogStack platform includes a stack of services that enable full-text clinical data searches, real-time risk prediction, and alerts for advanced patient monitoring [32]. Wang et al. used the CogStack platform to implement real-time psychosis risk detection and an alerting service in a real-world EHR system. This is the first study to create and use early-stage psychosis detection and alerting system in clinical practice [32].

MetaMap is a common natural language processing tool utilized in constructing IR systems [33]. MetaMap is a tool developed to retrieve relevant MEDLINE citations based on queries of the user. It allows one to search for the titles and abstracts of MEDLINE citations by mapping concepts in the text to the UMLS Metathesaurus. Researchers create simple hashes that map the Concept Unique Identifiers (CUI) from MetaMap to patient records [26, 34]. The US National Library of Medicine (NLM) manages the MEDLINE/PubMed database, which contains bibliographic references to biomedical articles. Users can download these MEDLINE/PubMed records for research purposes.

CDAPubMed is an open-source web browser extension developed in 2012 to incorporate EHR elements into biological literature retrieval methods [35]. The Retrieval And Visualization in ELectronic health records (RAVEL) project aims at retrieving relevant elements within the patient's EHR and visualizing them. They proposed implementing an extensive industrial research and development effort on the EHR while taking the following factors into account: IR, data visualization, and semantic indexing [21, 36]. Medreadfast is a hybrid browser designed specifically for combining an EHR keyword search over an automatically inferred hierarchical document index [37].

Although most of these tools were developed between 2005 and 2012, it can be observed that they are still used for clinical IR research. This suggests that more advanced clinical IR methods—utilizing advanced machine learning techniques—could be integrated into these already-established workflows to improve their efficiency and effectiveness.

### 5.3.3 Clinical Information Retrieval Methodologies

This section summarizes the methods used in the reviewed articles for the following three IR components: querying, indexing, and ranking.

**Query Methods** *Keyword search* is the simplest technique to search over free-text EHRs. It involves identifying and searching for the lexicalized (surface) forms of specific words or phrases within a collection of EHR documents or a clinical database. To perform a keyword search, the user enters a query containing one or more keywords into the search field of a search engine or database. The keyword search engine then looks for documents or records that contain those keywords and returns a list of results ranked according to the number of occurrences of these keywords. Early clinical IR systems used keyword search, which did not always return the most relevant or accurate results, particularly if the keywords used in the query were too broad [38]. Studies demonstrated that this method may not be well suited for searching for more complex clinical information as it relies on the surface form of query terms rather than the underlying semantics of the search query [37].

The limitations of keyword-based search led to the development of more advanced querying and ranking systems that could interpret the semantics of complex clinical texts in EHRs. One such limitation is the issue of negation, which can lead to retrieving irrelevant documents despite containing the query keywords. The presence of a query keyword does not always imply that the document is relevant. For instance, “no family history of cancer” could be retrieved for a query to search patients with “cancer”. This issue of negation has to be addressed to avoid retrieving EHRs that contain phrases in contexts that are not relevant to the query. Garcelon et al. tried to address this problem by extracting subtexts from each original patient record and classifying them into 4 categories: “patient–not negated”,”patient–negated”,”family history–not negated”, “family history–negated” [39]. By using contextual information, such as negation, temporality, and the subject of clinical mentions, semantic contexts can be incorporated into an Elasticsearch-based indexing/scoring system [40, 41].

Ontology is the formal representation of a set of concepts and their interactions within a domain. It helps to classify, annotate, and query biological data by organizing and standardizing the information within a certain area [42]. Ontologies and other knowledge-based resources are used to extract the semantic nature and associations of medical terms, which are then used at the record level to infer the patient's overall medical history [43–45]. *Semantic search* enhances the representation of both queries and free-text EHRs by expressing concepts and their contexts. In 2011, Gurulingappa et al. developed a computational platform for clinical IR with the aim of exploring clinical ontology-based semantic search techniques [46]. Afzal et al. proposed query generation from Medical Logic Modules (MLMs) [47] where they built different query sets from the concepts used in MLMs. These sets were then expanded with domain ontology derived from SNOMED CT. More details about semantic search will be discussed in later sections of this paper.

*Concept-based information retrieval (CBIR)* is a type of IR system that uses concepts, or high-level abstractions, to represent and index the content of documents. These concepts are typically derived from the words and phrases that appear in the documents and are organized into a hierarchy or ontology to provide a more intuitive and meaningful representation of the information. This method can be more effective than a traditional keyword-based search, as it offers less opportunity for ambiguity and vocabulary mismatch. In these systems, queries and documents are standardized from their original terms to concepts from medical ontologies. Early uses of CBIR for biomedical literature [48] have been ported to use for clinical IR using SNOMED CT concepts [7, 47, 49]. Researchers used MetaMap to identify UMLS concepts and to map the UMLS and SNOMED concept ID in the EHRs with the queries [49]. Formal concept analysis (FCA) is another method to derive the concept hierarchy and match it with the indexed documents [50, 51].

*Query expansion* is another mechanism through which concepts can be integrated into the query. Instead of altering the query to a concept-based representation, the sets of synonyms in an ontology accompanying the concepts found in the query are added as additional query terms. This has been used, for instance, to perform query expansion using the UMLS Metathesaurus [52–55]. Topic modeling is a technique used in natural language processing to identify and extract the main themes in a collection of text documents. It can be used to expand patient queries by identifying related concepts and keywords that are present in the EHR notes but not included in the original query [7]. As with UMLS and SNOMED-based query expansion, MeSH-based query expansion has also been utilized [56].

Clinical IR queries can be mapped to a common data model, like the Observational Medical Outcomes Partnership (OMOP) Common Data Model, to standardize queries. This involves the extraction of entity mention types from patient-level IR queries and mapping them to a subset of OMOP data fields [57]. Wen and colleagues proposed an empirical data model that is implemented to cover major entity mention types in cohort identification tasks [40]. They investigated the Clinical Data Repository tables from the Mayo Clinic and Oregon Health & Science University to map the corresponding fields in both a structured and an unstructured format to the proposed data model. In 2020, Shi et al. investigated the relationship between different querying approaches and the characteristics of the cohort definition structure or query taxonomy. But even after developing a 59-parameter taxonomy, they failed to find any significant associations [58].

Modern IR systems frequently utilize automatic query expansion to increase the search space, as the original query may be too narrow or ambiguous, or the search terms may not accurately capture the relevant information. The reformulated query with the expansion terms achieves better results than the original query. The expanded query can be used to obtain more accurate and relevant information from EHRs, which can aid in making better clinical decisions and improving patient outcomes. In clinical IR, researchers have proposed several methods for query expansion based on features of medical language and clinical needs [46]. *Semantic query expansion (SQE)* techniques use semantically similar terms to expand the queries [50, 51]. Based on the meaning of the words in the query, semantic query expansion seeks to develop useful candidate features suitable for query expansion. Utilizing the clinical associations between terms from ontologies, including knowledge of synonyms and hypernym/hyponyms, and semantic relationships among medical concepts, such as symptoms, exams and tests, diagnoses, and treatments, led to an improvement in the precision and recall values of the IR systems [59]. In a recent paper, Wang et al. [60] used a CANDECOMP PARAFAC-alternating least squares (CP-ALS) decomposition algorithm to identify latent variables or hidden factors within EHRs to enhance the initial query. These latent variables can be used to represent important concepts or patterns in the EHR data, such as disease progression, treatment effectiveness, or patient outcomes. In another study, Kreuzthaler et al. [61] used a log-likelihood–based co-occurrence analysis to identify patterns of co-occurrence between the ICD-10 codes and the related keywords. By comparing the log-likelihood of different pairs of terms, this method could identify terms that are most likely to be related to each other. The identified co-occurring terms were then used to identify possible candidates for expanding the initial query.

*Term weighting* is the process of assigning a weight to each term in a document in order to reflect the importance of that term in the document. This method can be used to improve the effectiveness of IR systems by helping them to identify and prioritize the most relevant terms and documents. Semantic term weighting is a type of term weighting that takes into account the meaning and context of the terms being used, rather than just their frequency within a document. There are a variety of techniques that can be used to calculate semantic term weights, including methods that take into account the co-occurrence of terms within a document, the relationships between terms, and the overall structure and content of the document. Yang et al. proposed an algorithm for SQE by improving expansion term weights [62] and their similarity calculation using Word2Vec, GloVe, and BERT [63–65]. Wang et al. proposed an automatic parts-of-speech–based term weighting scheme which iteratively calculates the term weight by utilizing a cyclic coordinate method. They used a golden section line search algorithm along each coordinate to optimize an objective function defined by mean average precision (MAP) [66]. Yang et al. weighted the terms with semantic similarities and assigned calculated category weights and co-occurrence frequencies between expansion terms and multiple query terms. If semantic term weighting is done on an index, instead of the query, we may have to deal with two challenges: to determine the meaning of a medical term in a given clinical text and to give semantic weights to a large number of terms in the indexed clinical texts [67]. Hence, term weighting is done mostly on search queries.

