Semantic Information Retrieval On Medical Texts: Research Challenges, Survey and Open Issues

LYNDA TAMINE, University of Toulouse Paul Sabatier, IRIT Laboratory, Toulouse 31062, France
LORRAINE GOEURIOT, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France

The explosive growth and widespread accessibility of medical information on the Internet have led to a surge of research activity in a wide range of scientific communities including health informatics and information retrieval (IR). One of the common concerns of this research, across these disciplines, is how to design either clinical decision support systems or medical search engines capable of providing adequate support for both novices (e.g., patients and their next-of-kin) and experts (e.g., physicians, clinicians) tackling complex tasks (e.g., search for diagnosis, search for a treatment). However, despite the significant multi-disciplinary research advances, current medical search systems exhibit low levels of performance. This survey provides an overview of the state-of-the-art in the disciplines of IR and health informatics and bridging these disciplines shows how semantic search techniques can facilitate medical IR. First, we will give a broad picture of semantic search and medical IR and then highlight the major scientific challenges. Second, focusing on the semantic gap challenge, we will discuss representative state-of-the-art work related to feature-based as well as semantic-based representation and matching models which support medical search systems. In addition to seminal works, we will present recent works that rely on research advancements in deep learning. Third, we make a thorough cross-model analysis and provide some findings and lessons learned. Finally, we discuss some open issues and possible promising directions for future research trends.

CCS Concepts: • **Information systems** → **Evaluation of retrieval results**; **Retrieval effectiveness**.

Additional Key Words and Phrases: Information Retrieval, Medical Texts, Knowledge Resources, Relevance, Evaluation

ACM Reference Format:
Lynda Tamine and Lorraine Goeuriot. 2020. Semantic Information Retrieval On Medical Texts: Research Challenges, Survey and Open Issues. *ACM Comput. Surv.* 1, 1 (April 2020), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 From Information Retrieval To Semantic Information Retrieval

"Information retrieval (IR) deals with the representation, storage, organization and access to information items" [6]. There are two main processes in IR. Indexing mainly consists of building computable representations of content items using metadata, while retrieval is the process of matching queries to documents to optimize relevance. Relevant pointers to previous introductory bibliographic resources that can help the reader get started in IR are [6, 33, 59, 97, 139]. IR models have been developed in stages over almost 60 years (since 1960s) evolving notably from the Boolean [138], vectorial [139], probabilistic [99, 135], language model (LM) [82], learning-to-rank (LTR) [92] models and more recently neural models [105]. Those models offer various forms of

Authors' addresses: Lynda Tamine, University of Toulouse Paul Sabatier, IRIT Laboratory, 118 route de Narbonne, , , Toulouse 31062, France; Lorraine Goeuriot, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, , Grenoble, 38000, France.

0360-0300/2020/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

formal support to set up a matching between query and documents for which the key output is an algorithmic score of document relevance. Traditionally, both query document indexing and retrieval rely on a lexical approach based on bag-of-words and this still operates as the main processes of modern search engines. Lexical representation and matching suffer however from semantic gap and vocabulary mismatch issues induced by the complex problem of language understanding. These issues are the most critical ones in search and have attracted much attention [12, 27, 88, 93, 141, 157, 186]. While the semantic gap refers to the difference between the computable representations of a document and its conceptual meaning, vocabulary mismatch refers to the difference between the lexical representations of two semantically close candidate texts, here a document and a query. The most prominent approach proposed so far to close the semantic gap and vocabulary mismatch falls under the umbrella of semantic search [12]. In the IR area, semantic search mainly consists of enhancing query and document representations to increase their level of understandability and performing a more meaningful query-document matching driven by semantics [88]. Many techniques are proposed for this purpose ranging from semantic query expansion to semantic relevance ranking that are based on a separate or combined use of key semantic resources. The latter fall into two main categories: (1) structured knowledge resources (e.g., ontologies, thesaurus, knowledge graphs) that provide human-established real-world knowledge which is crucial to relate words and their associated meanings. Examples of such resources make use of words/terms[1] and concepts/entities to organize relational knowledge in general domains, such as WordNet and DBPedia, or in specialized domains, such as MeSH used in the medical domain; (2) unstructured knowledge in the form of raw textual corpora from which semantic representations of words, phrases, and documents are automatically built and also semantic relationships between words automatically established.

1.2 Medical IR: Specificities and Challenges

Today, there is wider access to medical information whose forms are constantly increasing. We can distinguish between knowledge-based information, which is derived from observational research, and patient-specific information, such as radiology reports, progress notes etc. Among the specificities of medical IR, making it significantly different from other domain-specific IR applications (e.g., legal IR, mobile IR), we mention the following: (1) the high diversity of users including laypeople (e.g., patients and their families), medical professionals and clinicians. The diversity of user profiles is characterized by a significant difference in both their domain-knowledge and their level of expertise which has a huge impact on their vocabulary and assessments of system results [119, 159, 175]; (2) the high diversity of tasks such as diagnosis search, advice and support search and patient cohort search. All these tasks lead to different relevance requirements; (3) the strong relationship between users' web search activities and their personal daily life including, but not limited to, health care utilization (HCU) [177, 188], and psychological attitudes [176, 178]. Taking the peculiarities of medical search mentioned above, we identify three main challenges that encompass key-related issues:

- **Semantic gap**: the core issue in the representation of meaning is to deal with the *semantic gap* that refers to the difference between the low-level description of texts and their high-level sense mainly because of the complexity of human language. The semantic gap is particularly challenging in medical texts because of the high variability of language and spelling, frequent use of acronyms and abbreviations, and inherent ambiguity for automated processes to interpret concepts according to document contexts [42, 61], the presence of negation and time factors [5, 90]. Through a failure analysis performed on the evaluation results of a clinical search task for cohorts, Edinger et al. [42] reported that the reasons for most precision and recall failures are related to bad lexical representations, presence of negation and time constraints. Examples of failure include: "*Notes contain very similar term confused with topic*", "*Topic terms denied or rule out*", "*Topic symptom/condition/procedure done in the past*".

- **Vocabulary mismatch**: In IR, this issue also called the *lexical gap*, occurs when the vocabulary of the query does not match the vocabulary of relevant documents, causing low recall. Vocabulary mismatch occurs at the retrieval stage and is mostly induced by the semantic gap problem. Poor representation of the meaning of queries and documents at the indexing stage are likely to manifest their mismatch at the retrieval stage. Many studies outlined the effect of vocabulary mismatch in medical IR [38, 42, 56, 81].

[1] Compound words

The study performed by Edinger et al. [42] also revealed numerous reasons for failure that fall into the vocabulary mismatch problem category, such as "*Visit notes used a synonym for topic terms*", "*Topic terms not named in notes and must be inferred*". Another possible reason for vocabulary mismatch is the difference in expertise levels of users. Previous studies showed that medical domain expertise plays an important role in the lexical representation of information needs [119, 159, 175]. For instance, Tamine and Chouquet [159] found that experts formulate longer queries and make use of more technical concepts than novices leading to significant differences in system documents' rankings for similar information needs.

- **Complexity of result appraisal**: Appraisal of findings in the medical domain is the process of examining research evidence to judge its trustworthiness, value and relevance in a particular context (e.g., patient and his conditions). From the medical IR system perspective, result appraisal is the process of assessing the relevance of search results w.r.t. a particular query issued in a particular task context. Numerous studies showed that relevance assessment in the medical domain is a time-consuming and a cognitively expensive process [80, 159, 168]. Result appraisal complexity can be explained by: (1) the difficulty of interpretation of document content caused by: ambiguity in context, specificity of language, temporality on relevance [80, 159], low level of domain-expertise of users [119, 159, 175]; (2) the variability in the perception of relevance for a given task. Previous work argues that there is a strong relationship between the task (and the doer) peculiarities and the type of expected relevance leading to significant variability of performance levels across tasks [160, 189].

By focusing on semantic IR in the medical domain, this survey is restricted to review the prominent solutions proposed so far to specifically tackle the semantic gap and the vocabulary mismatch challenges.

1.3 Motivations, Scope and Contributions of this Survey

Motivations and Scope. These last years, both semantic search and medical search have become emerging topics, addressed in different scientific communities with a variety of perspectives and methods.

In the IR community, medical search represents a domain-oriented application research topic where the major problem studied is how to better understand complex information needs with a medical faceted intent and how to map them to documents to improve the likelihood of relevance. Medical search increasingly appears in special sessions of several IR and information-seeking conferences such as SIGIR[2], CIKM[3], and ECIR[4] as well as workshops such as the *Medical IR* workshops at SIGIR (2013-2015) and special issues of journals (e.g., JIR[5] [51], JASIST[6] [107]). In the health informatics field, medical search is viewed as the application of information science and information technology findings and principles to healthcare and everyday wellness. It is a fundamental research topic addressed within this scientific community through numerous special tracks at major conferences (e.g., "Ontologies and knowledge representation and access" track at AIME'19[7]) as well as workshops (e.g., "Semantic extraction from Medical Texts" at AIME 2017, "Knowledge Representation for Health" at AIME'19) and major journals in the area such as JAMIA[8][128], BMC Medical informatics & Decision Making [106, 120] and Medical informatics [19, 100, 187].

These research activities generally follow a mono-disciplinary approach which hampers the development of collective knowledge to enhance overall research progress in the broad area of medical information access. With this in mind, the main motivation behind this survey is to provide a comprehensive review on the hot topic of semantic search on medical texts, targeting varied audiences from the IR and health informatics disciplines. More precisely, we have two audiences in mind: (1) newcomers as well as mature researchers in the IR community interested in the medical application domain; (2) academic researchers from the health informatics community and companies interested in the design of clinical decision support systems. The ideas

[2] Special Interest Group on Information Retrieval
[3] Conference on Information and Knowledge Management
[4] European Conference on Information Retrieval
[5] Journal of Information Retrieval
[6] Journal of the American Society on Information Science and Technology
[7] Artificial Intelligence in MEdicine
[8] Journal of the American Medical Informatics Association

and methodologies presented in the survey may help industrial practitioners to turn the research results into products. To sum up, providing a single coherent resource as a point of reference for these two different audiences would be highly useful. Our additional motivation is to provide a compiled view of the current landscape of neural approaches of semantic search research in the medical domain.

Given, on the one hand, the wide range of tasks that can be performed on medical texts and, on the other hand, our aim to make a thorough review of information retrieval, we only consider in this survey the following: (1) raw textual documents as the focal units of search. This survey does not cover search on images and other documents that have an inherently non-textual representation; (2) ranking and similarity tasks, and to a lesser extent , classification and clustering tasks also studied in the IR discipline. Ranking or similarity tasks consist respectively of selecting from a corpus a ranked list of search units that better match a textual query, searching for similar or related items such as documents and concepts. Illustrations of these core tasks in the medical domain include, but are not limited to, search for cohorts, patient search and similarity, care episode search, diagnosis and concept classification and clustering. This survey does not cover major Natural Language Processing (NLP) tasks such as named entity recognition (NER), relation extraction, relation classification, entity/relation-based indexing and patient data de-identification which are core auxiliary tasks that support or can help medical IR tasks.

Related Surveys and Contributions. Although several valuable introductory readings, tutorials, and surveys [12, 43, 44, 60, 182] already exist for medical information access, management and mining at large on the one hand and semantic search [12, 88] on the other hand, we are not familiar with any existing literature review surveying research progress made in semantic IR within the medical domain. The primary objective in [44] is to present computational aspects of big data in health informatics. The authors provide a comprehensive overview of health data management covering data capturing, storing, sharing, analyzing, searching, and decision support. In their paper, search techniques are addressed from the data mining community view to find through unstructured and structured medical data a useful pattern. Other surveys [43, 182] focused on data mining techniques for modeling electronic health records (EHRs) and related standards that structure the clinical content for semantic inter-operability. In [182], the authors mentioned (Section 1. page 1:3) that techniques related to mining semi-structured or unstructured data through and IR are excluded from the survey because the technical challenges they pose are different from the challenges faced in mining structured EHR data. Closer to our survey but significantly different both in the scope and the focus, Hersh [60] surveyed methodologies of indexing and retrieving medical documents. However, the author emphasized lexical matching techniques based on bag-of-words text representations. Additionally, there has been considerable progress made in medical IR since the publication of the aforementioned survey. Regarding specifically the semantic view of IR, a significant body of work proposed: (1) new hybrid approaches for combining multiple contexts and types of knowledge to mine information and knowledge from medical texts; (2) new methodologies for building rich semantic representations of complex unstructured medical texts (e.g., patient records); (3) new learning-based approaches of relevance estimation and new related evaluation protocols. Therefore, a new systematic review of the state-of-the-art is needed.
A structured comprehensive review of semantic search on texts and knowledge bases is provided in [12] from the semantic web perspective. This survey is a relevant introductory reading of our survey providing the preliminary concepts and technologies that are required to understand the various approaches of semantic search in the medical domain, particularly through keyword search approaches.

Regarding the previous literature reviews mentioned above, we make several contributions in this survey. First, we perform a detailed and thorough investigation of the state-of-the-art in knowledge-base driven and data-driven approaches to medical text representation and matching for relevance ranking and relevance learning purposes. Second, we we make a through cross-model analysis and provide some findings and lessons learned. Finally, we provide a list of promising research directions intended for both IR and health-informatics audiences.

Survey Methodology. We adopt a top-down approach to surveying the literature about semantic search in medical texts. As a result, our survey is structured as a reflection of a well-established practice in the IR

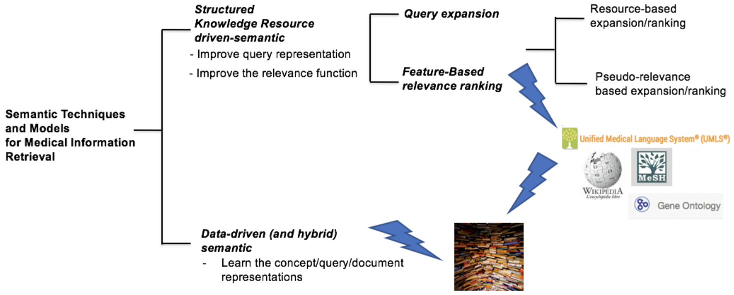

Fig. 1. Structure of the survey

community to use semantic knowledge resources with the goal of accurately either modelling or matching the textual units about the (medical) domain for which it is intended to be used. There are two distinct approaches used to extract and represent in-domain knowledge in the IR area. The first is driven by domain experts, using structured knowledge resources that provide human-established real-world knowledge. Such knowledge is crucial to relate words and their associated in-domain meanings (Section 3). The second is rather driven by domain-knowledge (e.g., word and concept meanings and associated relationships) that are captured from raw textual corpora through hidden linguistic regularities and patterns (Section 4).

As shown in Figure 1, we classify accordingly the research work carried out in the area into two main categories: (1) *structured knowledge resources-based approaches* that mainly include query expansion (QE) techniques and feature-based relevance ranking models; (2) *data-driven approaches* whose objective is either to learn from raw textual data item (word, concept, document, patient, etc.) representations that could be incorporated in a ranking, classification or relevance prediction model.

The motivation behind our categorization is twofold: (1) these categories rely on radically different approaches for both text representation and matching. While knowledge-driven models rely on the explicit use of items from structured knowledge, data-driven models rely on learning-based techniques of item representations over unstructured knowledge; (2) while the knowledge-driven model design is the earliest line of work that leverages IR theoretical results that have matured over time, data-driven models are still in their early development. Thus, we believe that this categorization is useful for providing generalizations, specifically explaining contributions that fall in the same approach and also identifying gaps and pointing out pending issues.

For each of these two lines of work, we collected using the appropriate keywords, publications in major conferences mostly in the period 2010-2020 (e.g., SIGIR, CIKM, ECIR), journals (e.g., IP&M[9], JASIST) and working notes from participants at major evaluation challenges (e.g., TREC[10], CLEF[11]) in IR, as well as in major conferences in health informatics (e.g., AMIA, AIME) and journals (e.g., Journal of the American Medical Informatics Association, BMC Medical informatics & Decision Making, Medical informatics and Journal of Biomedica Informatics). Based on each article, we enlarged the set of publications by following citation links and selecting representative works based on bibliometrics or work's relevance and specificity leading us to collect publications from other various sources such as ACL[12], IJCAI[13] and AAAI[14]. After a manual review of all the cited papers, we finally included in this review 204 among which 176 are key-related papers and articles.

The rest of this survey is organized into five sections. Section 2 introduces some resources for medical IR including information and evaluation tools. Section 3 reviews knowledge-resource driven approaches

[9] Information Processing & Management
[10] Text Retrieval Conference
[11] Cross-Language Evaluation Forum
[12] Association of Computational Linguistics
[13] International Joint Conferences on Artificial Intelligence
[14] Association for the Advancement of Artificial Intelligence

for semantic search on medical texts, while Section 4 puts the focus on data-driven approaches. Section 5 summarizes the article and then discusses promising research directions.

2 RESOURCES FOR MEDICAL IR: TOOLS AND BENCHMARKS

2.1 Medical Text Semantic Annotation and Resources

In this section, we describe the semantic resources and the main methods to enrich semantically text.

2.1.1 Structured Knowledge Resources. The medical community works towards building rich knowledge resources for several decades now. Knowledge resources have been built for several applications such as label each electronic health record issued by a hospital with the major disease/finding, index biomedical documents, identify named entities in a text, etc. We list in this section a few thesauri and then introduce a meta-thesaurus initiative. A detailed description can be found in [144].

The International Classification of Diseases (ICD)[15] is a "standard diagnostic tool for epidemiology, health management and clinical purposes"[16]. It is maintained by the World Health Organization (WHO) and is designed as a healthcare classification system: it aims at assigning diagnosis codes for various disorders, symptoms, etc. The ICD is used worldwide for statistics on morbidity and mortality, by the hospitals for billing and as a decision making support tool for medical professionals.

Medical Subject Headings (MeSH)[17] is a controlled vocabulary created to index biomedical literature. It has been created and is maintained by the National Library of Medicine (NLM). It is used to index all the documents contained in MEDLINE, a database of biomedical and life sciences articles, which contains more than 25 million references[18]. MeSH entries are defined with a unique identifier, a short description or definition, links to related descriptors, and a list of synonyms (known as entry terms). MeSH contains 28,000 descriptors (concepts or entities), with over 90,000 entry terms. It has 3 types of relationships: hierarchical, synonymous, and related. MeSH contains 16 categories, such as: *'anatomy'*, *'diseases'*, or *'chemical and drugs'*. **The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)** [19] is a multilingual clinical healthcare terminology. Its purpose is to enable clinicians to record data with enhanced accuracy and consistency. In 2020, the release of the SNOMED CT International Edition included 352,567 concepts.

The Unified Medical Language System (UMLS)[20] has been initiated in 1986 by the National Library of Medicine (NLM) to provide a mechanism to link existing medical thesaurus and controlled vocabularies. It contains a metathesaurus (integrating more than 100 thesauri/vocabulary), a semantic network, and NLP tools. UMLS contains over 1 million biomedical concepts and 5 million terms and covers multiple languages [20]. But most of the non-English languages are far from the coverage of English [109, 174, 197]. A study done in 2006 [98] shows that only 27% of UMLS concepts are non English ones. Several national or international initiatives have gathered or created knowledge resources in languages other than English, such as[21]: **UMLF**, a unified medical lexicon for French [204]; **UTCMLS**, a unified traditional Chinese medical language system [197]; **MedLexSp**[22], a unified medical lexicon for Spanish [23]; **TRIMED**[23], a trilingual terminological resource in English, French and Italian [166].

2.1.2 Semantic Annotation of Medical Texts. Semantic annotation is the process by which knowledge can be added to raw text. It consists in adding information (through annotations) in texts at several linguistic levels [86, 144]: grammatical (morphological and syntactical), semantic, pragmatic. The result of the semantic annotation is an enriched text, which serves as the essential basis of any semantic task.

Semantic annotation consists of linking documents to knowledge bases by identifying:

[15] https://www.who.int/classifications/icd/en/

[16] http://www.who.int/classifications/icd/en/

[17] https://www.ncbi.nlm.nih.gov/mesh/

[18] https://www.nlm.nih.gov/bsd/medline.html, accessed Dec. 2019

[19] http://www.snomed.org/snomed-ct/

[20] https://www.nlm.nih.gov/research/umls/index.html

[21] We list only some examples here. Providing a comprehensive list is out of the scope of the paper.

[22] http://www.lllf.uam.es/ESP/nlpdata/wp1/

[23] http://shiny.dei.unipd.it/TriMED/

- **Entities/concepts in the document**: this task is called *named entity recognition*. For example, the sentence "the female patient suffers from TIA and high blood pressure" contains the entities *female patient* (UMLS CUI C0150905), *suffers* (UMLS CUI C0683278), *TIA* (UMLS CUI C0007787), *high blood pressure* (UMLS CUI C0020538)
- **Relationships (implicit or explicit) between entities**: this task is called *relation extraction*, and often requires entities to be identified. For example, the sentence "HK1 gene involved in glycolytic process" contains 2 entities (*HK1 gene*, CUI C1415554 and *glycolytic process, CUI C0017952*), and their relationship is characterized by the verb *involved*.
- **Relationship between an entity and a document**: this task consists in assigning one or several entities/concepts to a global document. This task can be considered as a document classification task (e.g., MeSH entities for indexing documents on MEDLINE: *PMID:3207429 is indexed with* `Glucose/metabolism` *and* `Hexokinase/genetics`

Semantic annotation can be done manually: it is the case of data curation, that includes some annotation phase; manual indexing is pretty common in the medical domain, and is done for instance on MedLine documents [4]; ICD coding in hospitals is also manually assigned. It can also be done automatically, as a Named Entity Recognition task, relation extraction task, or classification task. In the remainder, we describe briefly the NER process and a few open-source tools. Biomedical NER aims at identifying automatically medical terms in text [144]. It is a three-step process: (1) determine the entity's substring boundaries; (2) assign the entity to a concept category; (3) assign the entity to the concept identifier in the knowledge base (called entity normalization).

There are several open-source softwares for biomedical named entity recognition with UMLS. **Metamap**[24] is one of the key players as it has been created by the National Library of Medicine to annotate biomedical literature [3]. CTakes is an open-source system for processing clinical free-text data [140]. **CTakes** is based on the Unstructured Information Management Architecture (UIMA) framework and includes a NER system specifically trained for clinical data[25]. The processing pipeline is very similar to Metamap, as it includes linguistic pre-processing of the text, span identification, and concept mapping with the UMLS. Metamap and CTakes are widely used by the community but require considerable time to annotate an entire corpus of biomedical text [148]. **QuickUMLS** is a python-based tool performing approximate matching to UMLS terms and allowing to annotate corpus more efficiently[26].

2.2 Evaluation of Medical IR

2.2.1 Evaluation methods. There are several ways to evaluate IR systems:

Laboratory-based evaluation consists of testing and comparing systems within a laboratory environment that does not change. Such an evaluation was introduced by Cyril Cleverdon in the Cranfield College of Aeronautics [34], where they were conducting retrieval experiments on test databases in controlled settings. These *Cranfield tests* are still used in most of the academic evaluation settings and are the founding basis of most of the evaluation challenges.

User-based evaluations come from the interactive IR domain and aims at measuring user satisfaction by getting feedback on the search systems from real users in laboratory settings [75].

Online evaluations is an alternative to laboratory-based evaluations. It consists of observing real users engaging with the system, and interpreting their actions to measure search systems effectiveness [67].

Both user-based and online evaluation allow getting direct feedback from users. This form of feedback is very useful to assess the *usability* of the systems, how users interact with it [40, 57]. They are used to measure user satisfaction, but can hardly evaluate a system's performances. In this survey, we will focus on search systems effectiveness, rather than usability. Therefore, we will focus on laboratory-based evaluations: they consist in comparing several systems on a fixed set of test collections. Test collections contain: *a set of topics* (users' query enriched with information such as a textual description of the information need); *a document*

[24] https://metamap.nlm.nih.gov

[25] https://ctakes.apache.org/

[26] https://github.com/Georgetown-IR-Lab/QuickUMLS

Venue	Task	Dataset	Activity
TREC	Genomics adhoc retrieval	Clinical information need Biomedical articles	2003-2005
	Genomics passage retrieval	Clinical information need Biomedical articles	2006
	Medical records	Patient cohort search	2011-2012
	Clinical decision support / Precision medicine	Case reports Biomedical articles	2014-
CLEF	ImageCLEF medical retrieval	Image and medical reports Collection of medical images	2003-
	CLEF eHealth consumer search	Health information need Large web crawl	2013-
	CLEF eHealth technological assisted reviews	Boolean queries Biomedical articles	2017-2019
Changing venue	BioASQ	Annotated biomedical abstracts and QA dataset	2013-

Table 1. Summarized view of the IR benchmarking activities in the medical domain

collection; *Relevance judgments* (manual construction of pairs of (d, q) indicating that document d is relevant to query q. This manual assessment is usually carried out on a pool of the document collection.)

Evaluation of search systems requires to measure (1) how well the system predicted which documents are relevant and which ones are not (precision and recall); (2) how well did it rank the resulting documents. Measures taking into account documents rank (2) are also based on the precision, integrating the rank of the documents predicted as relevant, e.g., Precision @r (noted P@r), Mean Average Precision (MAP) [167], normalized discounted cumulative gain (NDCG) [69]. The evaluation of medical IR uses the same metrics as for classical IR, which depends on the search task. Classically, IR systems have considered the topical relevance: if the document is on the same topic as the query, it is relevant. In reality, relevance has many other dimensions [193] such as reliability, novelty, readability, etc. Assessing a document's relevance wrt a topic, therefore, consists in assessing each relevance dimension considered. In the medical domain, search systems should provide patients with: readable and understandable documents and the information it contains should be reliable. Medical professionals should be provided with documents containing up-to-date information, and properly cover the topic searched. While the abovementioned metrics only take into account one dimension of relevance, some metrics such as the uRBP allow to integrate several dimensions, such as the understandability [202].

2.2.2 Benchmarking activities and Test Collections. Various conferences and organizations propose evaluation challenges, which purpose is to provide a framework for testing IR systems in a similar framework and with a similar dataset. To do so, each evaluation challenge tackles a particular search task, provides a test collection to its participants, and evaluates the results of the submitted systems. The datasets are often shared with the community once the challenge is over. These conferences are at the origin of the creation of many quality test collections that are widely used in the literature. We describe in this section the benchmarking activities, the test collections issued from them, as well as other test collections used in the literature.

Table 1 provides a summary of all the benchmarking activities focusing on medical IR. Most of these datasets are available for the community and broadly used for evaluating medical search systems in the literature. The tasks that were not IR centered (i.e. classification tasks, NLP tasks) are omitted.

In the remainder, we list the test collections used in the literature, issued from benchmarking activities or publicly released.

- *TREC Filtering Track* ran from 1996 until 2002. In 2000, the task introduced a medical dataset, with the purpose of improving the ability of systems to build persistent user profiles which successfully retrieve relevant documents [134].
  - *Documents*: a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). This document collection is known as the OHSUMED collection [61]

  - *Topics*: the 63 selected queries were manually built by medical experts.
  - *Relevance judgement*: three-point scale manual assessment of the relevance conducted by medical professionals
  - *Metrics*: P@N, linear utility, mean average precision

- *TREC Genomics*[27] *adhoc* ran from 2003-2007 [62–66] and proposed a range of tasks in the domain of genomics using IR, document classification, etc. The *Adhoc retrieval task* (2003, 2004, 2005) aimed at retrieving biomedical literature citations for varied clinical information needs.
  - *Documents*: 10 years of completed citations from the database MEDLINE from 1994 to 2003, with a total of 4,591,008 records.
  - *Topics*: built manually from interviewing experts on their information needs. Topics relate to gene names (2003), and information needs statement (2004, 2005).
  - *Relevance Judgement*: For each topic, documents are judged as definitely relevant (DR), possibly relevant (PR), or not relevant (NR)
  - *Metrics*: The primary evaluation measure for the task was mean average precision (MAP). As well as the binary preference (B-Pref), precision at the point of the number of relevant documents retrieved (R-Prec), and precision at varying numbers of documents retrieved (e.g., 5, 10, 30, etc. documents up to 1,000)

  TREC genomics aiming at studying very specific use cases, many team approaches fell back on semantics. While queries in 2003 focused on a single gene, they were in 2004 and 2005 concerning relations between biological objects such as genes, diseases, biological processes, etc. (description of the Generic Topic Types can be found in [64]).

  The *Passage retrieval task* (2006) intended to go beyond ad-hoc retrieval by challenging systems to retrieve short passages that specifically addressed an information need. The *Entity-based Question-Answering* (2007) was a continuation of the Passage Retrieval task where the questions were more precise and also required relevant passages as an answer.
  - *Documents*: a collection of biomedical articles from Highwire Press (full text in HTML format, which preserved formatting, structure, table and figure legends, etc.). It represents 162,259 documents.
  - *Topics*: biologically relevant questions.
  - *Relevance Judgement*: judges were instructed to break down the question into required elements and isolate the minimum contiguous substring that answered the question (definitely relevant, possibly relevant, not relevant)
  - *Metrics*: Multidimensional evaluation: passage retrieval, aspect retrieval, and document retrieval, MAP was used to measure all dimensions

  The passage retrieval queries were similar to the ad-hoc task and following the track generic topic types.

- *TREC Medical Records Track*[28] was organized in 2011 and 2012 [168]. The task consists of searching in a set of EHR to identify patient cohorts for (possible) clinical studies.
  - *Documents*: The document set used in the track is a set of de-identified clinical reports (Radiology Reports, History and Physicals, Consultation Reports, etc.). They are semi-structured reports with ICD coding, the chief complaint made available to TREC participants through the University of Pittsburgh NLP Repository[29]. EHR are grouped as "visits", the dataset contains 93,551 reports mapped into 17,264 visits.
  - *Topics*: description of the criteria for inclusion in a study. Topics were created by physicians (graduate students in OHSU)from a list of research areas the U.S. Institute of Medicine (IOM) has deemed priorities for clinical comparative effectiveness research. The dataset contains 35 topics in 2011, 50 in 2012.

[27] https://trec.nist.gov/data/genomics.html

[28] https://trec.nist.gov/data/medical.html

[29] Use of this dataset is now restricted to people having a license with Pittsburg NLP.

  - *Relevance Judgement*: conducted by 25 physicians(1–9 topics each). They rated each visit to determine whether such a patient would be a candidate for a clinical study on the topic (3-points scale)
  - *Metrics*: infNDCG, infAP, P@10

- *TREC Clinical Decision Support*[30] */ Precision Medicine Track*[31] started in 2014, as the Clinical Decision Supports (CDS) track, and since 2017 until now (2020) is named Precision Medicine (PM) track [130–133, 145]. The purpose of the task is to retrieve biomedical articles relevant for answering generic clinical questions about medical records: given a case report, participants have to find full-text biomedical articles that answer questions related to several types of clinical information needs.
  - *Documents*: Full biomedical articles: open access subset1 of PubMed Central (PMC), snapshot of 733,138 articles.
  - *Topics*: case reports and one of three generic clinical question types, such as "What is the patient's diagnosis?". Created by expert topic developers at the NLM. A case report typically describes a challenging medical case, and is often organized as a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to the case.
  - *Relevance Judgement*: manual assessment done by physicians and graduate students on a 3-points scale. 34,949 documents were judged across the topics, with a mean of 1265 documents judged per topic.
  - *Metrics*: infNDCG and P@10

  The use case and topics for this task varied: TREC CDS aimed to retrieve biomedical articles relevant for topics falling into a diagnosis, test, and treatment categories. Since 2017, TREC PM focuses on the oncology domain and provides participants with more complex patient cases. The purpose is to retrieve biomedical articles and clinical trials corresponding to the case.

- *TREC COVID*[32] was organized in 2020 and aimed at helping medical professionals needing to constantly search for reliable information on the virus and its impact. This presented a unique opportunity to study methods for quickly standing up information systems for similar pandemic [129].