Query expansion using a combination of multiple techniques has been shown to produce more effective results than relying on a single expansion system, as described in the previous section. Several studies have reported that combining different external resources can significantly improve the effectiveness of query expansion. For instance, some researchers have proposed a method that combines medical concept weighting and expansion collection weighting, which has been shown to improve retrieval effectiveness compared with uniform weighting methods [68, 69]. Specifically, the medical concept weighting approach assigns different weights to medical concepts based on their importance in representing the information needs of the query, while the expansion collection weighting approach assigns different weights to the expansion terms based on their relevance to the collection as a whole. The combination of these two approaches has been found to enhance the performance of the IR system by capturing both the query-specific and collection-specific aspects of relevance.

*Relevance feedback* is the process of incorporating feedback on the retrieved documents. Generally, this is done with manual user feedback (e.g., from data collected by users). Pseudo-relevance feedback, however, is an automatic feedback mechanism that often improves retrieval performance without manual interactions [7]. The Rocchio algorithm is a very popular relevance feedback algorithm which models the feedback information as a vector space model. Hyperspace analogue to language (HAL) is a method for representing and analyzing high-dimensional text data by mapping it into a lower-dimensional space, called a “hyperspace,” in a way that preserves the similarity relationships between the text data [70]. Researchers have also proposed a HAL-based Rocchio model, called HRoc, to better incorporate proximity information to query expansion [71]. Zhu et al. used mixture of relevance models (MRM) [55] for building a clinical IR system for discharge summaries. For query expansion, they derived related terms from a relevance model using pseudo-relevance feedback.

*Multi-modal search* enables searching using both text and visuals, as well as retrieval that includes images, charts, and other illustrations from relevant documents in addition to text. Both text and visual information are included in queries and document representation. The use of techniques from the fields of natural language processing, IR, and content-based image retrieval allows both the text and images to be embedded in queries and document representation. However, not many researchers have attempted to implement multi-modal search systems in the clinical domain. For the scope of time covered in this review, we could only find one such study: one by Demner-Fushman et al. [72] that used a combination of techniques and tools from the fields of NLP, IR, and content-based image retrieval.

**Indexing Methods** The index is one of the key components of an IR system. Indexing is the process of collecting and managing the data, including its storage, to facilitate the efficient IR. In this section, we review different methods for building an IR index found in the literature.

*Inverted indexes* are commonly used in IR systems because they allow for fast and efficient searching of large collections of documents. An inverted index acts as a map between the terms and the corresponding document to which they belong. Numerous papers have been published which used inverted indexing for clinical IR. It is particularly useful for handling full-text searches, in which users enter a keyword or phrase, and the system returns all documents containing that term. Elasticsearch is designed as an inverted index-based search engine to facilitate fast and accurate IR [19]. Technically, the projects built on Elasticsearch are indirectly using an inverted index-based indexing system [21, 22, 40, 73, 74]. In a recent paper, Dai et al. proposed an inverted index-based IR system to find cohorts of patients, with a special focus on family disease history [75].

*Rule-based indexing* is a method of indexing documents in an IR system based on a set of predefined rules or criteria. These rules can be used to classify the EHR documents into categories, or to extract specific information, such as keywords or metadata, from the documents. Rule-based indexing systems typically involve the use of software programs or scripts that are designed to parse the documents and apply predefined rules to extract the relevant information. Edinger et al. experimented with rule-based indexing, developing rules for identifying clinical document Sects. [25]. Rule-based indexing systems can be efficient and reliable, but they can also be inflexible and require significant manual effort to maintain and update the rules as the content of the documents changes. JointEmbed is an IR approach that automatically generates continuous vector space embeddings that implicitly capture semantic information, leveraging multiple knowledge sources such as free-text cases and pre-existing knowledge graphs [76]. JointEmbed was used for the medical CBR task of retrieving pertinent patient electronic health records, where the quality of the retrieval is crucial due to potential health implications.

**Ranking Methods** A ranking model matches queries with the relevant documents and scores each document's relevance with the query. In this section, we discuss about different ranking approaches, ranging from probabilistic models to deep learning–based ranking methods.

Clinical information can be retrieved and synthesized when using semantically similar terms from EHR vectors or embeddings. Vector search is a technique used in IR systems to find documents or other data items that match a given query based on their vector representation. In a vector search, documents are represented as vectors in a high-dimensional space. Various approaches, such as term frequency-inverse document frequency (TF-IDF) and word embeddings, can be used to generate these vectors. The vectors are then used to calculate the similarity between the query and the documents or data items, and the most similar documents or data items are returned as search results.

*Vector space models (VSM),* which use word vectors or embeddings, are used to select similar terms from multiple EHRs and evaluate their performance quantitatively and qualitatively across multiple chart review tasks [77]. VSMs have gained interest recently with the emergence of deep representation models and vector search techniques in IR systems. VSM methods have proved to be efficient in patient identification, which retrieves patient records corresponding to a specific treatment sequence [78]. In order to find similar terms to support chart reviews, researchers introduced a novel vector space model called the medical-context vector space model. It is a collection of clinical terms which are normalized with their frequencies in various medical contexts. VSMs are widely used in open-domain IR systems because they provide a simple and effective way to represent and compare documents and queries. They are also relatively easy to implement and can be used in a variety of different types of clinical IR tasks, including clinical document classification, text similarity, and search.

TF-IDF and BM25 are two of the popular VSM algorithms used in clinical IR. *TF-IDF* is a probabilistic model that reflects how relevant a query word is to a document in a corpus. It is calculated by multiplying the term frequency (TF) of a word by the inverse document frequency (IDF) of the word. The TF of a word is the number of times the word appears in a document, while the IDF is a measure of how common the word is across all documents in the corpus. TF-IDF has been widely used to identify the most important clinical terms or concepts within EHRs [67]. *Okapi BM25* is also a probabilistic ranking model, which compares each word of the query and its number of occurrences in the given document with its frequency in the entire document collection [79]. Although BM25 is based on the principle of TF and IDF, it takes into account factors such as the frequency of the query terms in the document, the length of the document, and the average length of documents in the corpus. It also includes a parameter called $k1$ and $b$ that can be adjusted to fine-tune the ranking function. By default, Elasticsearch uses BM25 ranking algorithm [22, 40, 73, 74], which ensures the scalability of the model by using Elasticsearch's distributed architecture [21]. Hristidis et al. compared a Clinical ObjectRank (CO) system using an authority-flow algorithm which exploits the entities associations in EHRs to discover the most relevant entities. Their results showed that CO outperformed BM25 in terms of sensitivity (65% vs. 38%) by 71% on average, while maintaining the specificity (64% vs. 61%) [38]. VSMs, such as TF-IDF and BM25, have been widely adopted in clinical IR systems due to their ability to effectively rank the relevance of documents to a query. However, it has been noted that these models have limitations in their ability to capture complex concepts and relationships within the text. One of the main limitations of vector space ranking models is their reliance on term frequency and inverse document frequency as the sole measures of relevance. This approach does not take into account the context in which words appear in the text, which can make it difficult to capture subtle nuances and relationships between concepts.

A class of techniques known as *learning to rank (LTR or LETOR)* uses supervised machine learning (ML) to address ranking issues. LTR ranks the document set based on the relative relevance of each document in the corpus [80, 81]. With the recent advancement of deep learning and pre-trained language models (PLM), neural LTR approaches have been adopted in latest clinical IR systems [82]. In their research, Arvanitis et al. proposed a k-nearest document search algorithm to efficiently compute the similarity between two EHRs [83]. In this algorithm, the similarity between two EHRs is measured by comparing their content, represented as a set of features, to the content of other EHRs in the corpus.

RankNet, one of the most popular LETOR algorithms, is a supervised learning algorithm that uses neural networks to learn the ranking function from the relevance judgments. AdaRank is an extension of this algorithm and is a sorting learning algorithm for IR that is particularly useful in the context of clinical IR [84]. It is designed to optimize the trade-off between relevance and diversity of the retrieved documents by iteratively adjusting the weights of the features used to rank the documents based on feedback from relevance judgments. AdaRank uses loss function to measure the difference between the predicted relevance scores and the actual relevance judgments, and it can take into account multiple features such as the text of the documents, the author, the publication date, the source, and many other relevant pieces of information to rank the documents. In many studies, the AdaRank algorithm has proved to outperform VSMs and to be capable of handling the complex and diverse nature of clinical documents like EHRs and improve the performance of clinical IR systems [85].