  - *Documents*: CORD-19, COVID-19 Open Research Dataset containing over 280,000 scholarly articles[33]
  - *Topics*: manually created topics by organizers with biomedical training, composed of a query, a question and a narrative
  - *Results submission*: 5 rounds of submissions allowing a cumulative set of relevance judgements
  - *Metrics*: trec_eval measures
- *ImageCLEF*[34] is one of the oldest CLEF tasks and has been running for 15 years [108]. ImageCLEF has mostly been organizing tasks with images, but a few tasks have also targeted textual data. The medical retrieval ad-hoc task (2003-2013) aims at retrieving images similar to the image query, and the collection contains textual cases along with related images.
  - *Documents*: images with multilingual textual notes
  - *Topics*: images
  - *Results submission*: Participants could submit results using only the images, only the text, or both
  - *Metrics*: MAP, BPref, P@5, 10, 30

  The purpose of the medical retrieval ad-hoc task is to search for images corresponding to a visual topic. The images from the document collection come with text, which allows basing the retrieval on visual, textual, or mixed approaches.

  The *case-based retrieval task* (2009-2013) aims at retrieving biomedical text from a given image. The purpose is to meet clinicians' information need when facing a specific patient case. The task is built as follows [71]:

[30] https://trec.nist.gov/data/clinical.html

[31] https://trec.nist.gov/data/precmed.html

[32] https://ir.nist.gov/covidSubmit/

[33] https://www.semanticscholar.org/cord19

[34] https://www.imageclef.org/

  - *Documents*: Full text biomedical articles
  - *Topics*: medical cases images
  - *Metrics*: MAP, BPref, P@5, 10, 30

- *CLEF eHealth*[35] is running since 2013 and proposed a variety of tasks focusing on health-related information extraction and retrieval [52, 54, 76, 77].

  The *patient-centered IR task/consumer health search task (CHS)* (2013-2019) targets the retrieval of relevant documents for consumer health search:
  - *Documents*: Set of medical articles and certified documents (2013-2015), large web crawl (2016-2019)
  - *Topics*: manually built patients queries built from real (or realistic) scenario.
  - *Relevance Judgement*: documents are manually assessed by experts on their topical relevance, on their readability since 2015, and also on their reliability since 2016.
  - *Metrics*: P@10, MAP, NDCG, rank-biased precision, including understandability and reliability

  The *Technologically Assisted Reviews in Empirical Medicine Task (2017-2018)* was running for two years and aimed to develop methods to retrieve relevant studies with high precision and high recall [74].
  - *Topics*: 20+30 topics for Diagnostic Test Accuracy (DTA) systematic reviews (training + test sets).
  - *Documents*: MEDLINE database documents
  - *Relevance Judgement*: Assessment was made based on systematic reviews provided by Cochrane Library
  - *Metrics*: AP, Recall @ k, Work saved oversampling at recall R@k, reliability, recall @ @threshold[36]

- *BioASQ Evaluation Challenge*[37] has been held in various venues including CLEF[38]. BioASQ is a challenge tackling large-scale biomedical semantic indexing and question answering [164]. BioASQ proposes 3 tasks:
  - Large-Scale Online Biomedical Semantic Indexing (running every year)
  - Biomedical semantic QA (running every year)

  The Semantic indexing aims at assigning MeSH concepts to biomedical abstracts (to be included in PubMed), which can be considered as multi-label classification.
  - *Data*: annotated PubMed abstract (training dataset), non-annotated PubMed abstracts (test dataset) in English and Spanish (2019)
  - *Metrics*: F-measure and variations

  The Biomedical QA task's purpose is to find the relevant answer to a question. It is done in 2 steps: find relevant articles and snippets (IR), and generate the answer (QA and summarization).
  - *Topics*: medical professionals questions, e.g. *Which 2 medications are included in the Qsymia pill?*
  - *Documents*: Subset of MEDLINE documents
  - *Metrics*: MAP, F-measure, accuracy, ROUGE

- *Other IR test collections.*

  Data sciences challenge platforms such as Kaggle and Glue Benchmark propose a range of medical tasks. However, to the best of our knowledge, most of them are centered on learning tasks (such as annotation and classification) rather than retrieval. The COVID-19 Open Research Dataset Challenge (CORD-19)[39] is an exception, with 17 tasks aiming at developing methods to provide specific types of information from a corpus of of over 200,000 scholarly articles [172]. Example of tasks include: *What do we know about COVID-19 risk factors?, Create summary tables that address therapeutics, interventions, and clinical studies, What has been published about medical care?*

  The NFCorpus [21] is a medical test collection on nutrition facts. It contains 3,244 natural language queries (collected from the NutritionFacts.org website) and 9,964 medical documents mostly from

[35] https://clefehealth.imag.fr/

[36] in 2017, organizers considered 13 evaluation measures

[37] http://bioasq.org/

[38] CLEF in 2013-2015 and 2020-2021, ECML PKDD in 2019, EMNLP in 2018, BioNLP in 2016-2017

[39] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/

Name	Description	Statistics
MIMIC III[1] [70]	Openly available dataset developed by the MIT Lab for Computational Physiology. It contains deidentified health records associated intensive care unit admissions.	58,976 admissions from 48,520 patients
STRIDE [94]	Unstructured clinical notes from 1.2 million patients, collected over a 19-year span. The notes comprise a combination of pathology, radiology, and transcription reports.	20 million notes from 1.2 million patients
Healthmap[2]	Corpus of public health-related news articles in English extracted from HealthMap, an online aggregator of news articles from all over the world for disease outbreak monitoring and real-time surveillance of emerging public health threats.	124,850 documents
DE-SynPUF[3]	Realistic set of claims data in the public health domain. Subset of the CMS limited datasets. It contains five types of data – Beneficiary Summary, Inpatient Claims, Outpatient Claims, Carrier Claims, and Prescription Drug Events.	229 million records from nearly 2 million patients
PMC open access subset[4]	Free full-text archive of biomedical and life sciences journal literature from PubMed at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). It contains more than 5 million full-text records.	More than 5 million full-text records
PubMed - MedLine[5]	Baseline set of MEDLINE/PubMed citation records in XML format for download on an annual basis by NLM.	Annual release with varying size
MedNorm[6][16]	Corpus of annotated textual descriptions extracted from biomedical and social media domains corpora	27,979 documents

Table 2. Non exhaustive list of medical and clinical corpora openly available and used in the papers cited in the survey

PubMed. The relevance judgement are automatically extracted from the NutritionFacts website, based on the hyperlinks (a direct link is a high relevance, a intermediary link is a moderate relevance, a topic/tag system connection is a low relevance).

2.2.3 Other Resources: Text Corpora and Word Embeddings. In addition to test collections, medical IR systems can rely on other sources, such as textual corpora for query expansion, or as training material; or word/concept embeddings, used as a knowledge resource or a representation model. Table 2 and Table 3 describe the major resources found in the literature along with their references.

3 STRUCTURED KNOWLEDGE RESOURCE DRIVEN APPROACHES FOR SEMANTIC SEARCH IN MEDICAL IR

This category of works primarily relies on predefined lists of senses inventoried in external knowledge resources (See Section 2.1) to tackle the semantic gap problem in medical IR. Knowledge resources include dictionaries, thesaurus, ontologies, etc. which could be combined with corpora (i.e., collections) of texts.

[1] https://mimic.physionet.org/
[2] http://www.healthmap.org
[3] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF
[4] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[5] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[6] https://github.com/mbelousov/MedNorm-corpus
[1] http://bio.nlplab.org/
[2] https://github.com/ncbi-nlp/BioWordVec
[3] https://github.com/dartrevan/ChemTextMining/blob/master/README.md
[4] http://diego.asu.edu/Publications/ADRMine.html
[5] http://www.ke.tu-darmstadt.de/resources/medsim
[6] https://data.mendeley.com/datasets/b9x7xxb9sz/1
[7] https://github.com/dmis-lab/biobert
[8] https://github.com/EmilyAlsentzer/clinicalBERT

Name	Description
Word2vecPMC[1] [124]	Word2vec embeddings built from PMC full text articles
Word2vecPubMed+PMC[1] [124]	Word2vec embeddings built from PMC full text articles and PubMed abstracts
Word2vecPubMed[1] [124]	Word2vec embeddings built from PubMed abstracts
Word2vecPubMed+PMC+wiki[1] [124]	Word2vec embeddings built from PMC full text articles, PubMed abstracts and Wikipedia
Word2vecOSHUMED [35]	Word2vec built from OHSUMED document collection
BioWordVecIntrinsic[2] [190]	FastText built from PubMed and MeSH
BioWordVecExtrinsic[2] [190]	FastText built from PubMed and MeSH
BioWordVec[2] [190]	FastText built from PubMed and MIMIC III
HealthVec[3] [142]	Word2vec built with health reviews
DrugTweetsVec[4] [114]	Word2vec built with drug related tweets
PubMedVec[5] [101]	AiTextML built with PubMed abstracts
Drug2Vec[5] [101]	Word2vec built with PubMed and DrugBank
MedNorm[6] [16]	DeepWalk and Word2vec SkipGram with MedNorm corpus
BioBERT[7] [85]	BERT pre-trained with biomedical articles (PubMed and PMC)
ClinicalBERT[8] [1]	BERT pre-trained with clinical data (MIMIC III)

Table 3. Major off-the-shelf embeddings in the medical domain. The list is an extended version of the list provided in [73]. If they are available in a public repository, resource's URLs are indicated.

The key idea of the works belonging to this category is to enhance, using external knowledge, either the query/document-representation or their matching as described in the following sub-sections.

3.1 Query Expansion

(Automatic) Query expansion (QE or AQE) refers to the act of automatically revising the query by adding new terms. QE has received a great deal of attention for several years in the literature and has been acknowledged as the most successful technique to deal with the vocabulary mismatch issue [25]. Medical search leverages today on the firmer theoretical foundations and a better understanding of the usefulness and limitations of a variety of QE approaches known in IR [13, 17, 18]. We review in what follows state-of-the art works and then discuss their effectiveness results obtained in similar benchmarks.

3.1.1 Overview of QE techniques. We further categorize the semantic QE used in medical IR based on two key impacting dimensions that have arisen from previous work [39, 203]: (1) the context of knowledge used to expand the query that might be global to the medical domain as held by the knowledge resource, local to the search at hand or hybrid by combining both of them. The most common local context used so far refers to the top-ranked query results retrieved in response to the initial user query and associated relevance feedback signals; (2) the number and the nature of the knowledge resource used in terms of domain specialization. Table 4 provides a detailed classification of representative works along these two key dimensions. The methods are ordered chronologically.

A critical stage of QE is the generation and ranking of potential candidate expansion terms to select a subset to expand the user's query. In practice, a score is computed for each candidate expansion term and the m terms with the higher scores are selected as the expansion terms. Term generation and ranking stage are generally run in two steps using respectively the local search context and the global knowledge context:

(1) *Local context:* first, an initial retrieval run is performed to build a list of ranked documents in response to the user's query. Most of the proposed QE methods are based on traditional document ranking models such as the language model [38, 115, 153, 181, 198] and the probabilistic model [151, 201]. Then, terms contained in the top-ranked documents are extracted using a blind or pseudo-relevance feedback approach. Additional document/query processing allows the mapping between these terms and the concept-based entries of one or multiple knowledge resources.

(2) *Global context:* candidate expansion terms are identified based on their explicit vs. implicit one-to-one or one-to-many semantic relationships with the query terms or the terms issued from the local context (if

Reference	Data Context	Knowledge Resources	
		Domain	*Number*
Dinh and Tamine [38]	Local & Global	Specialised (MEsH, SNOMED, GO, ICD)	Multiple
Zhu and Carterette [198]	Local & Global	Specialised (MeSH, medical collections)	Multiple
Limsopatham et al. [89]	Local & Global	Specialised (MeSH, MedDRA[40], DOID[41])	Multiple
Oh and Jung [115]	Local & Global	Specialised (Medical collections) & General (Medical collections)	Multiple
Shondi et al. [153]	Local (corpus and top ranked documents)	Specialised (MeSH)	Single
Martinez et al. [100]	Global (knowledge resource)	Specialised (UMLS)	Single
Wang and Akella [170]	Local & Global	Specialised (UMLS)	Single
Znaidi et al. [201]	Local & Global	Specialised (UMLS)	Single
Soldaini et al. (a) [152]	Local & Global	General (Wikipedia)	Single
Soldaini et al. (b) [151]	Local	General (Wikipedia)	Single
Xu et al. [181]	Local & Global	Specialised (MeSH)	Single
Balaneshinkordan and Kotov [10]	Local & Global	Specialised (UMLS, DGIdb[42], COSMIC catalog)[43])	Multiple
Fujita [46]	Local & Global	Specialised (LocusLink and MeSH)	Multiple
Ando et al [2]	Global	Specialised (LocusLink, GO, MeSH and SwissProt)	Multiple
Huang et al [68]	Global	Specialised (AcroMed and LocusLink)	Multiple
Demner-Fushman et al[37]	Global	Specialised (UMLS)	Single
Shen et al [143]	Global	Specialised (UMLS)	Single
Zhou et al [196]	Local & Global	Specialised (MeSH, Entrez Gene/LocusLink)	Multiple
Zhu et al [199]	Global	Specialised (UMLS)	Single

Table 4. A fine classification of representative AQE methods in medical IR.

> any). The semantic relationships are established within a global context provided by a single or multiple knowledge resources that could be either general (e.g., Wikipedia) or specialized (e.g., UMLS, MeSH). In practice, term ranking is based on a relevance score that estimates the strength of the semantic relationships of the candidate expansion terms with one or multiple query terms or terms from the local context.

Several key issues heavily impact the effectiveness of structured knowledge resources driven approaches to semantic search on medical texts [39, 203]. Among the most important ones: (1) concept mapping which consists of performing exact associations between mentions and knowledge resource entries. Regardless of the domain application, semantic search based on knowledge resources is highly dependent on the performance of concept mapping (or entity recognition) which is still a challenging problem; (2) knowledge resource characteristics: structure, specialization level, number of resources used and possible combinations. The concept mapping performance is out of the scope of this paper. We consider the knowledge resource characteristics and the number of resources to be the first level of our categorization of QE techniques. The motivation behind our choice is that using single vs. multiple resources leads to fundamental differences in the key stage of term generation and ranking in QE.

Mono-Resource Based QE.. The simplest form of term generation and ranking is based on the use of evidence provided by a single knowledge resource. In Table 5 we show the main term-ranking functions associated with the mono-resource based QE methods presented in Table 4.

Reference	Term Ranking/Selection Function
Shondi et al. [153]	Enhanced term query frequency if term t is associated to relevant types (e.g., disease)
Martinez et al. [100]	PageRank term score over the query graph Relationships between terms are provided by the resource
Wang and Akella [170]	Maximum likelihood estimation of term assignment to relevant types
Znaidi et al. [201]	Term score based on a recursive propagation algorithm through the query graph
Soldaini et al. (a)[152]	Likelihood of being health-related
Soldaini et al. (b) [151]	Likelihood of being health-related
Xu et al. [181]	Learning-to-rank -based score

Table 5. Main term ranking and selection functions used in AQE methods using a single resource.

To identify candidate expansion concepts, most previous work in the area of medical IR used either a specialized ontological resource (e.g., UMLS [170, 181] or MeSH [153, 201]) or a raw specialised textual corpus (e.g., health-related pages of Wikipedia [152]). Basic ontological-based approaches commonly identify candidate concepts: related concepts having *generalization/specialization* or *part of* relationships with the query concepts; terms generated from the local context [152, 201]. For example, with respect to the UMLS resource, the query *'cancer'* is likely to be expanded with the synonymous concept '*malign tumour*' and the query *'osteoporosis*' could be expanded with the (preferred) term of concept *'Boniva'* since UMLS states that *'Bonivia'* is the treatment of *'osteoporosis*. In [181], a LTR based term ranking method is adopted. The main characteristic of this method is the joint use of MeSH-based features and corpus-based features to identify semantically related candidate terms. A different class of methods for term selection and ranking was suggested in [100, 201] by considering the whole structure of the candidate term graph built from the relationships provided by the knowledge resource. In [100], a graph-based representation is proposed to structure the query concepts with the relations provided in the UMLS Meta-thesaurus. The concepts are then ranked using random walks over the graph, mainly using a traditional PageRank score. Using a similar approach, [201] generate the query sub-graphs for a PICO query structure with the MeSH concept-to-concept relationships. The authors compute a candidate term expansion score by performing a recursive propagation algorithm. The latter propagates the scores of the active query concepts to their sub-concepts considering each query sub-graph by iteratively summing the scores of the hyponym concepts.
Another term ranking method is unlikely based on the use of raw textual data as background knowledge to compute the candidate term selection scores [151, 152]. In [152], Soldaini et al. introduced the HTPRF (Health Terms Pseudo Relevance Feedback) score computed for each term appearing in the N top-ranked documents in response to the original user's query. The HTPRF score is estimated using the maximum likelihood of being health-related based on the odds ratio between term frequencies in health-related Wikipedia pages over the whole Wikipedia corpus.

Multi-Resources Based QE. Another common practice for term generation and ranking is to leverage multiple knowledge resources by using a variety of semantic relationships between terms. We mainly distinguish approaches that use semantic term relations provided by homogeneous data sources [10, 38], namely ontological knowledge, from approaches combining them with term relations inferred from local corpus statistics [89, 115, 198]. When QE is based on multiple resources, a key question is how to deal with term relations that overlap between resources. The approach generally adopted directly or indirectly assigns a relevance weight to each resource and then injects this weight into the computation of the candidate term score. It is expected that resources with little relevant informative content for the current query evaluation, in terms of semantic association, should have a lower impact on term score importance. Following this general paradigm, various term-ranking functions have been proposed that assign scores that best contextualize the candidate term relevance to both the query and the global knowledge provided by each resource.
In Table 7 we show the main term-ranking functions associated with the multi-resource based QE methods presented in Table 5.

Test Collection	Reference	Metric	Value
TREC Genomics 2004	Ando et al [2] (2005)	MAP	0.4552
	Dinh and Tamine [38] (2011)	MAP	0.4529
	Fujita [46] (2004)	MAP	0.4075
TREC Genomics 2005	Huang et al [68] (2005)	MAP	0.3011
	Ando et al [2] (2005)	MAP	0.2883
	Dinh et al [39] (2013)	MAP	0.2859
	Dinh and Tamine [38] (2011)	MAP	0.2685
TREC Genomics 2006	Zhou et al [196] (2006)	MAP	0.174 (passage) 0.537 (document)
	Demner-Fushman et al [37] (2006)	MAP	0.047 (passage) 0.379 (document)
	Xu et al [181] (2019)	MAP	0.0706 (passage) 0.2818 (document)
TREC Medical Records 2011	Zhu and Carterette [198] (2012)	BPREF	0.583
	King et al [78] (2011)	BPREF	0.5523
	Martinez et al [100] (2014)	BPREF	0.5469
	Limsopatham et al [91] (2013)	BPREF	0.5283
CLEF eHealth CHS 2014	Oh and Jung [115] (2015)	MAP	0.8478
	Thakkar et al [162] (2014)	MAP	0.4146
	Shen et al [143] (2014)	MAP	0.4069
OHSUMED	Oh and Jung [115] (2015)	MAP	0.1934

Table 6. Overview of the performances of the QE approaches presented in the paper, when they are comparable: tested on the same test collection with the same metric. The challenges' participant best run are underlined.

Given a set of terminologies (MeSH, SNOMED, ICD-10 and GO), data fusion techniques (e.g., CombMax, CombMNZ) are used in [38] to select among the concepts extracted from the N top-ranked documents in response to the query the best candidates for QE. To do so, resource importance is implicitly considered in the computation of concept scores. More precisely, for each feedback document among the N candidates, the authors combine the average weight of associated concepts issued from each terminology and the document rank in the ordered list of documents. Interesting extensions of the multi-resource based QE technique that use both ontological and corpus-based knowledge are described in [10, 89, 198]. The general underlying idea is to leverage from both explicit human-established term associations provided in ontological resources and hidden term associations inferred from textual corpora to select expansion terms. A basic method proposed in [89], consists of building statistical-based association rules using term co-occurrences in the corpus and lexical-based association rules based on the resource knowledge graph. However, the issue of term association overlap between resources is not addressed. Instead, all the resources are given equal importance weights. The same hypothesis related to the benefit of combining ontological and corpus-based knowledge is adopted in [10]. However, instead of using an additive function to compute the overall candidate expansion concept score as done in [38], the authors compute a joint feature-based posterior probability of relatedness of concept to the different resources. Instead of using static resource weights, Zhu et al. [198] propose a query-adaptive resource weighting strategy that relies on the hypothesis that a good expansion resource for a query allows building an expanded query topically close to the user's original query. Thus, a collection weight is simply the normalized similarity between the initial query and expanded query language models using the Jensen-Shanon divergence metric.

3.1.2 Discussion. Despite the relatively limited number of test collections in the domain, one cannot compare methods across papers, as they might not be evaluated on the same test collection or metrics. However, we provide in Table 6 a summarized view of the methods described in this section, when comparable. We only present in this table methods centered on AQE, without specific semantic ranking (presented in Section 3.2.2). Firstly, we can observe from Table 6 that the range of performance levels obtained across datasets varies greatly. This shows the diversity of the evaluated tasks and the differing effects of the QE techniques used on retrieval

Reference	Term Ranking/Selection Function
Dinh and Tamine [38]	CombRank: sum of terminological concept ranks
Zhu and Carterette [198]	Mixture of term relevance model
Limsopatham et al. [89]	Conceptual-based association within terminologies
Oh and Jung [115]	Feedback model
Balaneshinkordan and Kotov [10]	Conceptual-based association within terminologies

Table 7. Main resource selection functions used in AQE methods using multiple resources.

effectiveness. The results obtained on the TREC Genomics 2004 test collection show that a careful selection of the expansion candidates, as done by [38] gives better results. Ando et al [2] took part in the challenge in 2005 and presented results on 2004 and 2005 datasets. Their best run approach combined structural feedback with AQE, expanding terms with synonyms from LocusLink, GO, MeSH and SwissProt. Fujita [46] obtained the top ranked results (with MAP) in the evaluation campaign set in 2004 by combining blind relevance feedback with term and acronym expansion with synonyms from LocusLink and MeSH. The results obtained on the TREC Genomics 2005 test collection indicate that expanding queries with term variants obtained with lexical rules and knowledge bases (AcroMed and LocusLink) gives better results [68] than using term variants provided from multiple resources [38]. TREC Genomics 2006 provides a dataset for passage retrieval and a dataset for document retrieval. The method proposed by [37], based on UMLS synonyms expansion and ranking fusion, seems to be more efficient for documents than for passages. On the contrary, the method proposed by [181], based on Learning-to-rank applied to candidate expansion terms, appears to be more efficient on passages than on documents. The best submission (with MAP) to the Genomics 2006 challenge was Zhou et al [196] who combined AQE with semantic matching and ranking. This approach is described in Section 3.2.2.

All of the QE techniques evaluated in the TREC Medical Records track prove that the techniques help in improving cohort search. Regarding the these techniques, it seems that combining multiple knowledge resources [198], graph-based term selection [100] as well as concept-co occurrence based selection [91] are equally effective. The best run of the challenge was submitted by King et al [78], who expanded the queries with UMLS related terms and most similar terms from an encyclopedia, found with a PRF approach. They also filtered documents using the age, gender, race and admission status attributes of the records.

Oh and Jung [115] conducted their experiments on several datasets including CLEF eHealth and show that their clustered-based external expansion model is efficient on medical IR on various search tasks (TREC CDS, CLEF eHealth). However, their approach seems to be challenged by the OHSUMED test collection, which might be caused by the challenging search task and the relatively short queries. The results they obtain for MAP are surprising, since they obtained 0.3989 MAP in the challenge in 2014 [53] with a similar query-likelihood with Dirichlet smoothing approach. Their paper did not allow to understand this increase. Shen et al [143] propose a query expansion approach on CLEF eHealth CHS 2016 topics, expanding queries with synonyms listed in UMLS. This approach seems to be efficient at tackling circumlocution in the topics by expanding them with equivalent terms that better match the documents. The best run in CLEF eHealth (with MAP) was obtained by Thakkar et al [162] with the query likelihood model without any semantics.

3.2 Relevance Ranking

In its most intuitive form, the relevance ranking problem consists of estimating the document relevance based on some knowledge mined from the initial query and corresponding search results on the one hand and knowledge mined from knowledge resources on the other hand. In this section, we review existing approaches to relevance ranking models and techniques that involve external knowledge resources and then discuss their effectiveness results when obtained on the same benchmarks.

3.2.1 Overview of models and techniques. Based on how the document relevance estimation model is designed, we classify relevance ranking methods into two main categories in this section.

Inference-based relevance ranking. In inference-based relevance ranking [24, 56, 81, 183], document relevance is formally estimated using a mapping score between knowledge held by the query and knowledge

held by candidate documents. A typical inference-based relevance ranking framework includes either offline (document) vs. online (query and search results) topic processing and then semantic inference-based ranking that bridges the gap between queries and documents using knowledge resources. The key problems are how to map document knowledge (through concepts) with the noisy initial query and how to transform this mapping into strength scores to be incorporated into the relevance estimation function.
Following this general approach, different techniques have been proposed with different conditions of semantic inference. Koopman et al. [81] designed a model of relevance ranking based on two mechanisms of inference that are prevalent in the medical domain: (1) conceptual implication between concepts driven by knowledge resources (for instance general implications between instances of *organisms* and *diseases*, e.g. *Varicella zoster* ⟶ *Chicken pox*); (2) conceptual associations between concepts that reveal some natural dependencies (e.g., *anxiety* and *depression*) that are not necessarily provided by knowledge resources but derived using corpus statistics. Both types of inferences are applied to query concepts and document concepts to estimate relevance. More specifically, documents are first pre-processed as concept-based graphs using a knowledge resourceto highlight both conceptual and association relationships between concepts. Then, the retrieval model is designed as an inference process over the graph whose ultimate objective is to compute document relevance for a query as the cumulative strength of implications between their concepts similar to a probabilistic language modeling. Instead of using implications and associations between concepts of queries and documents, Yan et al. [183] argue that the analysis of topical granularities of both queries and documents provide signals of relevance. Granularity is estimated with the combination of generality and cohesion based on the depth of associated concepts and the strength of the semantic relationships between them in the knowledge resource (e.g., the Resnik [127] concept-to-concept similarity). Using a document re-ranking function, the relevance ranking model moderates the traditional query-document relevance score with the granularity gap between them. To reveal the hidden relationships between the query and document concepts, [24] propose a probabilistic inference model that operates on a probabilistic knowledge graph (called clinical picture and therapy graph CPTG) built upon corpus-based knowledge resources including EHRs and scientific articles (e.g., PubMed articles). The underlying motivation is to compute a probability distribution of concepts over all the concept dependencies (and implications) that can be observed in all the possible clinical pictures and therapies of patients. Hence, the semantic scope of concept dependencies in this work is higher than it is in [81, 183].

Feature-based relevance learning. This category of works relies on a learning approach of relevance [9, 91, 149, 180]. Formally, the relevance estimation is turned into a LTR problem which can be formulated as finding the optimal ranked list of documents using manually crafted features from query and documents and combining them with a ranking objective. The key problem is to identify the best set of intrinsic query or document features, as well as mutual features of query-document matching that is predictive of relevance. A basic approach is proposed by Soldaini and Goharian [149] who leverage a large pool of features including traditional statistic-based features (e.g., term frequency *tf*, inverse document frequency *idf*) and semantic features (e.g., concept frequency, semantic types, and word embeddings). Each document and query is first represented by a vector of features. They tested a set of loss functions such as LamdaMART, AdaRank, and ListNet to perform a supervised document ranking. Through a feature analysis, the authors acknowledged that statistical-based features have the most impact on retrieval performance. An interesting revised LTR model, called Latent-ListMLE, is proposed by Xiong and Callan [180]. It consists of extending the traditional ListMLE loss function by adding a latent layer in the ranking generation process to jointly learn the best representation of queries and document rankings. Query representation relies on query-object relationships inferred from external resources (e.g., MeSH). Traditional features are used to represent such relationships including textual similarity features and concept frequency mean similarity score between concepts. Feature-based learning is employed by Balaneshin-Kordan and Kotov [9] to learn the feature weights of concepts used in the computation of document relevance score. Traditional statistical and semantic features are used similar to those used in [149].

3.2.2 Discussion. We provide in Table 8 a summarized view of the methods described in this section, when comparable. We only present in this table methods centered on semantic-based ranking, and methods adopting both semantic-based AQE and ranking.

Test collection	Reference	Metric	Value
TREC Genomics 2006	Wang and Akella [171]	MAP	0.473 (document)
	Zhou et al [196]	MAP	0.174 (passage) 0.537 (document)
TREC CDS 2015 task A	Balaneshin et al [9]	infNDCG	0.3109
	Goodwin and Harabagiu [56]	infNDCG	0.311
TREC CDS 2015 task B	Song et al [155]	infNDCG	0.3821
	Balaneshin et al [9]	infNDCG	0.3690
	Goodwin and Harabagiu [56]	infNDCG	0.382
CLEF eHealth CHS 2013	Wang and Akella [171]	P@10	0.572
		NDCG@10	0.587
	Zhu et al [199]	P@10	0.5180
		NDCG@10	0.4665

Table 8. Overview of the best results obtained by each team for ranking methods and mixed methods (AQE + ranking) on the same test collection with the same metrics. Best runs submitted to the tasks are underlined.

Only two models were evaluated using similar datasets and metrics for ranking-only approaches [9, 56]. They both used the TREC CDS test collection, which proposed 2 subtasks. The performance of teams approaches are very close on both tasks, which shows that learning-based matching approaches [9] and inference-based approaches [56] can be efficient on a clinical decision use case.

As for methods using both AQE and ranking, if we first compare results with those described in Table 6, we observe that combined methods give higher results on TREC Genomics 2006. Wang and Akella [171] extract UMLS concepts from the queries and documents with Metamap, and propose a concept-based relevance model. Zhou et al [196] mixed a 2-dimensional ranking (word-based and concept-based) and introduce a semantic matching model. They combine this model with an AQE model using synonyms from EntrezGene/LocusLink and MeSH and some pseudo relevance feedback, by selecting the most similar concepts from the top 15 ranked documents. Wang and Akella's method proves its efficiency on a consumer health search scenario [171]. Team Mayoclinic [199] obtained the best results in CLEF eHealth CHS 2013 with a double index (word-based and concept-based) and some concept-based relevance feedback using UMLS, although the improvement was not statistically significant in comparison with the organizers BM25 baseline [50].

The best run (with MAP) in TREC Genomics 2006 was obtained by Zhou et al [196]. The best run (with infNDCG) in TREC CDS 2015 task A was obtained by [9]. While task B best run was submitted by Song et al [155], by expanding queries with MeSH terms extracted from the title and snippets of top ranked documents obtained by querying Google with the topic. Zhu et al [199] obtained the best results in the CLEF eHealth 2013 CHS challenge.

4 DATA DRIVEN APPROACHES FOR SEMANTIC SEARCH IN MEDICAL IR: FOCUS ON NEURAL APPROACHES

Generally speaking, data-driven approaches for semantic representation and matching in the medical domain, refer to a category of machine-learning (ML) based methods that learn semantics from raw textual medical corpora. Early methods, reviewed in the previous section [9, 91, 149, 180], are generally based on feature-based LTR methods combined with the use of symbolic knowledge provided by knowledge resources. So far, few works have adopted the ML approach in medical IR because of the difficulty in achieving a reasonable balance between the cost of human annotation and the gain through retrieval effectiveness.

Recently, the historical success achieved by deep learning approaches in a wide range of research disciplines such as computer vision and speech recognition has given rise to a surge of interest in IR [105, 116] and medical IR [36, 93, 150, 161]. Specifically, semantic search on medical texts leveraged on neural IR and NLP models and techniques. Following the general approach in neural IR, neural models for medical search learn semantics using unsupervised or semi-supervised matching models. Unsupervised models use pre-trained item embeddings (e.g., word embeddings) and then inject them in a retrieval model [32, 110]. Semi-supervised models (1) learn item embeddings from scratch such as representations of words and concepts [111] or more complex items such as patients [156] and employ them in a retrieval task [93, 111], and (2) learn either representations or relevance/semantic ranking in an end-to-end neural fashion by considering a target task (e.g., learning patient similarities [200]).