With the success of deep learning–based contextualized language models, neural IR systems have been developed, which facilitate the use of contextualized embeddings for the task of relevance ranking. *BERT (Bidirectional Encoder Representations from Transformers)* [65] is a contextualized language model, which makes use of the transformer encoder structure with self-attention mechanisms that learns contextual relations between words (or sub-words) in text. BERT-based clinical language models like BioBERT [86] and clinical BERT [87] have enabled researchers to contextualize query and document embeddings for different clinical IR applications including patient cohort retrieval. A query with a patients' target characteristics and document corpus are passed to these language models to retrieve the clinical reports of similar patients [82, 88]. Shi et al. [89] proposed an approach that used lexicon-driven concept detection to identify relevant concepts in sentences from EHRs, and then used these concepts as queries. These queries were used as input to train a Sentence-BERT (SBERT) model. In a recent study [90], the authors explored the use of masking techniques during the fine-tuning stage of BERT for a reading comprehension QA task on clinical notes. The results suggested that transformer-based QA systems may benefit from moderate masking during fine-tuning, likely by forcing the model to learn abstract context patterns rather than relying on specific clinical terms or relations.

*Re-ranking* refers to the process of adjusting the ranking of a subset of documents that were retrieved using an initial ranking function. The initial ranking function, such as TF-IDF or BM25, is applied to the entire corpus of documents. The re-ranking process then focuses on a specific subset of the top *N* documents that were retrieved by the initial ranking function. The goal of re-ranking is to improve the relevance of the top-ranking documents retrieved by the initial ranking function or by taking into account additional information or criteria that were not considered in the initial ranking. Based on expanded search terms and users' feedback, the

**Table 1** Distribution of research papers by methodology type and specific sub-types in clinical information retrieval

| Methodology type | Sub-type | Number of papers |
|---|---|---|
| Query methods | Keyword search | 30 |
| | Semantic search | 18 |
| | CBIR | 11 |
| | Query expansion | 43 |
| | Others | 5 |
| Indexing methods | Inverted index | 55 |
| | Rule-based | 19 |
| | JointEmbed | 3 |
| | Others | 2 |
| Ranking methods | Vector space model | 15 |
| | TF-IDF/BM25 | 72 |
| | Learning to rank | 17 |
| | Neural IR | 12 |
| | Others | 3 |

retrieved outputs are re-ranked to generate the new ranking scores [41]. Thus, clinical IR becomes a two-step process, where (1) the ranked documents are retrieved by the user query and (2) the retrieved documents are retrieved based on the expanded query [55]. Kullback–Leibler (KL) divergence, a measure of the difference between two probability distributions, which can be used as a way to compare the relevance of different documents to a user's query, was used in a study by Yang et al. to compare the similarity of an EHR document's content to the contents of other relevant documents in their clinical IR system [62]. The documents with the lowest KL divergence are considered to be the most similar to the other relevant documents and are ranked higher.

Table 1 presents the count of research papers that have employed various methodologies within Query Methods, Indexing Methods, and Ranking Methods categories in the field of clinical IR, based on the papers reviewed for this study.

While there is a growing interest in using deep learning and language model–based approaches, they are not yet widely adopted in the field. Out of the papers reviewed, only 12 used deep learning methods, and of those, only 5 employed pre-trained language models like BERT. In contrast, 39 papers represented machine learning–based IR methods, and TF-IDF and BM25 together constituted more than 70 papers. This suggests that there is a need for more research in the area of deep learning and language model–based IR in the clinical domain. Such approaches have the potential to improve the accuracy and relevance of retrieval results and thus can play an important role in supporting clinical decision-making.

### 5.4 Evaluation of Clinical IR systems

To measure the efficiency and effectiveness of clinical IR systems, we need the following components:

- A test document collection
- Test query set
- Relevance judgments—labels (relevant or non-relevant) for each query-document pair

Test collections are the most common way to evaluate how well IR technologies work. Test collections are made up of a list of topics or descriptions of information needs, a list of information objects that need to be searched, and relevance judgments that say which information objects are relevant for which topics [91]. The relevance judgments are manually annotated by the domain experts, by labeling each document as either relevant or non-relevant to a particular query. In this section, we first discuss the test collections available for evaluating clinical IR systems. Then we delineate the evaluation matrices used for assessing the performance of clinical IR systems.

The absence of publicly available EHR test collections is a significant barrier to clinical IR evaluation. Patient data cannot be utilized extensively in informatics research due to privacy protection regulations and institutional access restrictions. However, there are two publicly accessible EHR test collections for evaluating clinical IR systems:

- Cohort retrieval dataset from the University of Pittsburgh Medical Center (UPMC) [92]—which was released as a part of the Text Retrieval Conference (TREC) challenge in 2011 and 2012, which will be discussed in the next section. The collection contains 17,264 encounters with 93,551 documents on 34 topics in 2011 and 47 topics in 2012
- Medical Information Mart for Intensive Care-III (MIMIC-III) [93]—a publicly accessible hospital database providing de-identified patient information for about 40,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012

Evaluating the performance of an IR system requires taking into account the entire ranked list of documents returned by the system, instead of a single decision. Table 2 describes the commonly used evaluation metrics, including their definition, formula, and the number of papers using them. Precision@k measures precision, but only among the top k retrieved documents, as opposed to the conventional precision for the complete list. Our study shows that precision@k was used in 57 published clinical IR-related papers, which makes it the most commonly used metric for the evaluation of clinical IR systems. Physicians search only for a single query at a time, so there is only one true positive for each instance of a retrieved document, either relevant or non-relevant. Similarly, recall@k measures recall for top-k retrieved

**Table 2** Evaluation metrics used in clinical IR

| Evaluation metric | Description | Formula | Number of papers | Evaluation toolkit |
|---|---|---|---|---|
| Precision@k | Rate of relevant documents from top k retrieved documents | Precision@k = (True Positives@k)/(True Positives@k + False Positives@k) | 57 | Scikit-learn, Trec_eval, Ir_eval, Rank_eval, Searcheval, Ir_measures, Ndeval, Rankmetrics, Gensim |
| Recall@k/sensitivity/hit rate | Rate of actual relevant documents retrieved from all the relevant results | Recall@k = (True Positives@k)/(True Positives@k + FalseNegatives@k) | 31 | |
| F1@k | Harmonic mean of Precision@k and Recall@k | F1 score = 2•(Precision@k)(Recall@k)/(Precision@k + Recall@k) | 5 | |
| Binary preference-based measure (bpref) | Checks whether relevant documents are ranked above irrelevant ones | bpref = 1/R × $\sum$ [1 – (\|n " ranked higher than" r\|)/R] | 6 | |
| Average precision (AP) | Mean of precision scores after each relevant document is retrieved | AP = [sum of (Precision@k × Relevance of document k)]/ number of relevant documents for the query | 48 | |
| Mean average precision (MAP) | Mean of the average precision (AP) for all queries | MAP = sum of AP of all queries / total number of queries | 48 | |
| Inferred average precision (InfAP) | Average precision as the outcome of a random experiment using a sub sample of the dataset | InfAP = average of the estimated precisions for each relevant document | 25 | |
| Mean reciprocal rank (MRR) | Mean of the reciprocal rank, which is the reciprocal of the rank of the first correct relevant result | For Q queries, MRR = 1/Q × ($\sum$ 1/rank$_q$) | 2 | |
| Discounted cumulative gain (DCG) | Sum of the relevance score normalized by the penalty | DCG at rank position p, DCG$_p$ = $\sum$_(i = 1)^p [rel$_i$/log$_2$(i + 1)] | 7 | |
| Normalized discounted cumulative gain(NDCG) | Measure of the average performance of a search engine's ranking algorithm | NDCG at rank position p, NDCG$_p$ = DCG$_p$/max DCG$_p$ | 14 | |
| Inferred normalized discounted cumulative gain (infNDCG) | NDCG as the outcome of a random experiment using a sub sample of the dataset | InfNDCG = average of NDCGs for each relevant document | 14 | |

results. F1@k combines both precision@k and recall@k as a single metric and is defined as the harmonic mean of the two.

Researchers tend to use other metrics to measure the effectiveness of the retrieval. Average precision (AP) calculates the mean of the precision scores of a single query after each relevant document is retrieved. Since multiple queries are usually used to evaluate a clinical IR system, we use mean average precision (MAP) which is the mean of APs for a batch of queries. We could find 58 published clinical IR-related papers using MAP as one of their evaluation metrics. When working with a large document collection, if a significant number of top-ranked documents have not been judged, it is a challenge to evaluate of the retrieval system's performance using traditional metrics such as precision@k, AP, or MAP. These metrics may not be the best choice in this scenario because they heavily rely on the availability of complete relevance judgments. To overcome this limitation, inferred average precision (infAP), has been proposed as a more robust alternative. It measures the AP on the subset of the ranked list that has relevance judgments and uses those to infer the judgments on the remaining items. Furthermore, when complete judgments are available, infAP is equivalent to actual AP, making it a robust metric for evaluating IR systems in large document collections [94]. This was one of the evaluation metrics used for TREC cohort discovery shared tasks in 2012.