This section serves as an attempt to provide an overview of early research in both representation learning and neural semantic matching in medical search. Since our goal is to provide key insights into specific medical search problems the neural approaches can tackle, we scope our discussion to the representation learning models of concepts, documents, and patients having the following features: (1) models that have already been evaluated in ranking and similarity-based search tasks w.r.t. the article's scope (See Section 1.3). Examples of those tasks include document retrieval (e.g., scientific literature retrieval, care episode retrieval) and matching (e.g., patient similarity and concept relatedness); (2) models that could be deployed in ranking and similarity-based search tasks but to-date have been evaluated in other downstream tasks (e.g., diagnosis prediction and mortality prediction). Models that do not fulfill both requirements are not covered in this survey since we aim to provide a thorough examination of the fundamentals in neural medical search instead of providing a general picture of neural approaches to all medical text mining tasks. A reader interested in deep learning approaches used to mine knowledge from medical data is referred to dedicated surveys [72, 179].

In the following, we categorize the research works in the area with respect to the type of medical item learned that can give rise to conceptually different retrieval tasks that are developed in context. For each category of work, we present an overview of state-of-the art models and then discuss the trending results.

4.1 Representation Learning of Concepts

4.1.1 Overview of models. Medical concepts are traditionally encoded as discrete symbols based on their ontological identifiers (e.g., Concept Unique Identifiers CUI). Their similarities are generally measured using path-based measures and/or content-based measures [121]. In contrast with such symbolic representations, a concept embedding encodes a concept as a low-dimensional continuous dense vector such that it is ideally aligned with its ontological counterpart. Medical concept embeddings are typically built using large unlabelled medical corpora either from scientific literature [150], health-related data [28, 49] or both [35, 93]. Following the advancements in the design of neural language models, the earliest works applied and/or extended the shallow models Skip-gram and Continuous Bag of Words (CBOW) [102], while the more recent ones fine-tuned transformer-based including BERT model and its variants [165]:

- *Skip-gram-based models* [8, 32, 35, 48, 49]: Traditionally the Skip-gram variant builds word representations by optimising their ability to predict the representation of surrounding terms. For concept embeddings, most of the work relies on a prior text pre-processing where words or compound words are first associated with CUI from an external medical resource (e.g., UMLS). Then, training is performed using sequences of concepts to build concept vector representations that are predictive of nearby concepts. This was the general approach adopted by most of the representative works [32, 35], but using different types of medical corpora in the learning process: (1) medical clinical narratives in [35] using TREC Med track data, MCECN dataset [45]; (2) medical scientific literature in [32, 35] using OHSUMED dataset; (3) other medical raw text corpora such as patient claims [32] or health-related news corpora [48]. In general, concept vector representations highly depend on the medical training data used, since the only information for their learning is their distribution over the corpora. Thus, the impact of those representations on downstream tasks is highly variable.
  For the concept relatedness task, Devine et al. [35] found that learning concept embeddings from clinical narratives is less effective than learning them from the scientific literature. Their finding is based on a quantitative analysis of the correlation between the vectorial-based vs. human-based pair-wise similarity ranking of concept pairs. In contrast, through the analysis of concept vector clusters built-up on different learning algorithms and different textual corpora, Choi et al. [32] argue that the learning concept embeddings from clinical narratives lead to a significant qualitative and quantitative improvement whose levels depend on the variant of the learning algorithm used. These findings clearly show that the corpus characteristics have an impact on the quality of the embedding outputs. The need for learning from multiple medical sources therefore became apparent.
  Regarding disease classification, Gosh et al. [48] showed that learning concept representations from medical corpora is more effective than using pre-trained word embeddings. The classifier was effective for a wide range of diseases including endemic and rare diseases. Interestingly, Bai et al. [49] propose the

JointSkip-gram model which embeds both diagnosis medical codes and words from clinical narratives in the same embedding space. Thus, the learned representations benefit from both concept-to-concept, word-to-word, and word-to-concept similarities. These representations are effective for a diagnosis prediction task using a small EHR database. However, no comparison has been carried out over the use of simple Skip-gram models to generate the visit embeddings. This work has been recently extended to leverage multiple knowledge resources [8]. The key problem addressed is the diversity of medical coding ontologies present in claims provided by different providers. To tackle this problem, the authors proposed *typePPMI* that computes the joint frequency distribution of medical codes over multi-source data. TypePPMI is then used to update the Skip-gram objective function in the negative sampling step so as to better capture relationships between codes.

- *Continuous bag-of-words (CBOW)-based models* [22, 93, 185]: Unlike the Skip-gram model, the CBOW models predict the current word given its context. Similar to the Skip-gram-based learning models, the CBOW-based learning models require the pre-processing of medical documents to identify sequences of CUI from resources and then learn the representation of a CUI according to the context CUI frequently appearing around it. The basic approach of this learning model is used in [185]. Based on the idea that if a word is well generated from a given context, its related words should also be well generated from the same context, Liu et al. [93] proposed a variant of the CBOW model by regularising accordingly the objective function. The regularizer consists of a log-likelihood of the co-occurrence of contextual concepts given a target context. The authors evaluated the effectiveness of the concept representations using two search tasks by using the TREC filtering based on the OHSUMED collection and CLEF eHealth 2014-2015 datasets. By building query and document vectors based on the averages of related concept embeddings, the authors evaluated a document retrieval task based on a re-ranking approach. The results showed that the retrofitted vectors allow the achievement of higher performance levels over traditional search models and basic CBOW-based models. However, it is still unclear to what extent the level of improvement is impacted by the type of embedding, the nature of the task (document retrieval) or the way the embedding has been incorporated into the relevance score function (linear combination of exact retrieval score and neural retrieval score). A recent work by Cai et al. [22] incorporated the temporal scopes of medical concepts in the learning process of their embeddings. More specifically, based on the CBOW model, the authors used the attention mechanism to learn the importance of each concept in a time-based scope. The latter is defined as the largest number of time-units between the contexts and the target medical concepts. Comparative experiments have been carried out based on an ICD-concept clustering task using private and public health-related datasets. The results clearly showed that the time-aware attention was able to capture more accurately than the traditional CBOW model, concept-context pairs that lead to a better improvement over a KNN concept clustering task. However, more intensive experiments are required to assess whether the model performance is task-dependent or resource-dependent.

- *Fastext and Glove based models* [15, 191]: recently, new concept embedding models based on Fastext [191] and Glove [15] have been proposed. In [191], Zhang propose *BioWordVec* a new set of word embeddings based on the subword embedding model. By using both unstructured data (biomedical literature) and structured knowledge (MeSH term sequences), the subword embedding model learns the text sequences and MeSH term sequences in a unified n-gram embedding space. In contrast, the central idea in [15] is is similar to the one given in [8]. The authors extended the Glove model by introducing the PMI word-context measure into the negative sampling step, leading to develop the SPPMI matrix (*shifted positive pointwise mutual information matrix*) which is implemented with Glove.

- *Deep neural- based models*
  - *Transformer-based models* [1, 84]: Lee et al. [84] and Alsentzer et al.[1] recently introduced respectively BioBERT and Clinical BERT which are the first fine-tuned BERT models on large Biomedical corpora. Unlike, previous medical language models described above, these language models do not rely on any design extension of the BERT model but rather train the BERT-based embeddings on

medical corpora. For instance, the overall process of BioBert relies on a three-stage training: 1) pre-training on a general domain corpora, mainly Wikipedia and Books corpus; 2) fine-tuning on domain specific data, mainly PubMed abstracts and PMC full-text articles; 3) further fine-tuning on task-oriented datasets, including NER, relation extraction and question-answering. The results of the question-answering task on the BioASQ dataset clearly show the effect of fine-tuning BERT model on domain-specific knowledge.

- *Other types of deep neural models* [29, 150, 154]: other recent works use other neural architectures such as CNN and multi-layer perceptrons (MLP). To overcome the limitation of the low-occurrence of some medical concepts that impact negatively the quality of the learned outputs, Choi et al. [29] proposed the GRAM, a graph-based attention model for concept representation learning. The key underlying idea is to leverage the parent-child relationship provided by a resource to enhance the likelihood of co-occurrence of less-observed concepts in context. The GRAM model is based on a multi-layer perceptron architecture with an attention mechanism that learns in an end-to-end fashion the concept importance in context based on hierarchical relationships provided at the input of the learning process. The authors found that the GRAM's performance on a diagnosis prediction task using the MIMIC III dataset is slightly higher than the baselines but also found that the results do not reach the same level of performance depending on different settings including the downstream task evaluated. One possible explanation is that the embeddings of concepts are close to those of their ancestors, which might not be relevant for all tasks. Song et al. [154] propose the MMORE model that tackles this limitation by allowing the learning of multiple representations of concept ancestors from multiple ontologies that carry multiple semantic meanings. Learning multi-sense based representations of concepts is particularly relevant since co-occurrences statistics in the corpus used for their learning do not necessarily fit with the semantic held by a single resource. Following the GRAM model, the concept embeddings are built as the linear combination of the basic embeddings of the ancestors learned using each ontology where attention weights act as damping factors. Experiments on the MIMIC III dataset showed the relative performance of the MMORE model over the GRAM model using a diagnosis prediction task. However, experiments were performed on only two resources (ICD and CCS) and only using one downstream task (diagnosis prediction). The question of the variability of the model performance in additional downstream tasks with respect to a more general framework based on multiple resources is still open.

4.1.2 Lessons learned. Table 9 provides an overview of the reviewed representation learning models of medical concepts organized on the basis of the core language model used. To allow for a fair comparison, we highlight for each the intrinsic vs. extrinsic task as well as the datasets used for training and evaluation.

As can be seen from Table 9, two major types of tasks have been used for evaluation: (1) concept clustering and relatedness which are both intrinsic evaluation tasks leading to similar conclusions about the vectorial similarity of concept embeddings regarding a human-assessed medical knowledge resource; and (2) search (e.g., literature retrieval, search for cohort, question-answering) and diagnosis prediction as extrinsic tasks. We can also see that the MIMIC III dataset, as can be expected given its large size in comparison to other biomedical datasets (See Section 2.3), is widely used for both types of tasks. While concept clustering tasks make use generally of UMNSRS, CCS and ICD9 ground truths for evaluation, search tasks rather use limited human relevance assessments generally provided in benchmarks such as TREC OHSUMED, TREC CDS and CLEF e-health datasets (See Section 2.3). The differences observed on several impactful factors used for

[29] http://clinicalml.org
[30] https://www.healthmap.org
[31] https://mimic.physionet.org/
[32] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF
[33] https://www.hcup-us.ahrq.gov
[34] https://www.cms.gov

Reference	Neural model extended	Task	Used datasets
		(I: Intrinsic, E: Extrinsic)	(T: Training, E: Evaluation)
Devine et al. [35]	Skip-gram	Concept relatedness (I)	TREC-Med (T), TREC OHSUMED (T) Pd, Cav datasets (E)[121]
Liu et al. [93]	CBOW	Literature retrieval (E) Search for cohorts	TREC OHSUMED (T, E) CLEF e-health 2014-2015 (T,E)
Choi et al. [32]	Skip-gram	Concept relatedness (I)	UMLS (E), Clinicalml dataset[29] (T) Private Database of medical claims (T)
Gosh et al. [48]	*Dis2Vec:* Skip-gram	Diagnosis classification (E)	HealthMap[30] (T, E)
Bai et al. [49]	Skip-gram	Concept relatedness (I)	ICD9 (T), MIMIC III[31] (E)
Choi et al. [29]	*GRAM:* MLP with attention mechanism	Diagnosis prediction (E)	MIMIC III (T, E) Private database of EHRs (E)
Soldaini et al. [150]	CNN:	Search for cohort (E)	TREC CDS dataset 2014-2016 (T, E)
Yu et al. [185]	CBOW	Concept relatedness	PubMed collection (T) UMNSRS-Sim [118] (E) UMNSRS-Sim [118] (E)
Cai et al. [22]	*MCE:* CBOW	Disease code clustering (I)	Private Database of EHRs (T) De-SynPUF [32] (T) CCS[33] (E) ICD9 (E)
Song et al. [154]	*MMORE:* MLP with attention mechanism	Diagnosis prediction (E)	MIMIC III (T, E)
Peng et al. [122]	*MC2Vec:* CBOW with attention mechanism	Concept clustering (I)	MIMIC III (T), CMS[34] (T) CCS (E), ICD9 (E)
Bai et al. [8]	*typeSkip-Gram:* Skip-gram	Concept clustering (I)	Database of EHRs [173] (T) ICD9 (E) ICD9 (E)
Zhang et al. [191]	*BioWordVec:* FastText	Concept clustering (I)	PubMed collection (T) UMNSRS dataset [118] (E)
Beam et al. [15]	*CUI2vec:* GLOVE	Concept clustering (I)	SNOMED-CT (E) SNOMED-CT (E)
Lee et al. [84]	*BioBERT:* BERT	Question answering (E)	PubMed collection (T) Private database of medical claims (T) SQUAD dataset[126] (T)

Table 9. Overview of recent research on neural representation learning models of medical concepts.

evaluation make the comparison between neural models both in the general and medical domain difficult. Among them, it is worth mentioning differences in the sizes of the corpus used for training and evaluation, the diversity of dimensions of the embeddings, the differences in the sizes of context windows used while training the embeddings, and the number of sources used for training the embeddings [36, 73, 116]. However, by making an in-depth analysis of the reviewed works, we can make some fair consensual conclusions across models: (1) the size of the datasets used for training the embeddings is critical since large sizes are required to achieve high quality concept embeddings. However quality of the embeddings comes at the cost of the availability of annotated corpora which is still limited as can be seen from Table 2; (2) training the embeddings over multiple corpora lead to significantly better performance regardless of the types of embeddings and tasks [28, 84, 161]. The well-established practice is to train the model on a large out-of-domain dataset (e.g., Wikipedia) and subsequently fine-tune the parameters of the model on medical datasets (e.g., Pubmed, MIMICIII). The improvements observed can be intuitively explained by the fact that heterogeneous corpora may likely differ in terms of their vocabulary, content and contexts leading to higher capabilities of generalizability of knowledge across tasks; (3) transformer-based models (BERT-based models) seem to be more effective than other language models (e.g., ELMO) in similarity- based search tasks (e.g., question answering) [84, 87, 123]. The levels of improvements are particularly impacted by the domains of the datasets used for training. The more diverse the domains to cover complex implicit semantics as hold in the literature and patient data (e.g., Wikipedia, PubMed, MIMICIII), the more effective are the embeddings to capture matching signals. Other models such as CBOW and Glove exhibit lower performances but with unclear trends with regard to varying window sizes and ground truths used [22, 122]; (4) the effectiveness results of neural models are still low and unstable across search tasks (e.g., literature retrieval, search for cohort) [93, 150, 161]. While the low performance of neural matching models have already been outlined in previous work in IR [184], the effect of the ways of integrating embeddings in matching tasks, as done in neural medical IR, is still under-studied. Regarding results instability across tasks, one possible explanation is related to the complexity of relevance appraisal (See Section 1.3) that leads to neural model failure in capturing task-agnostic relevance matching signals. Another possible explanation is the very limited size of ground-truth in search tasks (human relevance assessments) that lead to sub-optimally trained embeddings; (5) time rises as a useful factor to either implicitly (e.g., order of contexts used for learning) or explicitly incorporate (e.g., temporal attention component) into traditional language models (e.g., CBOW, Skip-gram) or into MLP-based language models of concepts [22, 28, 122]. It allows enhancing the effectiveness of both concept relatedness and diagnosis prediction tasks.

4.2 Representation Learning of Documents

4.2.1 Overview of models. Documents and short texts (sentences, paragraphs) which are viewed as more complex textual units than concepts and words, have also been the focus of a large body of works in the field of neural-based representations [83, 104, 169]. A simple but efficient approach consists of inferring the document representation by averaging embeddings of its words. As an extension of word2vec, the Paragraph-Vector model (PV) [83] jointly learns paragraph (or document) and word representations within the same embedding space. This joint learning relies on the compositional assumption underlying document representation [104, 169] leading to a mutual benefit for learning the distributional semantics of both documents and words. Leveraging these previous works, a consistent body of work proposed neural models for the representation of medical scientific publications [101, 110, 161], patients visit reports [11, 28, 110] and care episode descriptions [58, 106].