Discounted cumulative gain (DCG) is another metric that considers the relevance and position of the retrieved documents. Manual relevance is assigned to the retrieved documents on a scale that can vary depending on the system being used. This scale ranges from a non-relevant score of 0 to a highly relevant score of 3, with intermediate scores indicating levels of relevance in between. Gain is predicated on the idea that the lower the rank of a relevant document, the less beneficial it is to the user. The value of gain is higher for the top-ranked documents, and it is discounted for lower-ranked documents. Hence, the name "discounted" cumulative gain. Ideal DCG (IDCG) is defined as the DCG value calculated after sorting documents in decreasing order of relevance. Normalized DCG (NDCG) is defined as the ratio of DCG to IDCG, over a set of queries. We observed 7 clinical IR research papers using NDCG as their evaluation metric. NDCG is similar to MAP, but its tail is heavier at higher ranks; it does not discount lower ranks as much as MAP does. Due to this, MAP is often preferred over NDCG for binary outcomes. The inferred NDCG (infNDCG) is defined in a similar way to InfAP, as the NDCG of a subset of the ranked list that has relevance judgments.

### 5.5 Clinical IR Shared Tasks

In recent years, numerous clinical IR-related shared tasks have been initiated to support clinicians and clinical research. All of these shared tasks have the common objective of evaluating clinical IR in as realistic a scenario as feasible and developing novel clinical IR application methodologies. In this section, we briefly describe those shared tasks due to their significant impact on IR research. Though the previous sections encompass the majority of articles on these shared tasks, Table 3 gives

**Table 3** Clinical IR shared tasks

| Shared task | Year | Brief description | No. of participants | Best participant performance | Website |
|---|---|---|---|---|---|
| TREC 2011 medical records track | 2011 | Ad hoc patient cohort retrieval | 29 | bpref = 0.658; P@10 = 0.727; Rprec = 0.500 | https://trec.nist.gov/data/medical2011.html |
| TREC 2012 medical records track | 2012 | Ad hoc patient cohort retrieval | 24 | infNDCG = 0.680; infAP = 0.366; P@10 = 0.749 | https://trec.nist.gov/pubs/trec21/t21.proceedings.html |
| CLEF eHealth shared task 3 | 2013 | Information retrieval to address patients' questions when reading clinical reports | 9 | P@5 = 0.4960; P@10 = 0.5180; NDCG@5 = 0.4391; NDCG@10 = 0.4665; MAP = 0.3108 | https://clefehealth.imag.fr/?page_id=253 |
| CLEF eHealth shared task 3a | 2014 | User-centered health information retrieval—monolingual | 14 | P@10 = 0.756; NDCG@10 = 0.7445 | https://clefehealth.imag.fr/?page_id=449 |
| CLEF eHealth shared task 3b | 2014 | User-centered health information retrieval—multilingual | 2 | P@10 = 0.7551; NDCG@10 = 0.7011 | |
| CLEF eHealth shared task 2 | 2015 | Retrieving information about medical symptoms | 52 | P@10 = 0.5394; NDCG@10 = 0.5086 | https://clefehealth.imag.fr/?page_id=430 |
| CLEF eHealth shared task 3 | 2016 | User-centered health information retrieval | 58 | Not available | https://clefehealth.imag.fr/?page_id=308 |

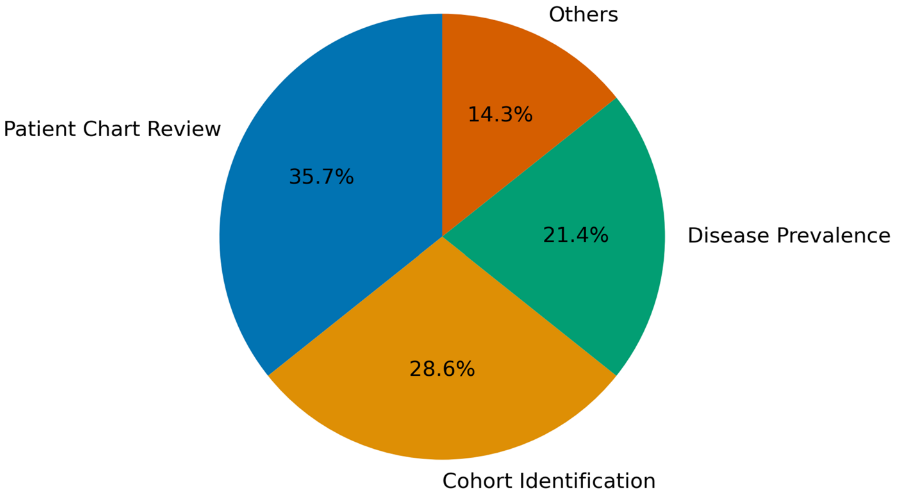

**Fig. 7** Categorization of clinical IR applications

a brief synopsis of the most well-known shared tasks associated with clinical IR research.

Clinical IR has been the topic of several IR conferences, including TREC and CLEF. In 2011 [95] and 2012 [96], TREC offered the Medical Records tracks. CLEF's e-health track had a clinical IR subtask from 2013 to 2016 [97–100]. Note that there have been many additional TREC biomedical IR tracks, some of which could be quite relevant within a clinical setting such as the TREC Clinical Decision Support [101–103] (Precision Medicine [104–107] and Clinical Trials tracks [108]). However, while relevant to clinicians, they do not search over EHR data and are therefore outside the scope of this review.

### 5.6 Applications of Clinical IR

This section will provide a high-level overview of the clinical applications of IR systems. Figure 7 shows the proportion of research papers dedicated to various applications in the field of clinical IR, as identified in our review.

#### 5.6.1 Patient Chart Review

In clinical chart review, clinicians go through EHR notes for searching a particular piece of information of interest [77]. This information could span from the medical history of a particular patient or searching for patients with a specific health condition. Chart reviews are time-consuming and costly because a patient's chart may be composed of hundreds of clinical notes, and the hospital database can contain thousands of patient records. IR techniques have been effectively used to improve the efficiency of this chart review task by using ad hoc search methods. For example, EMERSE, as introduced in a previous section, is a patient chart review system built on IR that has been widely used by clinicians, administrators, and clinical and translational researchers to find relevant information in free-text EHRs.

Recent studies demonstrate that retrieval and synthesis of clinical information can be accelerated by using semantically related terms from various embeddings [77]. In the OpenEHR Archetypes retrieval system, initial search word recommendations were used on a bespoke medical dictionary to find synonyms as replacements for the original search terms [109, 110]. Hanauer et al. [34] developed a MetaMap-based query recommendation algorithm that suggests semantically interchangeable terms based on an initial user-entered query.

Many systems also employ embedding and vector-based search term recommendation methods, which have proven to be more accurate at the expense of system speed. Ye et al. [111] proposed a novel vector space model, the medical-context vector space, to identify similar terms to support chart reviews. As a collection of normalized frequencies of clinical terms in various medical contexts, the medical-context vector space provides information on the relationships between clinical terms. Another study used multiple EHR-based word embeddings and evaluated their performance quantitatively and qualitatively across multiple chart review tasks. The refined terms outperformed the baseline method's (dictionary-based) IR performance (e.g., increasing the average P@5 from 0.48 to 0.60).

### 5.6.2 Cohort Identification and Patient Screening

Patient cohort retrieval refers to the process of identifying and selecting a group of patients from a larger population based on certain criteria or characteristics, such as their diagnosis, treatment history, or demographics. This can be useful in a variety of contexts, such as in clinical research, where patient cohort selection can help to ensure that the study sample is representative of the target population, or in clinical care, where patient cohort retrieval can help to identify patients who may be at risk for certain conditions or who may be candidates for specific treatments.

Cohort retrieval requires the extraction of relevant EHR notes on the basis of a given query. IR methods have made it possible to identify groups of patients in unstructured EHRs based on what the user needs. Li et al. [74] proposed a patient-screening tool using OpenEHR to transform screening conditions into expressions for queries on EHRs. The tool is designed to support queries on EHRs directly within a local context. The Elasticsearch-based tool helps resolve concept mismatches, especially for derived concepts. Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records (CREATE) is a cohort retrieval system that can execute textual cohort selection queries on both structured data and unstructured text, by leveraging the OMOP Common Data Model [112]. This system also uses Elasticsearch as the search engine of the retrieval model, where the data is indexed after identifying medical concepts in the documents using cTAKES (Apache Software Foundation) [113]. Goodwin and Harabagiu [114] proposed a learning patient cohort retrieval (L-PCR) system that uses a relevance model to enhance the quality of patient cohorts retrieved from EHRs by using feedback from physicians. Goodwin and Harabagiu used a learning relevance model (LRM) which exploited the relevance judgments provided by physicians to extract the features of the patient cohort descriptions and match it with the EHRs [114]. Their learning patient cohort retrieval (L-PCR) system can study how physician evaluations can be used to build relevance models that improve the quality of patient cohorts recovered from EHRs thanks to the paired learning-to-rank architecture that the LRM employs.

Recruit is an ontology-based IR system for clinical trials recruitment which uses ontologies to reconcile heterogeneous databases by merging data from structured EHRs with unstructured EHRs [115]. Richman et al. [116] utilized EMERSE to identify patients experiencing food or housing insecurity by utilizing specific keywords and phrases related to these issues. The search engine was used to scan EMRs and retrieve the notes containing specific social determinants of health (SDOH)-related keywords, enabling them to easily identify patients and study the interventions taken.

Siamese network–based embeddings have been successfully used for patient cohort retrieval [117, 118]. The Siamese network based on Time-attention Continuous Bag-of-Word Model (Siamese-Time-CBOW) model was used to obtain patient-phenotype embeddings by calculating the sentence embeddings of each patient's EHR using a time-attention strategy [117]. The model calculates cosine similarity scores between the embedding of a query and the embedding of a patient's EHR data.

Not much research has been done on cohort identification or patient screening using deep learning–based language models. The only work we found was by Soni and Roberts [88], where they proposed a framework for retrieving patient cohorts using transformer language models based on the BERT architecture without the need for explicit feature engineering and domain expertise.

### 5.6.3 Disease Prevalence

Clinical IR has also been applied for predicting the prevalence of a disease or a condition in a population of patients. Hammond et al. [119] used clinical IR on a collection of veteran medical records and demonstrated that text search improves the identification of persons who have attempted suicide in the past by eight to ten times. A similar study was conducted to screen glaucoma patients with poor medication compliance. They utilized EMERSE to search for the terms “noncompliant” and “noncompliance” in the physician notes of eligible patients [30].

Pharmacovigilance is another area in the clinical domain where IR has been effectively employed. Osmont et al. [120] used an IR method for detecting drug-induced anaphylaxis by querying both structured and unstructured data from a clinical data warehouse (CDW). In addition to the 25 cases already identified via spontaneous and DRG reporting for 2012, researchers could identify 41 additional cases using this method.

### 5.6.4 Other Applications

Clinical IR systems are expanding into personalized medicine, allowing for treatments to be tailored based on individual genetic data and personal health records [121]. In clinical decision support, these systems provide real-time, patient-specific information to clinicians, aiding in informed decision-making and improving patient outcomes [122]. Medical education leverages clinical IR systems to access a wide array of educational content, thereby enriching the learning experience for students and professionals [123]. Additionally, public health monitoring relies on clinical IR systems to track and analyze disease patterns and outbreaks, supporting proactive public health responses and surveillance [116]. These varied applications showcase the role of clinical IR systems as vital tools across multiple facets of healthcare.

## 6 Discussion

In this study, we have reviewed the clinical IR literature published between 2010 and 2022. While the literature shows a wide range of applications of IR systems in the clinical domain, a limited number of new research studies on retrieval or ranking methods have been carried out in this area in recent years.

A central issue in clinical IR is the highly complex nature of the clinical language embedded within free-text EHRs. The format, language, and quality of clinical information vary significantly among hospital systems and different users. For instance, one healthcare provider may use technical medical terminology to describe a patient's condition, while another may use simpler, more layman terms. This variation in medical terminology makes it difficult to create and implement large-scale clinical IR systems. Evaluation of IR systems is another bottleneck in the development of novel search or retrieval methods in the clinical domain, due to the limited availability of test collections.

Our review indicates that most clinical IR systems still rely on the BM25 ranking algorithm, with the Elasticsearch search engine supporting their retrieval system. With recent advancements in the field of neural IR, deep learning–based IR systems have shown huge potential to be used for more efficient and accurate retrieval in clinical settings. This study enabled us explore the opportunities for developing new methods for the clinical IR process, especially in querying, retrieval, and ranking. However, one possible obstacle to the wider adoption of methods being developed is the scarcity of good-quality datasets for clinical IR research and development. The TREC cohort retrieval dataset and MIMIC remain the only publicly available EHR datasets. Even though hospitals and research institutions could use internal data, evaluation of these clinical IR systems is still a big challenge. It takes quite a lot of time for annotators to go through the entire patient history, especially for negated conditions and treatments, such as checking if the patient does not have a specific disorder or procedure. This makes it difficult to evaluate the performance of the system with a large number of queries on a fully annotated patient cohort [112]. Moreover, most clinical systems are not tested on external datasets, which raises the question of the generalizability of these systems.

Second, even though the existing clinical IR systems using inverted indices and BM25 may not be the most efficient, they are robust and scalable enough to work on millions of EHR documents in hospital CDWs. The slow training and optimization mechanisms of neural IR and vector search decreases the applicability of these systems to large-scale clinical IR tasks. In clinical settings, the efficiency and accuracy of the retrieved documents can make up for the newer generation of neural IR systems' slower response time (e.g., for cohort retrieval, the relevance of the retrieved patient data is more important than the time taken for the task). Although there may be some initial hesitancy among practitioners and clinical IR researchers to adopt neural IR (which have both hardware and expertise barriers), the practical significance of clinical IR and the potential for a new generation of clinical IR systems makes it highly likely that researchers will adopt deep learning practices for clinical IR.

Third, we could find only a few papers related to the interoperability of clinical IR systems. Interoperability in clinical IR refers to the ability of different systems and applications to communicate and share information seamlessly and the integration of IR systems to fetch the data from these systems. It allows for the integration of data from multiple sources, such as EHRs, lab results, and prescription records from multiple sources. This can help to improve the accuracy and completeness of patient information and can also help to identify potential issues, such as drug interactions or other contraindications, that may impact a patient's care. Additionally, interoperability can help to reduce the risk of retrieving duplicative tests and treatments by ensuring that IR systems have access to a patient's complete medical history.

This study also finds that query expansion strategies dominate clinical IR research more than retrieval models or ranking algorithms. This is because query expansion enables the system to incorporate medical knowledge into the retrieval process. However, retrieval models and ranking algorithms play a critical role in clinical IR systems, along with query expansion strategies. They determine how the system represents and matches the query, which may contain complex clinical terms, and clinical documents and how the system orders and presents the retrieved documents to the user. Therefore, retrieval models and ranking algorithms should be studied with equal importance, along with query expansion strategies.

In addition, we discovered that no research has been conducted to evaluate the bias of clinical IR systems. Bias and fairness are crucial factors in the design and implementation of clinical IR systems. Bias in an IR system can develop when the system favors or disfavors specific user groups or types of information disproportionately. This might lead to unequal access to or representation of clinical information, which can have substantial effects on patient care and decision-making. Multiple sources of bias can influence clinical IR systems, including:

- Data bias: when the data used to train and evaluate the system is skewed, resulting in biased search results
- Algorithmic bias: when the IR system's ranking algorithms are biased, resulting in biased search results
- User bias: when the preferences of the system's users have an effect on the search results, especially during the process of relevance feedback. For instance, if

researchers or medical practitioners are more inclined to a particular gender or ethnic group, the algorithm may be biased to these results over others

### 6.1 Impact of Large Language Models on Clinical IR

The advent of large language models (LLMs) like GPT-3 [124] and LLaMA [125] and domain-specific adaptations such as BioBERT and ClinicalBERT has undeniably revolutionized the field of NLP, opening up new avenues for applications in clinical IR. These models' nuanced understanding of complex language structures and clinical terminology has the potential to greatly enhance IR tasks by providing more accurate, context-aware search results and facilitating the extraction of relevant information from vast repositories of unstructured clinical data. These models can understand queries in a way that mirrors clinical reasoning, taking into account the intricacies of medical conditions and treatments.

However, the integration of LLMs into clinical IR systems is not without challenges. One of the most pressing issues is compliance with healthcare regulations such as the Health Insurance Portability and Accountability Act (HIPAA), which sets stringent standards for the protection of protected health information (PHI). LLMs carry the potential risk of inadvertently disclosing PHI, which is particularly pronounced with API-based LLMs like GPT-3.5, which function remotely and processing data away from the user's control. As such, the development and deployment of LLMs in the clinical IR domain must prioritize the establishment of robust privacy-preserving practices.

The implementation of LLMs in clinical IR systems also introduces concerns regarding the interpretability and explainability of the models' outputs. The "black box" nature of deep learning models can be problematic in clinical settings, where decision-making processes need to be transparent and understandable to healthcare professionals [126]. Addressing these concerns is vital to building trust and ensuring the reliability of LLM-powered clinical IR systems.