Moen et al. [11, 106] adopted the simplest way to generate care episode vectors. For instance, a care episode embedding is obtained by averaging the w2vec embeddings of all the words that belong to the care notes. The similarity between care episodes is then computed with the cosine similarity between their vectors. Regardless of the domain of application, this basic approach has shown to be efficient but with limited improvement particularly for search tasks [192]. Choi et al. [28] propose the Med2Vec model that learns patient visits embeddings using an extended Skip-gram model. Given a visit, a multi-layer perceptron first converts a binary code-based visit representation to an intermediate representation that is concatenated with demographic data and then converts them to the final latent representation. The network is trained based on patient visit sequences to predict past and future visits. The experimental evaluation carried out on private patient health records using a disease prediction task shows both the model performance gains over the Skip-gram model and the good level of interpretability of the visit representations.

Other works use document neural models to jointly learn concepts and document embeddings [101, 110, 161]. While authors in [101, 110] adapt the PV algorithm to learn the connection between the documents and concepts, Tamine et al. [161] additionally leverage the concept-to-concept relationships provided by a resource to regularize the learning objective. Nguyen et al. [110] proposed an offline learning approach of document embeddings based on the PV model [83]. They formulate the document representation task as an optimization problem based on the assumption that either using word or concepts in the flat document descriptor, both representations might lead to the same latent semantic representation. Accordingly, word-based and concept-based document embeddings are first learned in separate spaces and then the optimal hybrid vector is inferred as the closest one to them. Experiments on a search task in medical documents (scientific literature and patient visit) on TREC OHSUMED and TREC Med datasets particularly showed the relevance of the learned concepts for a QE task. However, the quality of document embeddings was not evaluated since their impact on intrinsic task performance was not addressed. This work was followed by a recent on-line model whose goal is to obtain a shared vector space for documents, words, and concepts. To achieve this, a tripartite model extends the conventional PV model to consider distributed word-concept semantic relations that could be captured from raw hybrid representations [161]. More importantly, the model learns document embeddings such that they are predictive of relational information established in structured resources. The model constrains the distributional learning model towards better revealing similarity concept-to-concept relations even if they appear in an insufficient amount of similar training contexts. More precisely, the PV learning objective is regularised towards making the representations of concepts close by considering their explicit relations in the resource. An intensive experimental evaluation has been carried out on concept relatedness and document literature retrieval tasks using TREC Med and TREC OHSUMED data. Although significant improvements have been achieved over representative document representation models (e.g., model from [104, 169]), the level of gains significantly vary from one task to another depending on a wide range of factors including source used for training the documents, and the nature of the downstream task employing those embeddings.

Grnarova et al. [58] proposed a dedicated end-to-end neural architecture to tackle health-related specific tasks. By considering a mortality prediction task, they proposed a two-layer architecture to model clinical notes as a set of separate sentences. While the first layer builds sentence vectors, the second one combines the sentence vectors and incorporates additional information such as visit category (e.g., nursing) that would help the target task. Moreover, they compute an individual softmax mortality probability for every sentence and incorporate them into the objective function. Through experiments on the MIMIC III dataset, the authors showed that the proposed neural model outperforms a topical model based on LDA as well as neural basic models based on averaged CBOW-based and doc2vec-based vectorial representations of sentences. Beyond performance gains, the study lacks in-depth qualitative analysis. As such, the relationship of the sentence characteristics and the patient class being predicted by the whole model are not known.

4.2.2 Lessons learned. Table 10 provides an overview of the reviewed representation learning models of medical documents. We report for each of them the intrinsic vs. extrinsic task as well as the datasets used for training and evaluation. We can observe from Table 10 that most embeddings rely on the PV model as the core model [101, 106, 110, 161] of documents. All the models are based on a joint objective function that exploits word or concept co-occurrences (as done in representation learning of concepts reported in Table 9) and manually labelled document contexts (e.g., scientific publication as done in [101] or a cohort description as done in [110]). Given, on one hand, the limited number of works that used document embeddings in the medical domain, and on the other hand, the limited experimental evaluation in terms of datasets, tasks and baselines, it is difficult to generalise lessons about their effectiveness. We can conjecture that document embedding is another promising approach to leverage in medical data representations; however, since no experimental evaluation has been carried out using similar datasets for similar tasks, there is a lack of clarity as to what extent these methods perform better than concept-based methods reported in Table 9 and what the intuition is behind the interference of concepts and documents that include these concepts particularly in medical search tasks. Works in the field of learning medical document embeddings are considered as preliminary works before tackling larger medical document contexts such as patient records which include multiple documents. This space of research has recently attracted a significant body of research in the domain as reported below.

4.3 Representation Learning of Patients

4.3.1 Overview of models. In simple terms, patient embedding is a single dense vectorial representation that conflates all multi-modal patient data. Simple incremental works of patient embeddings extend word (e.g., Skipgram) and document (e.g., doc2vec) embedding models by learning from sequences of patients' visits [113, 117, 156, 158]. For instance, Stojaovic et al. [156] model patient embeddings by summing the vectorial representations of discharge records that have been learned using the Skip-gram model. Each discharge disease is a sequence of disease and procedure codes. The experimental evaluation using the prediction of total charges on a hospital database of patient stays showed that the model is promising. However, the proposed model has not been compared to strong baselines such as Med2Vec [28] or GRAM [29] models. Deep neural architectures have also been proposed to handle classification and prediction tasks based on patient-related information [14, 30, 41, 58, 103, 112, 113, 163]. Convolutional neural architectures (CNN) are proposed in [112, 200]. Nguyen et al. [112] proposed the Deepr model based on CNNs. Similarly to [156], Deepr uses at the bottom level of the network raw representations of patient discharge including diagnosis and procedures. Time

[35] https://www.merckmanuals.com/

Reference	Neural model extended	Task	Used datasets
		(I: Intrinsic, E: Extrinsic)	(T: Training, E: Evaluation)
Nguyen et al. [110]	*ConceptualDoc2vec:* doc2vec	Literature retrieval (E)	TREC OHSUMED collection (T, E)
		Search for cohort (E)	TREC Med collection (T, E)
Mencia et al. [101]	doc2vec	Concept relatedness (I)	BioASQ collection (T)
			UMNSRS-Res (E)
			UMNSRS-Sim (E)
Choi et al. [28]	*Med2Vec:* word2vec	Disease prediction (I)	Private database of EHRs (T, E)
Grnarova et al. [58]	CNN	Mortality prediction (E)	MIMIC III (T, E)
Moen et al. [106]	word2vec	Private Database of EHRs (T, E)	
		retrieval	
Hughes et al. [106]	doc2vec	Text classification (E)	PubMed collection (T)
	doc2vec		Merck Manual dataset[35] (E)
Wang & Koopman [106]	*Ariadne:* doc2vec	Ad-hoc document search (E)	MEDLINE collection (T)
			Database of medical guidelines (E)
Tamine et al. [161]	*SD2V:* doc2vec		
			DBPedia (T), TREC-Med collection (T)
		Search for cohorts (E)	DBPedia (T), TREC-Med collection (E)
		Literature retrieval (E)	TREC OHSUMED (E)

Table 10. Overview of recent research on neural representation learning models of medical documents.

gaps between diseases are discretised and coded in the network input to help the prediction task. Qualitative results obtained using the risk prediction task show that Deepr better discriminates between patients. However, the comparison is only performed to bag of word (BOW) representations.

A siamese based CNN neural architecture is proposed in [200]. The overall learning framework learns in an end-to-end fashion to map patient representation to vectors, which can then be employed to compute their similarity. Each patient is represented as a matrix of embeddings of patient visits. Limited experiments have been carried out using a private database of EHRs. The PSDML model [113] is based on a quadruple network architecture that includes an anchor patient, a positive patient, a negative patient and a similar patient which are separately fed into four deep neural networks. The top layer of the network is a metric layer relying on a quadruple loss function. The latter allows a fine-grained patient similarity by computing two margins between positive and similar patients, similar and negative patients. An in-depth analysis of the patient similarity results in two private patient datasets clearly show the usefulness of the quadruplet architecture. This architecture induces however issues related to the choice of the more accurate patient samples to ensure the model performance over siamese and triplet architectures.

The temporality of patient records has been explicitly considered using appropriate neural architectures [14, 163]. Baytas et al. [14] proposed a time-aware LSTM model (T-LSTM) to handle the issue of irregular time-stamps between diseases within the patient trajectory. The main idea is to model patient embeddings by learning a subspace decomposition that better reveals patient subtypes. The input of the network is the temporal patient data which is fed into T-LSTM. The latter decomposes the previous memory into long and short term components and employs the gap times to discount the short term effects. While quantitative analyses showed the performance gains of T-LSTM over representative baselines, qualitative analyses do not provide a clear picture of patient sub-types. Tran et al. [163] proposed eNRBM, a neural model based on a Restricted Boltzmann neural network architecture and in which an input layer is connected to an output layer. The input layer consists of clinically observed variables over time (e.g., disease occurrences) and the representation layer is composed of latent factors related to discovered health-related facets. More precisely, the model embeds multi-modal data that comprises static information (e.g., gender) and healthcare trajectory through real-value events (e.g., EEG signals). This information is represented in raw vectors and provided to the network for algebraic transformations. Experimental evaluation has been carried out using a private hospital database of patients that was assessed for suicidal risk. The results showed that the eNRBM model enabled the accurate clustering of risk factors for scientific literature findings. However, the authors acknowledged

Reference	Neural model extended	Task (I: Intrinsic, E: Extrinsic)	Used datasets (T: Training, E: Evaluation)
Tran et al. [163]	*eNRBM:* Boltzmann machine	Suicide risk prediction (E) stratification Comorbidity classification	Private Database of EHRs (T, E)
Miotto et al. [103]	*DeepPatient:* AutoEncoder	Disease classification (E)	Private Database of EHRs (T, E)
Nguyen et al. [112]	*Deepr:* CNN	Risk prediction (E)	Private Database of EHRs (T, E)
Zhu et al. [200]	Convolutional network	Patient similarity (I)	Private Database of EHRs (T, E)
Baytas et al. [14]	LSTM	Future Parkinson sequence prediction (E)	Private Database of EHRs (T, E)
OrmandyI et al. [117]	Skip-gram	Patient similarity (I)	MIMIC III (T, E)
Stojanovic et al. [156]	Skip-gram	Total charges prediction (E) Mortality prediction	Private Database of EHRs (T, E)
Ni et al. [113]	*PSDML:* Siamese neural network	Patient similarity (I)	Private Database of EHRs (T, E)
Sushil et a. [158]	doc2vec Autoencoder	Patient mortality prediction (E) Disease prediction	MIMIC III (T, E)
Dligach & Miller [41]	MLP	Disease prediction (E)	MIMIC III (T) I2b2 challenge dataset [55] (E)
Choi et al. [30]	*MIME:* MLP	Health failure prediction	Private Database of EHRs (T, E)
Choi et al. [31]	Graph convolutional transformer	Mortality prediction (E)	Collaborative research dataset[36] (T, E)

Table 11. Overview of recent research on neural representation learning models of patients.

that some of the retrieved clusters are not clinically relevant and that it is unclear if this result could be more crucially impacted when using other clinical data collections. Interestingly, Choi et al. [30] propose Multilevel Medical Embedding (MiME) which learn the multilevel embedding of EHR data while jointly performing prediction tasks. The tasks incorporate knowledge provided by EHR data into the embedding process such that the main task can leverage prediction power from related auxiliary tasks. Works previously described [28, 101] often conflate codes and treatment into a single visit representations. On the contrary, the MIME model explicitly discriminates the hierarchy (patient-level) between diagnosis codes (diagnosis level), treatment codes (treatment level) and their relationships in separate levels. MIME model has been evaluated using sequential disease prediction and showed significant improvement over the Med2Vec and GRAM models. Incorporating additional patient data such as demographic information and procedure instructions is left for the future.

4.3.2 Lessons learned. Table 11 provides an overview of the reviewed representation learning models of patients. We report for each of them the intrinsic vs. extrinsic task as well as the datasets used for training and evaluation. From Table 11, we can see that most of the research work in patient representation learning has been carried out on private databases of EHRs, making across-model comparison not possible. This observation can be expected from the nature of the task, since patient data is constrained by privacy-concerns due to its sensitive content. A close look at the proposed models lead us however to make the following general statements: (1) patient embeddings are generally processed based on a multi-level or multi-step training; a supervised training level on diagnosis codes within patient visits and a semi-or non supervised training level over the longitudinal structure of the patient record. This two-step process makes the models able to capture hierarchical patterns and dependencies in patient data allowing them to be effective for a wide range of downstream tasks including both similarity-based tasks (e.g., disease classification) and prediction tasks (e.g., risk prediction, disease prediction) [30, 31, 41, 103]; (2) Deep neural architectures seem to be more suited (than feed-forward networks) to capture patient-related data patterns as well as to detect patient similarities as outlined in comparative experiments [30, 31] and can be inferred from model architecture adopted in related works [14, 30, 31, 113]; (3) Model convergence and effectiveness of the models heavily depend on multiple factors including the balance in disease classes among patients [113, 163], patient sampling strategy, mainly in the case of end-to-end learning of patient similarities [113], and elapsed time irregularities to observe target disease codes [14]. This line of work is still at its early stage. Further research collaboration is needed to tackle the barriers of shareable data and tasks.

5 CONCLUSION AND CHALLENGES AHEAD

Summary. In this survey, we reviewed the literature on semantic IR in the medical domain. We started by introducing the medical domain, the type of data and the existing major semantic resources. Then, we described how the medical domain was particularly challenging for IR: the semantic gap, the vocabulary mismatch, and the complexity of result appraisal. We presented a comprehensive overview of the major models and techniques, among which we distinguished structured *knowledge resource-driven approaches*, and *data-driven approaches*. *Knowledge-resources driven approaches* fall into 2 categories. The first one relates to QE approaches, where terms are added to the query to retrieve more relevant documents. These expansion terms are identified in a local or a global context, using one or multiple knowledge resources. The second category includes feature-based relevance ranking methods, hand-crafted based learning methods are applied to representations of documents that include semantic information, to learn a new rank order for documents. A cross-model analysis allowed us to provide a comparative evaluation. *Data driven approaches* particularly include neural representation learning models, where either an item semantic embedding is learned from raw text or a relevance function is learned in an end-to-end fashion. Those embeddings are generally used within various intrinsic and extrinsic tasks, ranging from concept relatedness to disease prediction. From the deep analysis of the representation learning models and their performance results, we provided trending lessons.

The main conclusion that one can draw from the literature review is that the increasing interest in the IR and health-informatics communities in medical search have led to significant scientific progress and the creation of various resources, increasingly allowing shareable and reproducible evaluations. However, research collaboration and shared views between both communities are still at an early stage. Further investigation is required to tackle pending challenges among which we identify those summarized below.

Challenges ahead and future research directions.
Semantic reasoning beyond semantic search. While the IR and health informatics communities actively design tasks assisting medical practitioners' information needs, the semantic IR models designed are not necessarily closely related to the actual information flow in the medical practice. Particularly, for tasks such as diagnosis that are inherently multi-step search tasks and induce a hypothetico-deductive cognitive process, professionals require support that extends beyond a query-response paradigm. Such tasks are acknowledged as complex [79] since beyond the semantic query-document matching, they should incorporate the decision-making process of the user (here the medical professional). The real required assistance of medical professionals is not only to fill the semantic gap but also to accomplish a task. One promising research opportunity would be the design of models which better exploit the expressiveness power of knowledge resources by incorporating the reasoning ability they offer. Beyond using simple structural relationships between knowledge nuggets (e.g., concepts) to improve document and query representations, improved knowledge representation and reasoning models are required to design deductive processes.

Acceptability and explainability of neural models in medical IR. Neural IR in general and neural medical search specifically is not yet at a mature stage [116, 184]. The early research and progress made so far gave rise to several issues that could lead to research opportunities in the near future. First, the need for a very large amount of training data is still a bottleneck particularly in the medical domain where expert-based ground truth is costly to collect. Various methods including transfer learning are explored to compensate for the lack of training data [179]. However, in order to gain maturity, cross-disciplinary research is required to target the learning strategies to specific medical search tasks. Second, the other challenge of neural approaches is the black box effect: to date, we are not able to explain their results. AI regulations, such as the EU General Data Protection Regulation (GDPR) present the "principle of transparency" of intelligent algorithms and emphasize the fact that algorithmic decisions in AI systems should be explainable. Recent efforts have been made in the recommendation and IR communities. As an example, Zhang et al [194, 195] have organized in SIGIR the *Workshop on Explainable Recommendation and Search (EARS).* Studies in this new domain cover a wide range

[36] https://eicu-crd.mit.edu/about/eicu/

of topics, including the design of explainable models and systems [146], investigations of various types of explanations [47], studies aiming at understanding users needs and interaction with the results [26, 95]. In domains such as medicine, not understanding a system's output is particularly critical and effort needs to be exerted towards system explainability. Legal and ethical acceptability of medical search systems integrating deep learning strategies is still questionable. Acceptability is inherently related to the ability to explain and interpret the results and the confidence that can be placed in them by medical professionals. Research has shown that professional searchers value transparency more than ranking performances [137]. One relevant research opportunity is the design of hybrid models that combine human-driven and data-driven learning objectives. This research trend has recently been initiated [96] but is still at a very early stage. Qu et al [125] propose a novel retrieval model that emulates how medical experts make structured relevance judgments. To do so, they train document classifiers based on documents aspects (e.g., disease, gene, demographics) in order to mimic experts cognitive process. As shown by these approaches, a focus on interpretable systems is needed in the medical domain, to avoid providing users with automatically generated "explanations" of complex blackbox systems that can be unreliable and misleading [136]. In that sense, [26, 147] provide very insightful recommendations to make systems explainable. Sokol [147] publish a taxonomy to characterise and assess explainable systems along five key dimensions: functional, operational, usability, safety and validation. In order to gain maturity, we clearly need: (1) to empower system design with user-centered concerns (either professional or patient) such as fairness, privacy and transparency, and (2) to design suitable evaluation frameworks, including datasets and measures, which allow for the comparison of the effectiveness of system decisions with medical professional-style decisions.

Cross-discipline collaboration for low-resource languages. Semantic IR strongly relies on semantic resources and semantic annotation. The English language is well covered, with several thesauri focusing on specialized fields, and the UMLS metathesaurus gathering hundreds of thesaurus, and several semantic annotation tools such as MetaMap. However, languages other than English are far from being as covered as English, and substantial effort is required to be able to assist medical professionals across the globe. A huge effort is needed from various disciplines including IR, NLP, linguistics and health-informatics to cite but a few, toward to build shareable resources. Thus, another relevant investigation would be the study of transfer learning strategies to exploit the relationship between rich-resource and low-resource language lexicons. Recent neural translation models [7] offer promising opportunities but the need for a huge amount of annotated corpora and the lack of pivotal resources still hampers their success. Further work is needed toward automatically building training data based on expert-style weak supervision.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Rie Kubota Ando, Mark Dredze, and Tong Zhang. 2005. TREC 2005 Genomics Track Experiments at IBM Watson. In *Proceedings of TREC 2005*.

[3] Alan R Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *JAMIA* 17, 3 (2010), 229–236.

[4] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. 2004. The NLM indexing initiative's medical text indexer. In *Medinfo*. 268–272.

[5] Mordechai Auerbuch, Tom H Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. 2004. Context-Sensitive Medical Information Retrieval. *Studies in health technology and informatics* 107 (2004), 282–6.

[6] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[8] Tian Bai, Brian L. Egleston, Richard Bleicher, and Slobodan Vucetic. 2019. Medical Concept Representation Learning from Multi-source Data. In *Proceedings of IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 4897–4903.

[9] Saeid Balaneshin-kordan and Alexander Kotov. 2016. Optimization Method for Weighting Explicit and Latent Concepts in Clinical Decision Support Queries. In *ICTIR '16* (Newark, Delaware, USA). ACM, New York, NY, USA, 241–250.

[10] Saeid Balaneshinkordan and Alexander Kotov. 2019. Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine. *Journal of Biomedical Informatics* 98 (2019), 103238.

[11] Imon Banerjee, Matthew C. Chen, Matthew Lungren, and Daniel Rubin. 2017. Radiology Report Annotation using Intelligent Word Embeddings: Applied to Multi-institutional Chest CT Cohort. *Journal of Biomedical Informatics* 77 (2017).

[12] Hannah Bast, Bjorn Buchhold, and Elmar Haussmann. 2016. Semantic Search on Text and Knowledge Bases. *Foundations and Trends in Information Retrieval* 10, 2-3 (2016), 119–271.

[13] Holger Bast, Debapriyo Majumdar, and Ingmar Weber. 2007. Efficient Interactive Query Expansion with Complete Search. In *CIKM '07* (Lisbon, Portugal). 857–860.

[14] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *ACM SIGKDD*. 65–74.

[15] Andrew L. Beam, Benjamin Kompa, Inbar Fried, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2018. Clinical Concept Embeddings Learned from Massive Sources of Medical Data. *CoRR* abs/1804.01486 (2018). arXiv:1804.01486

[16] Maksim Belousov, William G. Dixon, and Goran Nenadic. 2019. MedNorm: A Corpus and Embeddings for Cross-terminology Medical Concept Normalisation. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. Association for Computational Linguistics, 31–39.

[17] Jagdev Bhogal, Andrew Macfarlane, and Peter Smith. 2007. A Review of Ontology Based Query Expansion. *Inf. Process. Manage.* 43, 4 (2007), 866–886.

[18] Bodo Billerbeck, Falk Scholer, Hugh E. Williams, and Justin Zobel. 2003. Query expansion using associated queries. In *Proceedings of CIKM 2003*. 2–9.

[19] Li Bin and K C Lun. 2001. The retrieval effectiveness of medical information on the web. *International Journal of Medical Informatics* 62, 2 (2001), 155 – 163.

[20] Olivier Bodenreider. 2005. The Unified Medical Language System Integrating Biomedical Terminology. (2005).

[21] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. *Proceedings of ECIR 2016*.

[22] Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. 2018. Medical Concept Embedding with Time-Aware Attention. In *Proceedings of IJCAI 2018*.

[23] Leonardo Campillos-Llanos. 2019. First Steps towards a Medical Lexicon for Spanish with Linguistic and Semantic Information. *BioNLP 2019* (2019), 152.

[24] YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J. Cimino, John Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics* 44, 2 (2011), 277 – 288.

[25] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* 44, 1 (2012), 1:1–1:50.

[26] Larissa Chazette and Kurt Schneider. 2020. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering* (2020), 1–22.

[27] Jianpeng Cheng, Zhongyuan Wang, Ji-Rong Wen, Jun Yan, and Zheng Chen. 2015. Contextual Text Understanding in Distributional Semantic Space. In *Proceedings of CIKM 2015*. 133–142.

[28] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of ACM SIGKDD 2016*. 1495–1504.

[29] Edward Choi, Taha Bahadori, Le Song, Walter Stewart, and J. Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proceedings of KDD 2017*. 787–795.

[30] Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In *Proceedings of NIPS'18* (Montr&#233;al, Canada). 4552–4562.

[31] Edward Choi, Zhen Xu, Yujia Li, Michael W. Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M. Dai. 2020. Graph Convolutional Transformer: Learning the Graphical Structure of Electronic Health Records. In *Association of Artificial Intelligence AAAI*.

[32] Sontag D. Choi Y, Chiu CY-I. 2016. Learning Low-Dimensional Representations of Medical Concepts. In *AMIA Summits on Translational Science Proceedings*. 41–50.

[33] Gobinda G. Chowdhury. 1999. *Introduction to modern information retrieval.* Library Association Publishing London. xix, 452 p. : pages.

[34] Cyril W Cleverdon, Jack Mills, and E Michael Keen. 1966. Factors determining the performance of indexing systems, (Volume 1: Design). *Cranfield: College of Aeronautics* (1966), 28.

[35] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical Semantic Similarity with a Neural Language Model. In *Proceedings of CIKM 2014*. 1819–1822.

[36] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical Semantic Similarity with a Neural Language Model. In *CIKM*. 1819–1822.

[37] Dina Demner-Fushman, Susanne M Humphrey, Nicholas C Ide, Russell F Loane, Patrick Ruch, Miguel E Ruiz, Lawrence H Smith, Lorraine K Tanabe, W John Wilbur, and Alan R Aronson. 2006. Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort. In *Proceedings of TREC 2006*.

[38] Duy Dinh and Lynda Tamine. 2012. Towards a context sensitive approach to searching information based on domain specific knowledge sources. *Web Semantics: Science, Services and Agents on the World Wide Web* 12, 0 (2012).

[39] Duy Dinh, Lynda Tamine, and Fatiha Boubekeur. 2013. Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine* 57, 2 (2013), 155 – 167.

[40] Ram Dixit, Deevakar Rogith, Vidya Narayana, Mandana Salimi, Anupama Gururaj, Lucila Ohno-Machado, Hua Xu, and Todd R Johnson. 2018. User needs analysis and usability assessment of DataMed–a biomedical data discovery index. *JAMIA* 25, 3 (2018), 337–344.

[41] Dmitriy Dligach and Timothy Miller. 2018. Learning Patient Representations from Text. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 119–123.

[42] Tracy Edinger, Aaron M. Cohen, Steven Bedrick, Kyle Ambert, and William Hersh. 2012. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track. *AMIA Symposium* 2012 (2012), 2012.

[43] Marco Eichelberg, Thomas Aden, Jörg Riesmeier, Asuman Dogac, and Gokce B. Laleci. 2005. A Survey and Analysis of Electronic Healthcare Record Standards. *ACM Comput. Surv.* 37, 4 (2005), 277–315.

[44] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. 2016. Computational Health Informatics in the Big Data Age: A Survey. *ACM Comput. Surv.* 49, 1 (2016), 12:1–12:36.

[45] Samuel G. Finlayson, Paea LePendu, and Nigam H. Shah. 2014. Building the graph of medicine from millions of clinical narratives. *Journal Scientific Data* 1, 1 (2014), 140032.

[46] Sumio Fujita. 2004. Revisiting Again Document Length Hypotheses TREC 2004 Genomics Track Experiments at Patolis. In *Proceedings of TREC 2004*.

[47] Abraham Gale and Amélie Marian. 2019. Metrics for Explainable Ranking Functions. In *Proceedings of the SIGIR EARS workshop 2019*.

[48] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S. Brownstein, and Naren Ramakrishnan. 2016. Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2Vec Approach. In *Proceedings of CIKM '16*. 1129–1138.

[49] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S. Brownstein, and Naren Ramakrishnan. 2017. Joint Learning of Representations of Medical Concepts and Words from EHR Data. In *Proceedings (IEEE Int Conf Bioinformatics Biomed)*. 764—-769.

[50] Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. 2013. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. *CLEF 2013 Online Working Notes* 8138 (2013).

[51] Lorraine Goeuriot, Gareth J. F. Jones, Liadh Kelly, Henning Müller, and Justin Zobel. 2016. Medical information retrieval: introduction to the special issue. *Information Retrieval Journal* 19, 1 (2016), 1–5.

[52] Lorraine Goeuriot, Liadh Kelly, Leif Hanlen, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. 2015. Overview of the CLEF eHealth Evaluation Lab 2015. In *Proceedings of CLEF 2015*.

[53] Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth JF Jones, and Henning Mueller. 2014. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proceedings of CLEF 2014*.

[54] Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. 2017. CLEF 2017 eHealth Evaluation Lab Overview. In *Proceedings of CLEF 2017*.

[55] Ira Goldstein and Özlem Uzuner. 2009. Specializing for predicting obesity and its co-morbidities. *J. Biomed. Informatics* 42, 5 (2009), 873–886.

[56] Travis R. Goodwin and Sanda M. Harabagiu. 2016. Medical Question Answering for Clinical Decision Support. In *CIKM '16* (Indianapolis, Indiana, USA). ACM, New York, NY, USA, 297–306.

[57] Nicolas Griffon, Gaétan Kerdelhué, Saliha Hamek, Sylvain Hassler, César Boog, Jean-Baptiste Lamy, Catherine Duclos, Alain Venot, and Stéfan J Darmoni. 2014. Design and usability study of an iconic user interface to ease information retrieval of medical guidelines. *JAMIA* 21, e2 (2014), e270–e277.

[58] Paulina Grnarova, Florian Schmidt, Stephanie L. Hyland, and Carsten Eickhoff. 2016. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. (2016). arXiv:1612.00467 [cs.CL]

[59] David A. Grossman and Ophir Frieder. 1998. *Information Retrieval: Algorithms and Heuristics.* Kluwer Academic Publishers, Norwell, MA, USA.

[60] William Hersh. 2010. *Information Retrieval - A Health and Biomedical Perspective.* Springer.

[61] William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proceedings of SIGIR 1994* (Dublin, Ireland). 192–201.

[62] William Hersh, Aaron Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe Roberts, and Marti Hearst. 2005. TREC 2005 Genomics Track Overview. In *Proceedings of TREC 2005.*

[63] William R Hersh and Ravi Teja Bhupatiraju. 2003. TREC genomics track overview. In *Proceedings of TREC 2003*, Vol. 2003. 14–23.

[64] William R Hersh, Ravi Teja Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron M. Cohen, and Dale F. Kraemer. 2004. TREC 2004 Genomics Track Overview. In *Proceedings of TREC 2004.*

[65] William R Hersh, Aaron M Cohen, Phoebe M Roberts, and Hari Krishna Rekapalli. 2006. TREC 2006 genomics track overview. In *Proceedings of TREC 2006.*

[66] William R Hersh, Aaron M Cohen, Lynn Ruslen, and Phoebe Roberts. 2007. TREC 2007 genomics track overview. In *Proceedings of TREC 2007.*

[67] Katja Hofmann, Lihong Li, Filip Radlinski, et al. 2016. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval* 10, 1 (2016), 1–117.

[68] Xiangji Huang, Ming Zhong, and Luo Si. 2005. York University at TREC 2005: Genomics Track. In *Proceedings of TREC 2005.*

[69] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[70] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[71] Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. 2014. Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at ImageCLEF 2004-2013. *Computerized Medical Imaging and Graphics* 39 (2014), 55–61.

[72] KS Kalyan and S. Sangeetha. 2019. SECNLP: A Survey of Embeddings in Clinical Natural Language Processing. *ArXiv* abs/1903.01039 (2019).

[73] Katikapalli Subramanyam Kalyan and S Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics* 101 (2020), 103323.

[74] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In *Proceedings of CLEF 2017 (Working Notes).*

[75] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[76] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. 2016. Overview of the CLEF eHealth Evaluation Lab 2016. In *Proceedings of CLEF 2016.*

[77] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, Guido Zuccon, and Joao Palotti. 2014. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In *Proceedings of CLEF 2014.*

[78] Benjamin King, Lijun Wang, Ivan Provalov, and Jerry Zhou. 2011. Cengage Learning at TREC 2011 Medical Track.. In *TREC.*

[79] Marijn Koolen, Jaap Kamps, Toine Bogers, Nicholas J. Belkin, Diane Kelly, and Emine Yilmaz. 2017. Report on the Second Workshop on Supporting Complex Search Tasks. *SIGIR Forum* 51, 1 (2017), 58–66.

[80] Bevan Koopman and Guido Zuccon. 2014. Why assessing relevance in medical IR is demanding. In *Proceedings of SIGIR 2014.*

[81] Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2016. Information Retrieval As Semantic Inference: A Graph Inference Model Applied to Medical Search. *Inf. Retr.* 19, 1-2 (2016), 6–37.

[82] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of SIGIR 2001* (New Orleans, Louisiana, USA). 120–127.

[83] Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of ICML 2014.* 1188–1196.

[84] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2019), 1234–1240.

[85] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4

(2020), 1234–1240.

[86] Geoffrey Leech. 2005. Adding linguistic annotation. (2005).

[87] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Medical Informatics* 7 (2019), e14830.

[88] Hang Li and Jun Xu. 2014. *Semantic Matching in Search.* Now Publishers Inc., Hanover, MA, USA.

[89] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013. Inferring Conceptual Relationships to Improve Medical Records Search. In *Proceedings of OAIR 2013* (Lisbon, Portugal). 1–8.

[90] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013. Learning to Handle Negated Language in Medical Records Search. In *Proceedings of CIKM '13* (San Francisco, California, USA). 1431–1440.

[91] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. 2013. Learning to selectively rank patients' medical history. In *Proceedings of CIKM '13* (San Francisco, California, USA). 1833–1836.

[92] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.* 3, 3 (2009), 225–331.

[93] Xiaojie Liu, Jian-Yun Nie, and Alessandro Sordoni. 2016. Constraining word embeddings by prior knowledge–application to medical information retrieval. In *Information Retrieval Technology*. Springer, 155–167.

[94] Henry J Lowe, Todd A Ferris, Penni M Hernandez, and Susan C Weber. 2009. STRIDE–An integrated standards-based translational research informatics platform. In *AMIA Annual Symposium Proceedings*, Vol. 2009. American Medical Informatics Association, 391.

[95] Francisco Lupiáñez-Villanueva, George Gaskell, Pietro Tornese, José Vila, Yolanda Gómez, Anthony Allen, Giuseppe Veltri, and Cristiano Codagnone. 2018. Behavioural study on the transparency of online platforms. (2018).

[96] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk Prediction on Electronic Health Records with Prior Medical Knowledge. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018).

[97] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

[98] Kornél Markó, Robert Baud, Pierre Zweigenbaum, Lars Borin, Magnus Merkel, and Stefan Schulz. 2006. Towards a multilingual medical lexicon. In *AMIA Annual Symposium Proceedings*, Vol. 2006. American Medical Informatics Association, 534.

[99] Melvin Earl MARON and John Larry KUHNS. 1960. On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM* 7, 3 (1960), 216–244.

[100] David Martinez, Arantxa Otegi, Aitor Soroa, and Eneko Agirre. 2014. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *Journal of Biomedical Informatics* 51 (2014), 100 – 106.

[101] Eneldo Loza Mencia, Gerard De Melo, and Jinseok Nam. 2016. Medical concept embeddings via labeled background corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* 4629–4636.

[102] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[103] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T Dudley. 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. In *Scientific reports.*

[104] Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *ACL*. 236–244.

[105] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 1 (2018), 1–126.

[106] H Moen, F Ginter, E Marsi, L-M Peltonen, Salakoski T, and Salanterä S. 2015. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making* 15 (2015).

[107] Robert Moskovitch, Fei Wang, Jian Pei, and Carol Friedman. 2017. JASIST special issue on biomedical information retrieval. *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2525–2528. https://doi.org/10.1002/asi.23972

[108] Henning Mueller, Paul Clough, Thomas Deselaers, and Barbara Caputo. 2010. *ImageCLEF – Experimental evaluation of visual information retrieval.* Springer.

[109] Aurélie Névéol, Julien Grosjean, Stéfan Jacques Darmoni, Pierre Zweigenbaum, et al. 2014. Language Resources for French in the Biomedical Domain. In *Proceedings of LREC 2014.* 2146–2151.

[110] Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, and Nathalie Souf. 2017. Learning Concept-Driven Document Embeddings for Medical Information Search. In *Proceedings of AIME 2017.* Vienna, Austria.

[111] Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, and Nathalie Souf. 2018. A Tri-Partite Neural Document Language Model for Semantic Information Retrieval. In *Proceedings of ESWC 2018*, Vol. 10843. Heraklion, Crète, Greece, 445–461.

[112] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics* 2 (2017), 22–30.

[113] Jiazhi Ni, Jie Liu, Chenxin Zhang, Dan Ye, and Zhirou Ma. 2017. Fine-grained Patient Similarity Measuring Using Deep Metric Learning. In *Proceedings of CIKM 2017* (Singapore, Singapore). 1189–1198.

[114] Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22, 3 (2015), 671–681.

[115] Heung-Seon Oh and Yuchul Jung. 2015. Cluster-based query expansion using external collections in medical information retrieval. *Journal of Biomedical Informatics* 58 (2015), 70 – 79.

[116] Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, Maarten de Rijke, and Matthew Lease. 2018. Neural information retrieval: at the end of the early years. *Information Retrieval Journal* 21, 2 (2018), 111–182.

[117] Christopher Ormandy, Zina M. Ibrahim, and Richard J. B. Dobson. 2017. Learning Patient Similarity Using Joint Distributed Embeddings of Treatment and Diagnoses. In *Proceedings of IJCAI 2017*. 30–35.

[118] Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, Vol. 2010. American Medical Informatics Association, 572.

[119] Joao Palotti, Allan Hanbury, Henning Müller, and Charles E. Kahn, Jr. 2016. How Users Search and What They Search for in the Medical Domain. *Inf. Retr.* 19, 1-2 (2016), 189–224.

[120] Min Pan, Yue Zhang, Qiang Zhu, Bo Sun, Tingting He, and Xingpeng Jiang. 2019. An adaptive term proximity based rocchio's model for clinical decision support retrieval. *BMC Medical Informatics Decis. Mak.* 19-S, 9 (2019), 251.

[121] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* 40, 3 (2007), 288 – 299.

[122] Xueping Peng, Guodong Long, Shirui Pan, Jing Jiang, and Zhendong Niu. 2019. Attentive Dual Embedding for Understanding Medical Concepts in Electronic Health Records. In *Proceedings of IJCNN 2019*. 1–8.

[123] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of BioNLP 2019*. 58–65.

[124] Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of LBM* (2013), 39–44.

[125] Jiaming Qu, Jaime Arguello, and Yue Wang. 2020. Towards Explainable Retrieval Models for Precision Medicine Literature Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1593–1596.

[126] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP 2016*.

[127] Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Peng, Xueping and Long, Guodong and Pan, Shirui and Jiang, Jing and Niu, ZhendongIJCAI'95*. 448–453.

[128] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *JAMIA* 27, 9 (2020), 1431–1436.

[129] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. *Journal of the American Medical Informatics Association* (2020).

[130] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, and William R Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *Proceedings of TREC 2016*.

[131] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *Proceedings of TREC 2017*.

[132] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2018. Overview of the TREC 2018 Precision Medicine Track. In *Proceedings of TREC 2018*.

[133] Kirk Roberts, Matthew S Simpson, Ellen Voorhees, and William R Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of TREC 2015*.

[134] Stephen Robertson and David A Hull. 2000. The TREC-9 filtering track final report. In *Proceedings of TREC 2000*, Vol. 10.

[135] Stephen Robertson, Cornelis J. van Rijsbergen, and Martin F. Porter. 1981. Probabilistic Models of Indexing and Searching. In *Proceedings of SIGIR '80* (Cambridge, England). 35–56.

[136] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[137] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[138] Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended Boolean Information Retrieval. *Commun. ACM* 26, 11 (1983), 1022–1036.

[139] Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA.

[140] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *JAMIA* 17, 5 (2010), 507–513.

[141] Hinrich Schütze and Jan O. Pedersen. 1995. Information Retrieval Based on Word Senses. *Computational Linguistics* 24, 3 (1995), 97–123.

[142] Miftahutdinov Z Sh, Tutubalina EV, and Tropsha AE. 2017. Identifying disease-related expressions in reviews using conditional random fields. *Computational Linguistics and Intellectual Technologies* (2017), 155–166.

[143] Wei Shen, Jian-Yun Nie, Xiaohua Liu, and Xiaojie Liui. 2014. An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014eHealthTask 3. *Proceedings of CLEF 2014 (Working Notes)* (2014).

[144] Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: a survey of recent progress. In *Mining text data.* Springer, 465–517.

[145] Matthew S Simpson, Ellen M. Voorhees, and William Hersh. 2014. Overview of the TREC 2014 Clinical Decision Support Track. In *Proceedings of TREC 2014.*

[146] Jaspreet Singh and Avishek Anand. 2018. Posthoc Interpretability of Learning to Rank Models using Secondary Training Data. *CoRR* abs/1806.11330 (2018).

[147] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 56–67.

[148] Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *Proceedings of the MedIR workshop, SIGIR.*

[149] Luca Soldaini and Nazli Goharian. 2017. Learning to Rank for Consumer Health Search: A Semantic Approach. In *Advances in Information Retrieval*, Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). 640–646.

[150] Luca Soldaini, Andrew Yates, and Nazli Goharian. 2017. Denoising Clinical Notes for Medical Literature Retrieval with Convolutional Neural Model. In *Proceedings of CIKM 2017.* 2307–2310.

[151] Luca Soldaini, Andrew Yates, and Nazli Goharian. 2017. Learning to reformulate long queries for clinical decision support. *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2602–2619.

[152] Luca Soldaini, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. 2016. Enhancing Web Search in the Medical Domain via Query Clarification. *Inf. Retr.* 19, 1-2 (2016), 149–173.

[153] P Sondhi, J Sun, C Zhai, R Sorrentino, and MS Kohn. 2012. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. *JAMIA* 19, 5 (2012), 851–858.

[154] Lihong Song, Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. 2019. Medical Concept Embedding with Multiple Ontological Representations. In *IJCAI-19.* International Joint Conferences on Artificial Intelligence Organization, 4613–4619.

[155] Yang Song, Yun He, Qinmin Hu, Liang He, and E Mark Haacke. 2015. ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. In *Proceedings of CLEF 2015 (Working Notes).*

[156] Jelena Stojanovic, Djordje Gligorijevic, Vladan Radosavljevic, Nemanja Djuric, Mihajlo Grbovic, and Zoran Obradovic. 2017. Modeling Healthcare Quality via Compact Representations of Electronic Health Records. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 14, 3 (2017), 545–554.

[157] Nicola Stokes, Yi Li, Lawrence Cavedon, and Justin Zobel. 2009. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval* 12, 1 (2009), 17–50.

[158] Madhumita Sushil, Simon uster, Kim Luyckx, and Walter Daelemans. 2018. Patient representation learning and interpretable evaluation using clinical notes. *Journal of Biomedical Informatics* 84 (2018), 103 – 113.

[159] Lynda Tamine and Cecile Chouquet. 2017. On the Impact of Domain Expertise on Query Formulation, Relevance Assessment and Retrieval Performance in Clinical Settings. *Inf. Process. Manage.* 53, 2 (2017), 332–350.

[160] Lynda Tamine, Cecile Chouquet, and Thomas Palmer. 2015. Analysis of Biomedical and Health Queries: Lessons Learned from TREC and CLEF Evaluation Benchmarks. *JASIST* 66, 12 (2015), 2626–2642.

[161] Lynda Tamine, Laure Soulier, Gia-Hung Nguyen, and Nathalie Souf. 2019. Offline Versus Online Representation Learning of Documents Using External Knowledge. *ACM Trans. Inf. Syst.* 37, 4, Article 42 (2019), 34 pages.

[162] Harsh Thakkar, Ganesh Iyer, Kesha Shah, and Prasenjit Majumder. 2014. Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered Information Retrieval System for Clinical Documents.. In *CLEF (Working Notes).* 248–259.

[163] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. 2015. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of Biomedical Informatics* 54 (2015), 96 – 105.

[164] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics* 16, 1 (2015), 138.

[165] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[166] Federica Vezzani and Giorgio Maria Di Nunzio. 2020. On the Formal Standardization of Terminology Resources: The Case Study of TriMED. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 4903–4910.

[167] Elen Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR 1998*. 315–323.

[168] E. M. Voorhees and W. Hersh. 2012. Overview of the TREC 2012 medical records. In *Proceedings of TREC 2012*.

[169] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of SIGIR 2015*. ACM, 363–372.

[170] Chunye Wang and Ramakrishna Akella. 2015. Concept-Based Relevance Models for Medical and Semantic Information Retrieval. In *Proceedings of CIKM '15*. 173–182.

[171] Chunye Wang and Ramakrishna Akella. 2015. Concept-Based Relevance Models for Medical and Semantic Information Retrieval. In *CIKM*. ACM, 173–182.

[172] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* (2020).

[173] Joan L Warren, Carrie N Klabunde, Deborah Schrag, Peter B Bach, and Gerald F Riley. 2002. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical care* (2002), IV3–IV18.

[174] Gesa Weske-Heck, Albrecht Zaiss, Matthias Zabel, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rudiger Klar. 2002. The German specialist lexicon. In *AMIA Symposium*. American Medical Informatics Association, 884.

[175] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of WSDM '09*. 132–141.

[176] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Trans. Inf. Syst.* 27, 4 (2009), 23:1–23:37.

[177] Ryen W. White and Eric Horvitz. 2012. Studies of the Onset and Persistence of Medical Concerns in Search Logs. In *Proceedings of SIGIR 2012*. 265–274.

[178] Ryen W. White and Eric Horvitz. 2013. Captions and Biases in Diagnostic Search. *ACM Trans. Web* 7, 4, Article 23 (2013), 28 pages.

[179] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA* 25, 10 (2018), 1419–1428.

[180] Chenyan Xiong and Jamie Callan. 2015. EsdRank: Connecting Query and Documents Through External Semi-Structured Data. In *Proceedings of CIKM '15*. 951–960.

[181] Bo Xu, Hongfei Lin, and Yuan Lin. 2019. Learning to Refine Expansion Terms for Biomedical Information Retrieval Using Semantic Resources. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 3 (2019).

[182] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv.* 50, 6 (2018), 85:1–85:40.

[183] Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. 2011. Toward a Semantic Granularity Model for Domain-specific Information Retrieval. *ACM Trans. Inf. Syst.* 29, 3 (2011), 15:1–15:46.

[184] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of SIGIR C2019*. 1129–1132.

[185] Zhiguo Yu, Byron C Wallace, Todd Johnson, and Trevor Cohen. 2017. Retrofitting Concept Vector Representations of Medical Concepts to Improve Estimates of Semantic Similarity and Relatedness. *Studies in health technology and informatics* 245, Article 42 (2017), 657–661 pages.

[186] Hamed Zamani and W Bruce Croft. 2016. Estimating embedding vectors for queries. In *ICTIR*. ACM, 123–132.

[187] Qing T Zeng, Sandra Kogan, Robert M Plovnick, Jonathan Crowell, Eve-Marie Lacroix, and Robert A Greenes. 2004. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics* 73, 1 (2004), 45 – 55.

[188] Yan Zhang. 2010. Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community. In *Proceedings of IHI '10* (Arlington, Virginia, USA). 210–219.

[189] Y. Zhang. 2013. Searching for specific health-related information in MedlinePlus: behavioral patterns and user experience. *JASIST* (2013).

[190] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data* 6, 1 (2019), 1–9.

[191] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Improving biomedical word embeddings with subword information and MeSH. In *Scientific Data, 2019*. International Joint Conferences on Artificial Intelligence Organization.

[192] Ye Zhang, Md. Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Hong Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. 2016. Neural Information Retrieval: A Literature Review. *ArXiv* abs/1611.06792 (2016).

[193] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of SIGIR '14*. 435–444.

[194] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2018. SIGIR 2018 Workshop on ExplainAble Recommendation and Search (EARS 2018). In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1411–1413. https://doi.org/10.1145/3209978.3210193

[195] Yongfeng Zhang, Yi Zhang, Min Zhang, and Chirag Shah. 2019. EARS 2019: the 2nd international workshop on explainable recommendation and search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1438–1440.

[196] Wei Zhou, T Yu Clement, Vetle I Torvik, and Neil R Smalheiser. 2006. A Concept-Based Framework for Passage Retrieval at Genomics. In *Proceedings of TREC 2006*, Vol. 8. 8–2.

[197] Xuezhong Zhou, Zhaohui Wu, Aining Yin, Lancheng Wu, Weiyu Fan, and Ruen Zhang. 2004. Ontology development for unified traditional Chinese medical language system. *Artificial Intelligence in Medicine* 32, 1 (2004), 15–27.

[198] Donqqing Zhu and Ben Carterette. 2012. Improving health records search using multiple query expansion collections. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*. 1–7.

[199] Dongqing Zhu, Stephen T Wu, James J Masanz, Ben Carterette, and Hongfang Liu. 2013. Using Discharge Summaries to Improve Information Retrieval in Clinical Domain. In *Proceedings of CLEF 2013 (Working Notes)*.

[200] Zihao Zhu, Changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. 2016. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. In *Proceedings of ICDM 2016*. 749–758.

[201] Eya Znaidi, Lynda Tamine, and Chiraz Latiri. 2016. Aggregating semantic information nuggets for answering clinical queries. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing 2016*. 1041–1047.

[202] Guido Zuccon. 2016. Understandability biased evaluation for information retrieval. In *Proceedings of ECIR 2016*. 280–292.

[203] Guido Zuccon and Bevan Koopman. 2018. Choices in Knowledge-Base Retrieval for consumer health search. In *ECIR'2018*.

[204] Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyere, et al. 2005. UMLF: a unified medical lexicon for French. *International Journal of Medical Informatics* 74, 2-4 (2005), 119–124.