Moreover, while LLMs offer sophisticated modeling capabilities, they are also characterized by a trade-off between optimality and generalizability, as indicated by the variations in model performance across different datasets and IR tasks. Machine learning–based models, when trained on specific tasks, tend to outperform heuristic methods but often at the expense of their ability to generalize to new, unseen datasets or tasks without additional training. This highlights the need for a continuous evaluation and fine-tuning of these models to maintain their performance across various clinical contexts and IR tasks.

The potential for LLMs to alter the landscape of clinical IR is clear, yet the path forward must be navigated with caution. As we stand on the cusp of integrating these advanced models into clinical IR, it is imperative to engage in a multidisciplinary dialogue that includes data scientists, clinicians, legal experts, and policymakers. Together, these stakeholders can forge a path that harnesses the strengths of LLMs while ensuring adherence to ethical standards and regulatory requirements, ultimately leading to the development of next-generation clinical IR systems that are both powerful and trustworthy.

### 6.2 Limitations

This study examines the clinical IR literature published during the past 13 years (2010–2022), comprising clinical IR techniques and applications. There are a few limitations to this review. First, the search terms and databases chosen for the review may not have been adequate, which may have introduced inadvertent bias into the review. Second, the search terms yielded papers on clinical recommendation systems, which are distinct from conventional clinical IR systems. Therefore, we excluded these papers after the manual screening process. Thirdly, the review is restricted to English-language articles only and clinical data sources in English only.

## 7 Conclusion

Clinical IR is an important field of study given the enormous amounts of unstructured data generated by modern healthcare, and a number of methods and technologies exist to facilitate this process. There have been significant advances in clinical IR in the last 13 years, driven by the increasing availability of EHRs and other digital health tools. Many healthcare organizations now use EHRs to store and manage patient data, and these systems often include search and recommendation features to help clinicians access relevant information. Despite these advances, there are still challenges in clinical IR. For example, some EHR systems may have limited search functionality or may be difficult to use, making it difficult for clinicians to find the information they need. The Okapi BM25 ranking algorithm is used by the vast majority of clinical IR systems, and there has not been much study into developing more sophisticated ranking tools. While these systems can handle vast amounts of patient data, the trade-off is a compromise in the accuracy and relevance of the retrieved clinical information. With the recent advancements in NLP and pre-trained language modeling in the open-domain, it would seem desirable to explore the integration of such technologies in order to improve upon the current clinical IR systems. We also observed that not much effort has been made to study the evaluation and ranking methods, with the majority of existing studies concentrating on query expansion methods.

Our findings show that more research needs to be done on a next-generation clinical IR system that can use fast semantic vector search and neural IR techniques. The following are characteristics that these systems are expected have:

1. Quick and reliable retrieval: One of the primary reasons why researchers and clinical practitioners continue to use traditional IR is its rapid retrieval capability. Vector search and neural IR must be robust enough to manage millions of EHR records and obtain results in a short amount of time.
2. Interoperability: These systems need to be able to interoperate with other clinical systems and data sources, allowing users to access a wide range of relevant information from multiple sources.
3. Vector search and Neural IR: These systems could use machine learning and deep learning techniques to continuously improve their performance and adapt to new clinical information and user needs.

4. Fair and representative retrieval: Bias estimation and fairness are particularly important in the development of clinical IR, as we need to ensure that retrieved results should be representative of all categories of the patient population.

The state of clinical IR is evolving as new technologies and approaches are developed and adopted. However, there is still room for improvement in terms of the accessibility, usability, and reliability of clinical information. Further study is required to continue enhancing the accuracy and efficacy of current approaches and to design and implement next-generation clinical IR systems.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s41666-024-00159-4.

**Acknowledgements** The authors would like to acknowledge the support from the National Center for Advancing Translational Sciences (NCATS) U24TR004111, the National Library of Medicine (NLM) R01LM014306, the University of Pittsburgh Clinical and Translational Science Institute (CTSI) Pilot Award, the University of Pittsburgh Momentum Funds, and the School of Health and Rehabilitation Sciences Dean's Research and Development Award. Authors KR, WH, and HL would like to acknowledge the support from the NLM R01LM011934.

**Author Contribution** SS: conceptualized the study, wrote the manuscript; HAM: conducted data analysis; DO: conducted data analysis, edited the manuscript; KR: edited the manuscript; WH: edited the manuscript; HL: edited the manuscript; DH: edited the manuscript; SV: edited the manuscript; YW: conceptualized the study, wrote the manuscript.

**Funding** National Center for Advancing Translational Sciences (NCATS) U24TR004111.
University of Pittsburgh Clinical and Translational Science Institute (CTSI) Pilot Award.
University of Pittsburgh Momentum Funds.
School of Health and Rehabilitation Sciences Dean's Research and Development Award.
National Library of Medicine R01LM014306, R01LM011934.

**Data Availability** Data and materials are available in the supplemental files.

## Declarations

**Ethical Approval** Not applicable.

**Competing Interests** The authors declare no competing interests.

## References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit 3(1):1–10
2. Botsis T, Hartvigsen G, Chen F, Weng C (2010) Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010:1
3. Clark KD, Woodson TT, Holden RJ, Gunn R, Cohen DJ (2019) Translating research into agile development (TRIAD): development of electronic health record tools for primary care settings. Methods Inf Med 58(1):1–8
4. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA. 309(13):1351–1352

5. McGowan J, Grad R, Pluye P, Hannes K, Deane K, Labrecque M et al (2009) Electronic retrieval of health information by healthcare providers to improve practice and patient care. Cochrane Database of Syst Rev 3
6. Hersh WR (2020) Information retrieval: a biomedical and health perspective. Springer
7. Zheng J, Yu H (2015) Key concept identification for medical information retrieval. In: Conference on empirical methods in natural language processing, EMNLP 2015. Association for Computational Linguistics (ACL)
8. Ceri S, Bozzon A, Brambilla M, Valle ED, Fraternali P, Quarteroni S (2013) An introduction to information retrieval. Springer, Web information retrieval, pp 3–11
9. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press
10. Tamine L, Goeuriot L (2021) Semantic information retrieval on medical texts: research challenges, survey, and open issues. ACM Computing Surveys (CSUR) 54(7):1–38
11. Himani S, Vaidehi D (2017) A survey on medical information retrieval. International Conference on Information and Communication Technology for Intelligent Systems, Springer
12. Gudivada A, Tabrizi N (2018) A literature review on machine learning based medical information retrieval systems. In: 2018 IEEE symposium series on computational intelligence (SSCI). IEEE
13. Lopes CT (2022) Health information retrieval--state of the art report. arXiv preprint arXiv:220509083
14. Montani S, Striani M (2019) Artificial intelligence in clinical decision support: a focused literature survey. Yearbook of medical informatics 28(01):120–127
15. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F (2019) A survey of word embeddings for clinical text. J Biomed Inform 100:100057
16. Moher D, Liberati A, Tetzlaff J, Altman DG, Group* P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151(4):264–269
17. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N et al (2018) Clinical information extraction applications: a literature review. J Biomed Inform 77:34–49
18. Wongsuphasawat K, Plaisant C, Taieb-Maimon M, Shneiderman B (2012) Querying event sequences by exact match or similarity search: design and empirical evaluation. Interact Comput 24(2):55–68
19. Gormley C, Tong Z (2015) Elasticsearch: the definitive guide: A distributed real-time search and analytics. O'Reilly Media, Inc
20. Grainger T, Potter T (2014) Solr in action. Manning Publications Co
21. Chen DQ, Chen Y, Brownlow BN, Kanjamala PP, Arredondo CAG, Radspinner BL et al (2017) Real-time or near real-time persisting daily healthcare data into HDFS and elasticsearch index inside a big data platform. IEEE Trans Ind Inform 13(2):595–606
22. Filho IB, Sampaio SC, Tenorio JCA, Filho EVDC, Pessoa MEDC, Malaquias RS et al (2020) Development of a health dashboard for an electronic health record system. In: 20th International Conference on Computational Science and Its Applications, ICCSA 2020. Institute of Electrical and Electronics Engineers Inc
23. Chen J, Yu P, Ge H (2005) UNT 2005 TREC QA participation: using Lemur as IR search engine. TREC
24. Ide NC, Loane RF, Demner-Fushman D (2007) Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 14(3):253–263
25. Edinger T, Demner-Fushman D, Cohen AM, Bedrick S, Hersh W (2017) Evaluation of clinical text segmentation to facilitate cohort retrieval. AMIA Annu Symp Proc 2017:660–669
26. Bretonnel Cohen K, Christiansen T, Hunter LE (2011) MetaMap is a superior baseline to a standard document retrieval engine for the task of finding patient cohorts in clinical free text. In: 20th Text REtrieval conference, TREC 2011. Gaithersburg, MD
27. Moen H, Ginter F, Marsi E, Peltonen L-M, Salakoski T, Salantera S (2015) Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. BMC Med Inf Decis Mak 15(Suppl 2):S2
28. Yadav N, Poellabauer C (2012) An architecture for personalized health information retrieval. In: Proceedings of the 2012 International workshop on smart health and wellbeing. Association for Computing Machinery, Maui
29. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K (2015) Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in

developing and using the electronic medical record search engine (EMERSE). J Biomed Inform 55:290–300

30. Hamid MS, Brenneman B, Niziol L, Stein JD, Newman-Casey PA (2020) Identification of glaucoma patients with poor medication compliance from the electronic health record. Investiga Ophthalmol Vis Sci Conf 61(7)
31. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X et al (2018) CogStack-experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Medical Inform Decis Mak 18(1):1–13
32. Wang T, Oliver D, Msosa Y, Colling C, Spada G, Roguski L et al (2020) Implementation of a real-time psychosis risk detection and alerting system based on electronic health records using Cog-Stack. J Vis Exp, JoVE (pagination)
33. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association
34. Hanauer DA, Wu DTY, Yang L, Mei Q, Murkowski-Steffy KB, Vydiswaran VGV et al (2017) Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. J Biomed Inform 67:1–10
35. Perez-Rey D, Jimenez-Castellanos A, Garcia-Remesal M, Crespo J, Maojo V (2012) CDAPub-Med: a browser extension to retrieve EHR-based biomedical literature. BMC Med Inf Decis Mak. 12:29
36. Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N et al (2012) RAVEL: retrieval and visualization in electronic health records. Stud Health Technol Inform 180:194–198
37. Gubanov M, Pyayt A (2012) MEDREADFAST: A structural information retrieval engine for big clinical text. In: 2012 IEEE 13th international conference on information reuse and integration, IRI 2012, Las Vegas
38. Hristidis V, Varadarajan RR, Biondich P, Weiner M (2010) Information discovery on electronic health records using authority flow techniques. BMC Med Inf Decis Mak. 10:64
39. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A (2017) Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc 24(3):607–613
40. Wen A, Wang Y, Kaggal VC, Liu S, Liu H, Fan J (2019) Enhancing clinical information retrieval through context-aware queries and indices. In: 2019 IEEE International Conference on Big Data, Big Data 2019. Institute of Electrical and Electronics Engineers Inc
41. Yang S, Zheng X, Xiao Y, Yin X, Pang J, Mao H et al (2021) Improving Chinese electronic medical record retrieval by field weight assignment, negation detection, and re-ranking. J Biomed Inform 119:103836
42. Bard JB, Rhee SY (2004) Ontologies in biology: design, applications and future challenges nature reviews genetics 5(3):213–222
43. Barcellos Almeida M, Farinelli F (2017) Ontologies for the representation of electronic medical records: the obstetric and neonatal ontology. J Assoc Soc Inf Sci Technol 68(11):2529–2542
44. Bonacin R, Dos Reis JC, Perciani EM, Nabuco O (2018) Exploring intentions on electronic health records retrieval: studies with collaborative scenarios. Ing Syst Inf 23(2):111–135
45. Goodwin TR, Harabagiu SM (2018) Knowledge representations and inference techniques for medical question answering. ACM Trans Intell Syst Technolog 9(2)
46. Gurulingappa H, Müller B, Hofmann-Apitius M, Fluck J (2011) A semantic platform for information retrieval from E-health records. TREC
47. Afzal M, Hussain M, Ali T, Khan WA, Lee S, Kang BH (2014) MLM-based automated query generation for CDSS evidence support. In: Hervas R, Bravo J, Lee S, Nugent C. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer Verlag, p 296–299
48. Hersh WR (1991) Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Med Decis Mak 11(4_suppl):S120–S1S4
49. Koopman B, Bruza P, Sitbon L, Lawley M (2012) Towards semantic search and inference in electronic medical records: an approach using concept--based information retrieval. Australas Med J 5(9):482–488
50. Curé O, Maurer H, Shah N, LePendu P (2013) Refining health outcomes of interest using formal concept analysis and semantic query expansion. In: Proceedings of the 7th international workshop

on data and text mining in biomedical informatics, San Francisco, Association for Computing Machinery

51. Cure OC, Maurer H, Shah NH, Le Pendu P (2015) A formal concept analysis and semantic query expansion cooperation to refine health outcomes of interest. BMC Med Inf Decis Mak. 15(Suppl 1):S8
52. Alonso I, Contreras D (2016) Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. Expert Sys Appl 44:386–399
53. Cureí O, Maurer H, Shah NH, Le Pendu P (2013) Refining health outcomes of interest using formal concept analysis and semantic query expansion. In: 6th International Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS 2013. CEUR-WS
54. Martinez D, Otegi A, Soroa A, Agirre E (2014) Improving search over electronic health records using UMLSbased query expansion through random walks. J Biomed Inform 51:100–106
55. Zhu D, Stephen W, James M, Carterette B, Liu H (2013) Using discharge summaries to improve information retrieval in clinical domain. In: 2013 cross language evaluation forum conference, CLEF 2013. CEUR-WS
56. Aravazhi R, Chidambaram M (2019) An enhanced semantic similarity based information retrieval system in mesh and EMR. J Adv Res Dyn Control Syst 11(9 Special Issue):993–998
57. Liu S, Wang Y, Hong N, Shen F, Wu S, Hersh W et al (2017) On mapping textual queries to a common data model2017. Institute of Electrical and Electronics Engineers Inc
58. Shi W, Kelsey T, Sullivan F (2020) Efficient identification of patients eligible for clinical studies using case-based reasoning on Scottish Health Research register (SHARE). BMC Med Inf Decis Mak 20(1):70
59. Jain H, Thao C, Zhao H (2012) Enhancing electronic medical record retrieval through semantic query expansion. Inf Syst e-Bus Manage 10(2):165–181
60. Wang N, Qi H, Deng Y, Yu W, Chen Z (2022) Transmission and drug resistance characteristics of human immunodeficiency Virus-1 strain using medical information data retrieval system. Comput. 2022:2173339
61. Kreuzthaler M, Pfeifer B, Schulz S (2022) Terminology expansion via co-occurrence analysis of large clinical real-world datasets. In: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)
62. Yang S, Zheng X, Yin X, Mao H, Zhao D (2020) An algorithm of query expansion for Chinese EMR retrieval by improving expansion term weights and retrieval scores. IEEE Access 8:200063–200072
63. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781
64. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)
65. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805
66. Wang Y, Wu S, Li D, Mehrabi S, Liu H (2016) A part-of-speech term weighting scheme for biomedical information retrieval. J Biomed Inform 63:379–389
67. Matsuo R, Ho TB (2018) Semantic term weighting for clinical texts. Expert Sys Appl. 114:543–551
68. Chamberlin SR, Bedrick SD, Cohen AM, Wang Y, Wen A, Liu S et al (2020) Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. JAMIA open 3(3):395–404
69. Zhu D, Carterette B (2012) Improving health records search using multiple query expansion collections. In: 2012 IEEE international conference on bioinformatics and biomedicine, BIBM2012, Philadelphia
70. Rohde DL, Gonnerman LM, Plaut DC (2006) An improved model of semantic similarity based on lexical cooccurrence. Commun ACM 8(627–633):116
71. Pan M, Zhang Y, Zhu Q, Sun B, He T, Jiang X (2019) An adaptive term proximity based rocchio's model for clinical decision support retrieval. BMC Med Inf Decis Mak. 19(Suppl 9):251
72. Demner-Fushman D, Antani S, Simpson M, Thoma GR (2012) Design and development of a multimodal biomedical information retrieval system. J Comput Sci Eng 6(2):168–177
73. Duren R, Smith R, Tackes N, Neeley S, Welsh J, Shirley LX (2018) Scalable assembly of individual patient profiles for clinical trials accrual and research. Cancer Research Conference 78(13 Supplement 1)

74. Li M, Cai H, Nan S, Li J, Lu X, Duan H (2021) A patient-screening tool for clinical research based on electronic health records using OpenEHR: development study. JMIR Med Inform 9(10):e33192
75. Dai X, Rybinski M, Karimi S (2021) SearchEHR: A family history search system for clinical decision support. In: 30th ACM International Conference on Information and Knowledge Management, CIKM 2021. Association for Computing Machinery
76. Metcalf K, Leake D (2018) Embedded word representations for rich indexing: A case study for medical records. In: Cox MT, Funk P, Begum S (eds) 26th international conference on case-based reasoning, ICCBR 2018. Springer Verlag, pp 264–280
77. Ye C, Fabbri D (2018) Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. J Biomed Inform 83:63–72
78. Syed H, Das AK (2016) Vector space models for encoding and retrieving longitudinal medical record data. In: Khan A, Luo G, Weng C, Wang F, Mitra P, Yu C (eds) 1st International Workshop on Data Management and Analytics for Medicine and Healthcare, DMAH 2015 and Workshop on Big-Graphs Online Querying, Big-O(Q) 2015 held in conjunction with 41st International Conference on Very Large Data Bases, VLDB 2015. Springer, Verlag, pp 3–15
79. Robertson S, Zaragoza H (2009) The probabilistic relevance framework: BM25 and beyond. Foundations and trends®. Inf Retr 3(4):333–389
80. Jin M, Li H, Schmid CH, Wallace BC (2016) Using electronic medical records and physician data to improve information retrieval for evidence-based care. In: 2016 IEEE international conference on healthcare informatics, ICHI 2016. Institute of Electrical and Electronics Engineers Inc
81. Huang HH, Lee CC, Chen HH (2014) Mining professional knowledge from medical records. In: 2014 International Conference on Brain Informatics and Health, BIH 2014. Warsaw: Springer Verlag, pp 152–163
82. Mutinda FW, Yada S, Wakamiya S, Aramaki E (2021) Semantic textual similarity in Japanese clinical domain texts using BERT. Methods Inf Med 60(S 01):e56–e64
83. Arvanitis A, Wiley M, Hristidis V (2014) Efficient concept-based document ranking. In: 17th international conference on extending database technology, EDBT 2014. OpenProceedings.org, University of Konstanz, University Library
84. Xu J, Li H (2007) Adarank: a boosting algorithm for information retrieval. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval
85. Zhang P, Wu J (2021) Research on search ranking technology of chinese electronic medical record based on AdaRank. In: 18th international computer conference on wavelet active media technology and information processing, ICCWAMTIP 2021. Institute of Electrical and Electronics Engineers Inc
86. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
87. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al (2019) Publicly available clinical BERT embeddings. arXiv preprint arXiv:190403323
88. Soni S, Roberts K (2020) Patient cohort retrieval using transformer language models. AMIA Annual Symposium Proceedings/AMIA Symposium 2020:1150–1159
89. Shi L, Syeda-mahmood T, Baldwin T (2022) Improving neural models for radiology report retrieval with lexicon-based automated annotation. In: Proceedings of the 2022 conference of the north American chapter of the Association for Computational Linguistics: human language technologies
90. Moon S, He H, Fan JW (2022) Effects of information masking in the task-specific Finetuning of a transformers-based clinical question-answering framework. In: 2022 IEEE 10th international conference on healthcare informatics (ICHI)
91. Scholer F, Kelly D, Carterette B (2016) Information retrieval evaluation using test collections. Inf Retr J 19(3):225–229
92. Chapman W, Saul M, Houston J, Irwin J, Mowery D, Karkeme H et al (2011) Creation of a repository of automatically de-identified clinical reports: processes, people, and permission. AMIA Summit on Clinical Research Informatics, San Francisco, CA
93. Johnson A, Pollard T, Shen L, Lehman L, Feng M, Ghassemi M et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035. https://pubmed.ncbi.nlm.nihgov/27219127
94. Yilmaz E, Aslam JA (2008) Estimating average precision when judgments are incomplete. Knowl Inf Syst 16(2):173–211

95. Bedrick S, Ambert KH, Cohen AM, Hersh WR (2011) Identifying patients for clinical studies from electronic health records: TREC medical records track at OHSU. TREC
96. Voorhees EM, Hersh WR (2012) Overview of the TREC 2012 medical records track. TREC
97. Goeuriot L, Jones GJ, Kelly L, Leveling J, Hanbury A, Müller H et al (2013) ShARe/CLEF eHealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In: CLEF 2013 online working notes, p 8138
98. Goeuriot L, Kelly L, Li W, Palotti J, Pecina P, Zuccon G, et al (2014) Share/clef ehealth evaluation lab 2014, task 3: user-centred health information retrieval. Proceedings of CLEF 2014
99. Palotti JR, Zuccon G, Goeuriot L, Kelly L, Hanbury A, Jones GJ et al (2015) Clef ehealth evaluation lab 2015, task 2: retrieving information about medical symptoms. CLEF (Working Notes)
100. Zuccon G, Palotti J, Goeuriot L, Kelly L, Lupu M, Pecina P et al (2016) The IR task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval
101. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR (2016) Overview of the TREC 2016 clinical decision support track
102. Roberts K, Simpson MS, Voorhees EM, Hersh WR (2015) Overview of the trec 2015 clinical decision support track. TREC
103. Simpson MS, Voorhees EM, Hersh W (2014) Overview of the trec 2014 clinical decision support track. Lister Hill National Center for Biomedical Communications, Bethesda MD
104. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR (2020) Overview of the TREC 2020 precision medicine track
105. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ (2018) Overview of the TREC 2018 precision medicine track
106. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ et al (2017) Overview of the TREC 2017 precision medicine track. In: The text retrieval conference: TREC text REtrieval conference. NIH Public Access
107. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ et al (2019) Overview of the TREC 2019 precision medicine track. In: The text retrieval conference: TREC text REtrieval conference, p 2019
108. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR (2021) Overview of the TREC 2021 clinical trials track. In: Proceedings of the thirtieth text retrieval conference (TREC 2021)
109. Min L, Wang L, Lu X, Duan H (2015) Case study: applying OpenEHR archetypes to a clinical data repository in a Chinese hospital. Studies in health technology and informatics 216:207–211
110. Sun B, Zhang F, Li J, Yang Y, Diao X, Zhao W et al (2021) Using NLP in openEHR archetypes retrieval to promote interoperability: a feasibility study in China. BMC Med Inf Decis Mak. 21(1):199
111. Ye C, Malin BA, Fabbri D (2021) Leveraging medical context to recommend semantically similar terms for chart reviews. BMC Med Inf Decis Mak. 21(1):353
112. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F et al (2020) Implementation of a cohort retrieval system for clinical data repositories using the observational medical outcomes partnership common data model: proof-of-concept system validation. JMIR Med Inform 8(10):e17376
113. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC et al (2010) Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
114. Goodwin TR, Harabagiu SM (2018) Learning relevance models for patient cohort retrieval. JAMIA open. 1(2):265–275
115. Patrão DF, Oleynik M, Massicano F, Morassi SA (2015) Recruit-an ontology based information retrieval system for clinical trials recruitment. In: MEDINFO 2015: eHealth-enabled health. IOS Press, pp 534–538
116. Richman EL, Lombardi BM, de Saxe ZL, Forte AB (2022) What do EHRs tell us about how we deploy health professionals to address the social determinants of health. Soc. 37(3):287–296
117. Kong N, Wang Y, Wang J, Tao X, Zhou Y (2020) Time-attention medical concept embedding and query representation for cohort selection. Basic Clin Pharmacol Toxicol 126(Supplement 4):10–11
118. Xiao C, Gao J, Glass L, Sun J (2020) Patient trial matching using pseudo-siamese network. J Clin Oncol Conf 38(15)
119. Hammond KW, Laundry RJ, O'Leary TM, Jones WP (2013) Use of text search to effectively identify lifetime prevalence of suicide attempts among veterans

120. Osmont MN, Bouzille G, Triquet L, Rochefort-Morel C, Polard E, Cuggia M (2017) Drug safety and big clinical data: detection of drug-induced anaphylactic shocks (BREIZH project). Fundam Clin Pharmacol 31(Supplement 1):32
121. Selvan NS, Vairavasundaram S, Ravi L (2019) Fuzzy ontology-based personalized recommendation for internet of medical things with linked open data. J Intell Fuzzy Syst 36(5):4065–4075
122. Dentino B, Davis D, Chawla NV (2010) HealthCareND: leveraging EHR and CARE for prospective healthcare. In: Proceedings of the 1st ACM international health informatics symposium
123. Orenstein EW, Rasooly IR, Mai MV, Dziorny AC, Phillips W, Utidjian L et al (2018) Influence of simulation on electronic health record use patterns among pediatric residents. J Am Med Inform Assoc 25(11):1501–1506
124. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P et al (2020) Language models are few-shot learners. Adv Neural Inf Proces Syst 33:1877–1901
125. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T et al (2023) Llama: open and efficient foundation language models. arXiv preprint arXiv:230213971
126. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29(8):1930–1940

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Sonish Sivarajkumar[1] · Haneef Ahamed Mohammad[2] · David Oniani[3] · Kirk Roberts[4] · William Hersh[5] · Hongfang Liu[4] · Daqing He[2] · Shyam Visweswaran[1,6,7] · Yanshan Wang[1,3,6,7]**

✉ Yanshan Wang
yanshan.wang@pitt.edu

[1] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

[2] Department of Information Science, University of Pittsburgh, Pittsburgh, PA, USA

[3] Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA

[4] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

[5] Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

[6] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[7] Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA