Abu-Salih *et al. Journal of Big Data* (2023) 10:81
https://doi.org/10.1186/s40537-023-00774-9

Journal of Big Data

RESEARCH Open Access

Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities

Bilal Abu-Salih[1*], Muhammad AL-Qurishi[2], Mohammed Alweshah[3], Mohammad AL-Smadi[4,5], Reem Alfayez[1] and Heba Saadeh[1]

*Correspondence:
Bilal Abu-Salih
b.abusalih@ju.edu.jo
[1]The University of Jordan, Amman, Jordan
[2]King Saud University, Riyadh, Saudi Arabia
[3]Al-Balqa Applied University, Salt, Jordan
[4]Jordan University of Science and Technology, Irbid, Jordan
[5]Qatar University, Doha, Qatar

Abstract
The incorporation of data analytics in the healthcare industry has made significant progress, driven by the demand for efficient and effective big data analytics solutions. Knowledge graphs (KGs) have proven utility in this arena and are rooted in a number of healthcare applications to furnish better data representation and knowledge inference. However, in conjunction with a lack of a representative KG construction taxonomy, several existing approaches in this designated domain are inadequate and inferior. This paper is the first to provide a comprehensive taxonomy and a bird's eye view of healthcare KG construction. Additionally, a thorough examination of the current state-of-the-art techniques drawn from academic works relevant to various healthcare contexts is carried out. These techniques are critically evaluated in terms of methods used for knowledge extraction, types of the knowledge base and sources, and the incorporated evaluation protocols. Finally, several research findings and existing issues in the literature are reported and discussed, opening horizons for future research in this vibrant area.

Keywords Knowledge graph, Knowledge graph construction, Healthcare Knowledge Graph, Drugs, Diseases, Biomedicine, Survey

Introduction

The emergence of big data has opened up new possibilities and ushered in significant changes in various disciplines. Healthcare industry is one of such areas in which advanced and sophisticated data analysis is required to accommodate and properly understand the growing volume of healthcare data, thereby optimising healthcare delivery. However, healthcare data is still regarded as a by-product [1], thus massive healthcare data sources remain neglected and underutilised [1, 2]. Attaining meaningful and actionable knowledge from such data sources could positively affect patient care and enable more accurate diagnosis, prevention of disease, personalised treatment, and better decision-making. Primary obstacles for analysts include heterogeneity of healthcare

data sources and formats, lexical disparities, and the lack of comprehensive and integrated healthcare knowledge libraries [3].

Knowledge Graphs (KGs) have evolved into a new type of knowledge representation that serves as the cornerstone for a variety of applications ranging from general to specialised industrial use [4, 5]. The fundamentally abstract structure of this technology, which efficiently promotes domain conceptualisation and data management, is one of the key factors driving the growing interest in it. The KG, in particular, displays an integrated collection of real-world entities linked by semantically associated relationships. In this case, data annotation put the available semantic content in a machine-readable format, minimising ambiguity and generating relevant information particular to the domain of an application. KGs can furnish an efficient and effective technical solution to conceptualise a healthcare domain and thus be used for several downstream tasks. Therefore, incorporating this technology into healthcare data analytics has emerged as a solution capable to mitigate such issues as data island's complexity, heterogeneity, and sheer size. However, constructing healthcare KGs with unproven methodologies raises concerns regarding their quality and robustness and whether sufficient assessment measures have been applied, especially for KGs obtained from unstructured data sources (such as scientific medical literature or social media). Furthermore, the dynamic nature of healthcare data is strongly linked to context, and numerous facts that characterise clinical and medical entities may vary or change over time. Disregarding the flexibility of knowledge lowers the quality and accuracy of facts embedded in the KGs, thereby leading to substandard decision-making based only on such data sources. As a result, it is critical to perform a detailed analysis of current state-of-the-art methodologies for healthcare KG creation in order to identify such difficulties and open new avenues for pursuing potential solutions.

This survey offers a bird's eye view of the current construction techniques and possible applications of KG technology in the healthcare domain. First, a taxonomy of healthcare KG construction is formulated to illustrate the scope of usage of KG in healthcare, levels of knowledge extraction, different types of knowledge bases and sources, and existing evaluation procedures. Next, we examined significant state-of-the-art KG generation approaches relevant for critical healthcare applications, including (i) drug discovery, repurposing and adverse reactions; (ii) diseases and disorders; (iii) biomedicine; and (iv) other miscellaneous healthcare applications. These approaches are scrutinized, with a summary created for each domain demonstrating specific KG functionalities, the incorporated knowledge extraction techniques (at both entity and relation levels), type of the knowledge base, the resources needed to construct it, relevant KG statistics, the measurements used to assess the KG construction methodology, and the limitations and shortcomings of each approach. This paper is distinguished from similar works that tend to focus too narrowly on specific healthcare subdomains [6, 7] or generic applications of KG in healthcare [8]. In particular, the following are the key contributions of this paper:

- To the best of our knowledge, this survey is the first to provide a bird's eye view of healthcare KG construction.
- A new representative taxonomy is outlined to facilitate easier KG construction in the healthcare domain.

- An in-depth analysis of state-of-the-art KG construction methodologies is provided, and their main strengths and weaknesses are discussed.
- A summary of the research findings and remaining issues is presented, paving the way for future research.

In Sect. 2, taxonomy of KG construction in healthcare is presented and analyzed from multiple perspectives. Several KG construction approaches relevant for various healthcare domains are reported in Sect. 4. Section 5 summarizes the major flows of the existing techniques, and the observed research gaps, and offers suggestions to overcome them.

Survey methodology

This paper aims to review the recent KG construction approaches for healthcare applications. Thus, we attempt to cover all papers that describe mechanisms for KG construction to benefit the healthcare domain. We focus on articles that were published in the past five years (2018–2022). PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [9] is followed to guide this systematic review. As demonstrated in Fig. 1, around 560 articles were selected in the first stage from various databases including Elsevier, ACM Digital Library, Multidisciplinary Digital Publishing Institute (MDPI), IEEE Xplore digital library, and Google Scholar. The collected articles were all in English and were retrieved using the following keywords used in this query: "Knowledge Graph Construction", "Healthcare", "biomedicine", "medicine", "drug discovery", "drug repurposing", "adverse drug reaction", "disease(s)", "disorder", etc. An additional 83 articles were identified and added to the corpus by reviewing the citations map of the tentative collected set of papers. The first stage resulted in a total of 643 records. Another round of inspection was carried out in the screening stage to eliminate any redundant or irrelevant articles. This was accomplished by examining both the title and the abstract of each paper. In this way, 440 records were excluded in the screening stage as they did not meet the inclusion criteria. In particular, many of the articles discussed approaches for KG embeddings that are applied to existing KGs, thus no construction of new healthcare KGs was proposed. Another array of articles reported KG construction for other domains of knowledge yet indicated "healthcare" as an example of the popularity of KGs to tackle industrial applications. The eligibility phase was then carried out by examining the full text of papers and eliminating the irrelevant ones (102 records). In

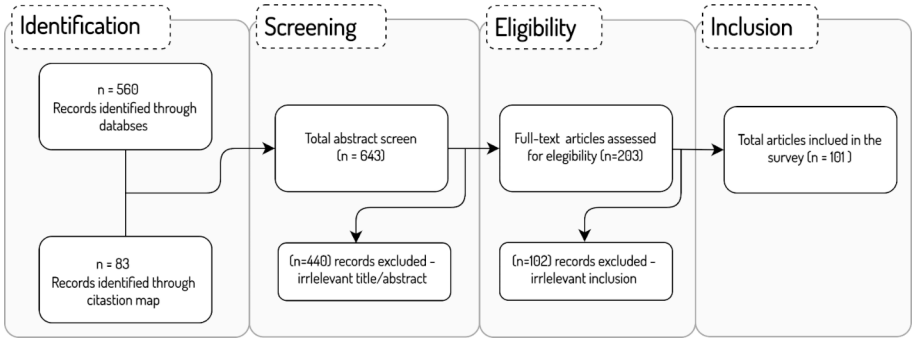

Fig. 1 The article selection strategy for the literature review (PRISMA model)

the final stage, a total of 101 papers were deemed to be qualified to be included in this review.

Figure 2 presents the volume distribution of the selected articles over the past years, clearly showing the growing interest in this technology.

Groundworks

An overview of KG

A KG is a multidimensional graph that contains entities (nodes) and relations (edges) that describe the interrelation of one or more domains. Hence, the KG displays a unified collection of real-world objects connected by semantically relevant relationships. The concept of semantic interlinking is framed by Semantic Web technology whereby data can be annotated in a machine-interpretable format. This is commonly accomplished through the use of ontologies, which define concepts (representing a collection of entities), the relations between entities, and semantic rules, thereby giving a formal and explicit representation of that domain's knowledge [10, 11]. These efforts are fostered by using KGs, an abstract data model that captures a single standard representation of semantically related data (i.e., a graph).

A KG is a directed graph (G), where $G = (V, E)$. This notation depicts the relationship between entities, as well as the interactions between these entities, in terms of graph vertices (V) and edges (E) connecting these vertices. The edges reflect relationships between real-world things whereas the vertices represent real-world entities. The edges of the graph connect the vertices/entities/nodes, and facts can be represented as an RDF[1] triple (*head, relation, tail*), which is also notated as <h,r,t>. As a result, a fact can be inferred by the relationship that connects two interrelated entities. Figure 3 demonstrates a sample KG demonstrating the semantic representation of entities captured from different interrelated healthcare domains, namely *Disease, Gene, Drug, and Compound*. The figure shows how a KG can be used to expand one domain by semantically interlinking it with another domain. Also, various facts can be inferred from the abstract

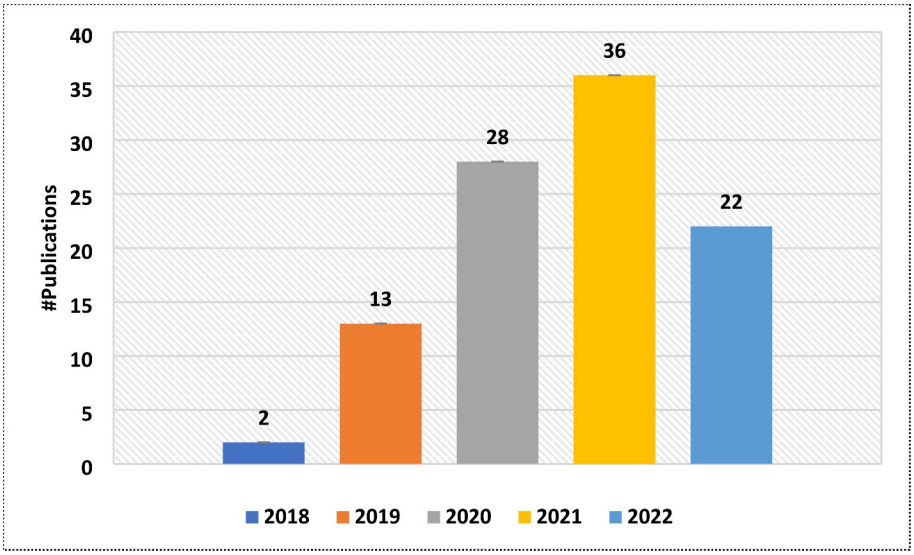

Fig. 2 #Publications about KG construction for healthcare in the past years

[1] https://www.w3.org/TR/rdf11-concepts/.

Fig. 3 A sample healthcare KG

structure of the KG. For example, the fact "*a Disease is associated with a Gene*" represents an abstract fact that comprises two abstract concepts (i.e. *Disease* and *Gene*), and the relation "*is associated with*" builds the triple <" *Disease", "associatedWith",*" *Gene*">. These abstract concepts can be then replaced with real-life entities to provide a specific domain representation. For example, the triple <" *Sjogren's Syndrome", "associatedWith", "HLA-DR3*"> indicates a fact about the Sjogren's Syndrome disorder which can be associated with HLA genes, namely HLA-DR3 [12].

The sample KG depicted in Fig. 3 can be further expanded and linked with other datasets and vocabularies to extend the understanding of these real-world entities which belong to one or different domains.

Generic and domain specific KG

There are two types of KGs: generic and domain-specific KGs. Since the Semantic Web's inception, generic KGs (also called domain-independent, cross-domain, or open-world) have been constantly expanded. As a natural representation of interconnected entities, generic KGs have been related to linked data [13]. Cyc[2], BabelNet[3], NELL[4], CliGraph[5],

[2] https://www.cyc.com/.

[3] https://babelnet.org/.

[4] http://rtw.ml.cmu.edu/rtw/kbbrowser/.

[5] http://caligraph.org/ontology/Scientist.

YAGO[6], and DBPedia[7] knowledge bases are examples of generic KGs, and the number of such KGs is rising rapidly. Domain-specific KGs are defined as "an explicit conceptualisation to a high-level subject-matter domain and its specific subdomains represented in terms of semantically interrelated entities and relations" [14]. These KGs are important to conceptualise specific domains, such as health, sports, social science, engineering, travel, etc. Examples of domain KGs include: HKGB [15], K12EduKG [16], SoftwareKG [17], ClaimsKG [18].

A taxonomy of healthcare KG construction

To better understand the overall paradigm of healthcare KG construction, we design a taxonomy that illustrates key activities and aspects of this process. Figure 4 shows the schematic representation of the taxonomy that was designed after careful examination of all significant state-of-the-art KG creation approaches relevant to critical healthcare applications, including (i) drug discovery, repurposing and adverse reaction; (ii) diseases and disorders; (iii) biomedicine; and (iv) other miscellaneous healthcare applications. This taxonomy aims to ensure that the process of constructing a typical KG in healthcare must demonstrate the intended primary use of KG, levels of knowledge extraction

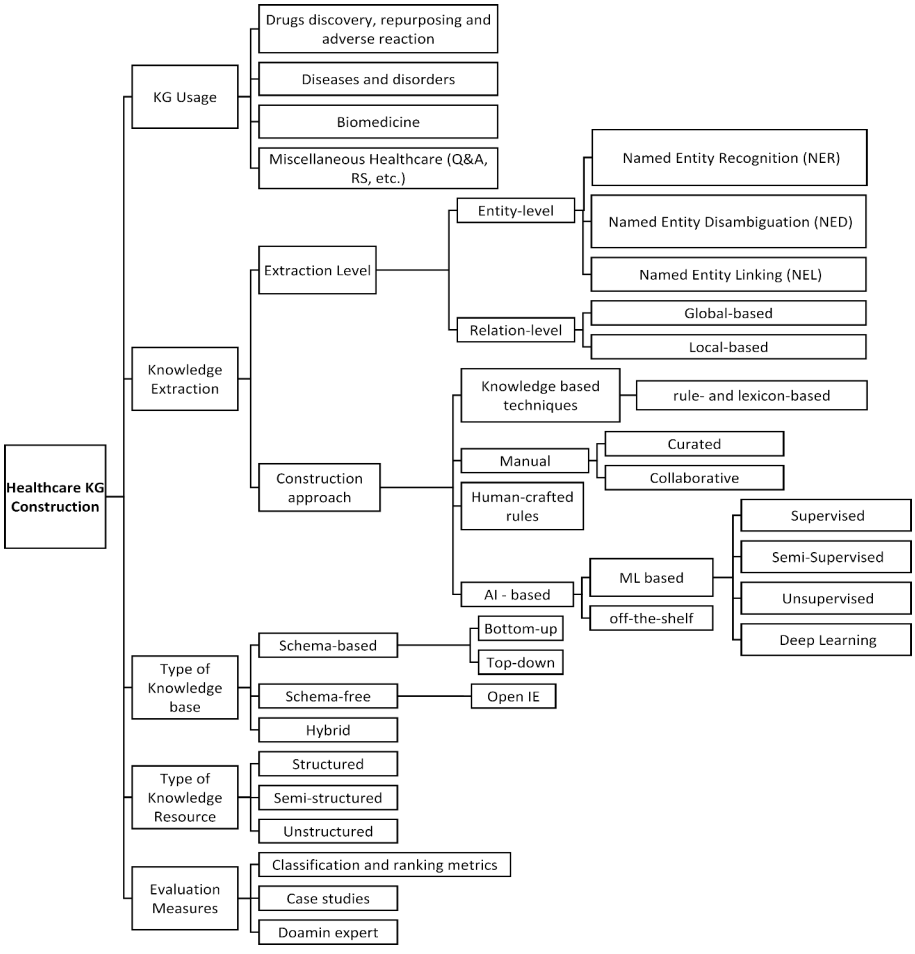

Fig. 4 A taxonomy of healthcare KG construction

[6] http://www.foaf-project.org/.

[7] https://wiki.dbpedia.org/.

(entity level and relation level), different types of knowledge bases and sources, and evaluation metrics and criteria. The following sections provide detailed descriptions of each of the aforementioned elements.

Levels of knowledge extraction

The mechanism used to build a typical healthcare KG includes extracting entities and relations that can be captured from various heterogeneous healthcare data sources using a range of extraction methods. This section discusses the knowledge extraction procedures at both the entity level and the relation level.

Entity-level

Entities in healthcare KGs represent the nodes of the graph, which correspond to real-world entities such as drugs, diseases, diagnoses, patients, hospitals, events, etc. There are three main approaches used for entity extraction [19, 20]; (i) Named Entity Recognition (NER); (ii) Named Entity Disambiguation (NED); and (iii) Named Entity Linking (NEL). NER techniques aim to analyse textual data, thereby identifying factual names of various real-world objects. For example, the "*Pfizer*" entity in the following text snippet "*Clinical trials showed that Pfizer is effective.*" refers to the name of BioNTech vaccine that protects against COVID-19. The techniques used in NER can be classified into (a) knowledge-based techniques that rely on domain-specific knowledge and (b) advanced machine learning techniques that benefit from annotated data (in case of supervised learning), or partially annotated data (in case of semi-supervised learning), or derive knowledge from the structural or distributed nature of data (in case of unsupervised learning) to carry out an entity recognition task. Examples of ML-based techniques include Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Fields (CRF) and variations, and Decision Trees [21–23].

Although NER techniques can identify potential entities, some of these units can be difficult to link to their corresponding entities that are located in the same or different KGs. For example, the expression "*Pink eye*" captured from any textual snippet could possibly refer to *conjunctivitis* and thus should be linked to a corresponding entity in a medical KG; or could simply refer to a cosmetic makeup term (*eyeshadow*) that relates to a completely different domain. The spectrum of currently used techniques in named entity disambiguation spans from rule-based approaches to advanced machine learning approaches, serving to clarify the results of NER and separate similar cases. Finally, NEL aims to link an identified entity (using a NER method) with an unambiguous manifestation (using a NED method) of the same entity captured from textual content, and frame it within a fixed context by linking it to a KG. As a result, NEL is the process of locating an entity mentioned in an (unstructured) text and linking it to a (structured) KG entry. The reader can refer to [19, 20] for detailed discussions on entity extraction mechanisms and technical issues related to their practical implementation.

Relation-level

A relation between two entities conveys the semantic relationship between these entities. Extracting the relations between entities in KG requires such links to be identified, thus establishing a tuple that connects two potential entities. The aim of relation extraction is to figure out in which ways the identified and disambiguated entities are related

semantically. This operation can be performed using either a local or a global strategy. The former denotes a mention-level relationship that is frequently inferred from short textual contents, while the latter seeks to infer relationships that span multiple knowledge bases and may involve numerous local relationships. Further information on relation extraction methods can be found in the related literature [24, 25].

Types of knowledge base

The course of construction of a healthcare KG is dependent on whether a predetermined ontology schema is used (schema-based), no predefined schema is used (schema-free) [26], or a combination of schema-based and schema-free techniques is employed. Based on the selection of data sources and ontology [27, 28], the first class of methods (schema-based) can be divided into two groups: (i) the bottom-up methods, in which the structural framework of an ontology is used as a foundation to construct the KG (e.g. Wikipedia is established by using the predefined ontology model, i.e. DBpedia [29]); and (ii) the top-down technique (e.g., YAGO) [30]), in which the ontology schema is inferred from the underlying structured data, or the taxonomies (hierarchy) which are developed based on information on the Web [31]. Schema-free methods are generally based on open information extraction strategies that rely on the open access to information on the Internet; as a result, data is gathered with diverse knowledge extraction techniques without particular concern for fitting the data into a unifying ontology design (e.g. OpenIE [32]). Hybrid knowledge-based approaches: are flexible strategies for obtaining knowledge that partially rely on a specified ontology but integrate new information in a flexible way (e.g. KnowledgeVault [27], NELL [33]).

Types of knowledge resources

Building a consolidated healthcare KGs requires extracting and integrating data from a variety of sources. The integration step is necessary in order to harmonise the data and provide a consistent big-picture view. There are three types of healthcare knowledge resources; (i) unstructured data sources (such as EMRs, medical literature, discharge summaries, and radiology reports); (ii) semi-structured or tree-structured data sources like JSONs and XMLs (e.g., Bio2RDF[8]); and (iii) structured databases that organize information in tabular formats such as relational medical databases (e.g., MEDLINE[9]).

KG evaluation metrics

The sudden growth of demand for healthcare KGs and the corresponding rush to produce them raises concerns about the quality of embedded information (i.e., entities and relations) and whether these elements accurately transmit the intended real-world facts behind the numbers. Assessing the completeness and veracity of information contained within a KG is the key to determining its "fitness of purpose" [34] for various downstream applications, as well as ascertaining data quality [35–38].

The lack of a complete and accurate KG in a particular domain makes the evaluation process difficult. This is due to the fact that compiling all factual data regarding a particular topic is a massive undertaking that may never be actually finished. As a result, several attempts have been made to augment and dynamically updateknowledge graphs

[8] https://bio2rdf.org/.

[9] https://www.nlm.nih.gov/medline/medline_overview.html.

with new facts derived from new entities and/or relations, usually referred to as KG Augmentation/Completion approaches. These efforts are subjected to correctness and completeness evaluation procedures to assure data quality. The evaluation can be performed by tracking classification accuracy and ranking metrics such as Hits@N and Mean Reciprocal Rank (MRR), Accuracy, Precision, Recall, and F-score [39, 40], based on a comparison between data in the KG and ground truth. These metrics are among a number of tools that can be used to assess the KG's construction quality and factuality of the described entities and relationships. Case studies and domain experts have also been occasionally used in the evaluation of KG structures [41, 42].

State of the art review

Recently, the Healthcare sector has gained much public attention, particularly with the coronavirus (COVID-19) pandemic that started in 2019 and continues to rattle the world. Therefore, there is a notable consensus between industry and academia that it is critical to consolidate the efforts of all stakeholders to overcome the challenges of this vital sector [43]. KGs offer the technical means to the healthcare sector to derive meaningful insights from voluminous and heterogeneous healthcare data contained in clinical and academic sources [44, 45]. The examined papers relevant to healthcare are classified into four different categories: (1) Drugs: This category comprises studies that incorporate KG technology for drug discovery, drug repurposing, and adverse drug reactions. (2) Diseases and disorders: Which includes studies that benefited from KG technology to conceptualise various diseases and conditions, such as stroke, subarachnoid haemorrhage, hepatitis, etc. Also, it includes papers about mental illnesses, such as depression, anxiety, autism, etc. (3) Biomedical studies: These include the fields of biomedicine, microbiology, etc. (4) Miscellaneous healthcare: These are works that span different categories, or those that incorporate KGs to model a specific healthcare solution.

Drug discovery, repurposing and adverse reaction

Drug discovery KGs are receiving a lot of attention from researchers who are involved in the drug development studies. The necessity to construct specific KG for the drug sector has several key motivating factors [46]; prescribing a particular drug to treat a certain disease might involve some non-medical factors including the demographics, insurance policy, drug availability, etc. Further, in some instances, healthcare professionals who are not qualified to prescribe drugs, might act upon an emergency, thereby initiating a treatment that in a normal situation has to be initiated by a specialist doctor. Such complications illustrate the need for an intelligent platform that can actively guide the search for the optimal drug to prescribe. In this context, Mann et al. [46] attempted to create such a platform that can assist in finding a valid treatment considering the known symptoms or identified disease. In this study, the authors integrated existing medical knowledge resources, thereby building a KG to benefit the entire domain. In the same line of research, Che et al. [47] proposed a method to integrate six knowledge bases into one coherent KG. The resulting KG is then embedded into Graph Convolutional Network with an Attention mechanism for Drug–Disease Interaction (DDI), which is used to predict and discover potential drugs capable of effectively treating COVID-19. The prediction of Drug-Target Interaction and Drug-Drug Interaction are important aspects of the development of new drugs. In another interesting study, Zhang et al. [48] constructed two designated KGs

describing drugs captured from a biological dataset, namely Bio2RDF[10]. This is followed by developing a learning model based on graph representation (MHRW2Vec), whose output was fed to a neural network model (TextCNN-BiLSTM Attention Network (TBAN)). The ultimate objective was to predict potential interactions of various COVID-19 drugs. Ye et al. [49] developed KGE_NFM, an integrated framework comprising both a KG and a recommender system to predict DDI. The components of the KG were embedded in a low-dimensional space, after which a neural factorization machine was tasked to build the recommender system for drug target discovery. Drug discovery incorporating KG technology was also discussed in [50–53].

Drug repurposing Drug repurposing (a.k.a. reprofiling, redirecting, rediscovery, or repositioning) is an interesting domain that has come into focus recently. It aims to reuse existing drugs to treat emerging diseases (such as COVID-19) thereby reducing both drug development timelines and the associated costs [54]. Therefore, various studies attempted to provide intelligent solutions for the challenges inherent in drug repurposing. Regarding the use of KGs, there is a direction of research aimed at constructing KGs that can be used for drug repurposing. BenevolentAI's proprietary KG [55] is amongst the most successful approaches in this research line. The BenevolentAI KG integrates an assortment of medical data obtained from structured and unstructured scientific repositories (including literature). It is queried by various algorithms to identify new relationships between entries, thereby suggesting new ways of treating diseases. In the same context, Wang et al., [56] proposed a framework called COVID-KG which aimed to construct a KG from multimodal data found in scientific literature into one actionable KG that can be used for drug repurposing. The proposed KG is built on an ontology of 77 entity subtypes and 58 event subtypes, as defined in the Comparative Toxicogenomic Database (CTD) (Davis et al., 2016), and entities are linked using Medical Subject Headings (MeSH) framework [57]. also proposed a multimodal drug repurposing KG for COVID-19 that was built with data harvested from scientific literature, and aimed to provide an overview of pathophysiology related to COVID-19. The construction of this graph was carried out manually using Biological Expression Language, and evaluated based on multiple case studies. Drug repurposing is further discussed in [58–63].

Adverse drug reactions (ADRs) ADRs refer to undesired reactions that occur after the use of a certain medical product [64]. ADRs carry significant risks to both patients and the hospital system [65], thus serious attention is required to tackle this issue and develop optimal technological solutions to mitigate it. The sophisticated structure of KGs presents an opportunity to define this problem conceptually and provides new ways to predict potential ADRs. Many studies were conducted in this direction, notably Bean et al. [66] benefited from access to two drug resources (namely DrugBank[11] and SIDER[12]) to build a KG that can predict ADRs. This KG contains four types of nodes and three types of edges, and is consolidated with a prediction model (similar to linear regression). Authors of [67] introduced a KG to represent drugs and ADRs, with data embedded using the Word2Vec model. On top of this model, logistic regression was used to predict whether a given drug

[10] https://bio2rdf.org/.

[11] https://go.drugbank.com/.

[12] http://sideeffects.embl.de/.

causes any ADRs. Tumor-Biomarker Knowledge Graph (TBKM) [68] is another attempt to design a KG with four node classes (namely Tumor, Biomarker, Drug, and ADR) based on data from scientific biomedical literature. The aim of the KG is to discover ADRs of antitumor drugs as well as provide explanations why they occur. Predicting and discovering ADRs have been further reported in [69–74]. Zhao et al. [75] designed their drug action mechanism KG after extracting information from 770,000 abstracts of medical papers. Despite the poor approach used to extract entities and their relationships, the paper managed to cover a large number of drugs and mechanisms of action. Table 1 illustrates a summary of currently proposed KG construction approaches for drug discovery, drug repurposing, and adverse drug reaction.

Diseases and disorders

Topographic and anatomic KGs have accelerated the pace of scientific discovery that aims to better understand diseases affecting the human body. For example, Zhang et al. [15] developed Health Knowledge Graph Builder (HKGB), which is a framework that can be used to construct a Health KG (HuadingKG) from multiple sources (namely EMRs, medical standards, and expert knowledge) to be used in the cardiovascular domain. To conceptualise subarachnoid haemorrhage stroke, the authors of [45] developed a comprehensive framework that allowed them to construct a KG from heterogeneous data automatically. In particular, the authors incorporated semantic analysis for entity and relation extraction, and implemented a knowledge prediction model based on the association rule and ensemble machine learning. KGHC [76] is a KG designed specifically for Hepatocellular Carcinoma. It brings together and connects entities captured from 5 different unstructured and structured data sources and extracted using information extraction techniques such as BioIE and SemRep. Yin et al. [77] constructed a KG for diagnosing and treating viral hepatitis B by adopting a top-down approach where a domain ontology was used to build the KG. The authors did not provide adequate details on the mechanism utilised to construct the KG or the evaluation metrics. Yet, they claimed that the designed KG benefits intelligent recommender systems that can be used to diagnose and treat viral hepatitis B. Another research direction identified the role of genes in human disease [78]. The authors built a convolutional neural network-based model on top of a biological KG to classify the genes highly correlated with cancer. While the construction of the KG itself was not adequately validated, the resultant embedding model was evaluated on downstream tasks. An attempt at preventing Myopia using KG technology was described in [79]. The authors developed a KG from various Chinese websites to provide intelligent Q&A services to users interested in Myopia prevention. However, the finalised KG lacks multimodal data that can be captured from medical databases and domain-relevant question answering systems. Conceptualising Stroke and its causes and effects is an extensively covered subject in the literature. For example, Yang et al. [80] constructed an integrated KG, named StrokeKG, that portrays various stroke-relevant relationships inferred from various medical datasets. Designing KGs to benefit the victims of stroke was also examined in [81, 82]. COVID19-related disease discovery using KGs was reported by Huang et al. [83]. Relying on a pipeline approach, the authors surveyed from relevant scientific papers related to COVID-19 and used the collected data to construct a KG that can identify diseases and drugs associated with COVID-19. The accuracy of the extracted knowledge was then verified using the time-slicing method. The use of KGs in

the healthcare domain was discussed with a focus on disease identification and prediction in [84–87], detecting the association between miRNA and disease in [88], chronic disease management in [89], and syndromes diagnosis in [90].

Mental disorders Yuan et al. [91] constructed a KG with minimal supervision to frame autism spectrum disorder diseases, using the articles obtained from the PubMed dataset. Entities were extracted using MinHash lookup/ UMLS [92] and formed into pairs which were then clustered using kmeans++based on similarity between entities. Constructing KGs that can describe depression was undertaken by Huang et al. [93]. In particular, they attempted to generate a sub-graph that describes depression disorder, obtained by parsing data from a variety of major knowledge sources such as PubMed, Medical Guidelines, DrugBank, Unified Medical Language System (UMLS) etc. In the same line if research, Li et al. [94] proposed a UMLS-based semantic prediction program, known as SemRep, as well as SemMedDB to construct a KG for describing depression by using a bottom-up approach. Depression and its association with metabolism is also discussed in [95]. The authors developed MDepressionKG KG that integrates data about human microbial metabolism network, human diseases, microbes, etc., to offer semantic-based rational reasoning and establishing probable relations between depression and comorbid diseases. Although the authors furnish a useful online website to demonstrate utility of MDepressionKG, the knowledge inference mechanism is ineffective due to the incorporated traditional rules of logic. Furthermore, automatic extraction methods are required to enrich the functional diversity of the proposed depression KG. The conceptualisation of various mental disorders through graphs was also presented in [96–99]. Table 2 shows a summary of KG construction approaches for diseases and disorders.

Biomedicine

Generic biomedicine KG's have been used with success for modelling both biological systems and pathologies, providing the means to understand this interplay between them. Several studies reported significant advances in this direction while incorporating KG technology. PharmKG [100] is a comprehensive KG built upon integrating six interrelated knowledge bases, with nodes representing genes, chemical compounds, and diseases. Entities of PharmKG are labeled with domain-specific information, keeping semantic and biomedical characteristics of the data. Percha et al. [101] compiled a basic KG known as the Global Network of Biomedical Relationships (GNBR) from biomedical literature. The process of populating GNBR with data was performed using PubTato (named entity annotator) tool, as well as Ensemble Biclustering for Classification (EBC) algorithm to annotate entities captured from Medline abstract. Wood et al. [102] developed RTX-KG2, an integrated KG that includes biomedical data captured from 70 biomedical knowledge bases. The aim of RTX-KG2 is to offer an open-source KG that can be used as a biomedical translational reasoning engine. Zhang et al. [103] reported the extraction of biomedical causality from the scientific literature. In their work, the authors constructed a biomedical knowledge graph to discover causal relationships in the biomedicine field. Authors of [82] developed a marine Chinese medicine KG using a top-down approach that takes guidance from a domain ontology. Developing an integrated KG to benefit the biomedical domain has also been discussed in [104]. The authors presented BioKG, a KG of drug-drug and drug-protein interactions data collected and compiled using modu-

Table 1 A Summary of KG construction approaches for drug discovery, drug repurposing, and adverse drug reaction

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[46]	Drug discovery	Manual and fuzzy matching		Schema-based	Wikidata, DrugBank[14], WedMD, and GoodRx	N/A	R, P	• Lack of statistics on the resultant KG. • Limited discussion on the Ontology design • The evaluation of the proposed model emphasized on KG embedding rather than the resultant integrated KG.
[47]	Drug discovery for COVID-19	Manual construction based on six KGs obtained from the literature		Schema-based	Literature on COVID-19	#n: 100,00 #e: 670,000	AUC, and AUPRC	• Insufficient discussion on the mechanism followed to integrate the incorporated KGs, • The evaluation of Att-GCN-DDI is limited and not detailed.
[48]	Drug discovery	Manual extraction based on Bio2RDF KG		Hybrid	Bio2RDF[15]	#n: 2,947,140 #e: 10,131,654	AUC, AUPR, F1	• Inadequate discussion on the construction of drug KG.
[55]	Drug repurposing	Algorithms developed at BenevolentAI[16] and part of their IP		Hybrid	Structured and unstructured resourced including Literature on COVID-19	#n: millions #e: hundreds of millions	Case study	• There is no detailed discussion on the mechanism followed to construct BenevolentAI graph. • The evaluation was merely measured by case study.
[56]	Drug repurposing	Coarse- and fine-grained entity extraction	Manually based on CTD and MeSH	Schema-based	Multimodal scientific literature (CTD[17])	#n: 67,217 #e: 77,844,574	Case study on Drug Repurposing Report Generation	• Although the proposed framework demonstrated success in tackling the quantity issue of relevant KG resources, the quality issue was not properly evaluated to demonstrate its effectiveness. • Observed bias in training and development data, source, and test queries.

[14] https://go.drugbank.com/.

[15] https://bio2rdf.org/.

[16] https://www.benevolent.com/.

[17] http://ctdbase.org/downloads/.

Table 1 (continued)

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[57]	Drug repurposing	Manually encoded in Biological Expression Language		Schema-free	PubMed, LitCovid[18], EuropePMC, etc.	#n: 4,016 #e: 10,232	Case study (Gene Expression Analysis)	• The mechanism followed to construct the KG (manual-based) is poor in terms of scalability.
[58]	Drug repurposing	Cross-referencing		Schema-based	PharmGKB, TTD, KEGG DRUG, DrugBank, SIDER[19], and DID	N/A	Case study (Finding drug–disease pairs)	• The proposed data model that was used for data integration can be improved by using formal domain ontology toward better conceptualizing the domain.
[67]	Prediction of adverse drug reactions	Direct construction from structural databases		Schema-free	DrugBank database and SIDER database	#n: 12,473 #e:154,239	P, R, F1, AUC, and a case study on Drug-induced liver injury	• The KG skips information of drugs and protein target, • The scope of information perceived by entities can be enlarged by using longer path in the KG as the input of Word2Vec model.
[66]	Prediction of adverse drug reactions	Direct construction from structural databases		Schema-free	DrugBank, SIDER	#n: 5,828 #e: 70,382	AUC and case study(Validation in EHRs and Eudravigilance)	• No clear discussion on KG construction approach, • Insufficient discussion on the methodology followed in the ML benchmark comparison.
[68]	Discovery of adverse drug reactions	cTAKES[20]	naive Bayesian model	Schema-based	MEDLINE	#n: 9,699 #e: 139,254	co-occurrence analysis and Case study (Osimertinib)	• The computed drug-biomarker groupings cannot differentiate between a drug-treatment relationship, • The study lacks the attention to drug-drug interaction, • lack of rationale on using the entity extraction method
[75]	Drug action	Automatically using rule-based approach		Schema-free	Medical papers	#n: 40,963 #e: 57,865	R, and accuracy	• Lack of verification to the textual prio KG construction. • Limited comparison with currently exiting similar KGs.

[18] https://www.ncbi.nlm.nih.gov/research/coronavirus/.

[19] http://sideeffects.embl.de/.

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

Table 2 A Summary of KG construction approaches for diseases and disorders

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[15]	Cardiovascular domain	LSTM-CR	pattern-based and supervised learning methods	Hybrid	UMLS, EMRs, medical standards, and expert knowledge.	#n: 8,293,284 #e: 32,256,360	The evaluation is conducted in the embedded modules	• The overall framework requires a detailed case study to evaluate the effectiveness of integrating the proposed modules.
[45]	Subarachnoid hemorrhage	Semantic analysis (Ontologies: LBO, IAO, etc.,)	Automatic (Rule-based)	Shema-based	clinical notes and brain angiograms	N/A	P, R, F1, and AC	• Limited discussion on the KG statistics • The overall framework requires a detailed case study to evaluate the effectiveness of integrating the proposed modules.
[76]	Hepatocellular carcinoma	SemRep[14], rule-based method,and BioIE(with Att-BiLSTM-CRF)		Schema-based	PubMed, SemMedDB, UpToDate, and Clinical Trials[15]	#n: 5,028 #e: 13,296	Accuracy	• The KG was not properly evaluated on real-life case study that addresses hepatocellular carcinoma. • There has been no detailed discussion on the mechanism followed to address the presented disagreements.
[80]	Stroke	DNorm[16], tmChem[17], GNormPlus[18], PWTEES[19]	NLTK, PKDE4J, and Bio-BERT	Shema-free	CID[20], TCMID[21], EU-ADR[22], ETCM[23]	#n: 46 k #e: 157 k	P, R, F1	• The constructed KG is limited to Chinese context and hard to replicate and build a more comprehensive map of medical knowledge.

[14] https://semrep.nlm.nih.gov/.

[15] https://clinicaltrials.gov/.

[16] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/DNorm/.

[17] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmchem/.

[18] https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/.

[19] https://github.com/chengkun-wu/PWTEES.

[20] http://www.cbs.dtu.dk/services/.

[21] http://bidd.group/TCMID/.

[22] https://biosemantics.erasmusmc.nl/index.php/resources/euadr-corpus.

[23] http://www.tcmip.cn/ETCM/.

Table 2 (continued)

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[77]	Diagnosis and treatment of viral hepatitis B	N/A	N/A	Schema-based	EMR (8544 patients in China)	#n: 8,563 #e: 96,896	N/A	• No proper evaluation was conducted. • No discussion on mechanism followed to construct the KG
[83]	Coronavirus pneumonia-related diseases,	CRF	Bio-BERT	Shema-free	COVID-19 scientific literatures	#n: 10,993 #e: 1,204,234	Specificity, P, R, F1, and AC	• The entity and relation extraction datasets are provided with lack of discussion on the mechanism followed to conduct the experiments on these datasets.
[78]	Identifying disease-gene associations	N/A	N/A	Shema-free	CTD, BioGrid[24], MalaCards[25]	#n: 103,625 #e: 3,273,215	N/A	• No discussion on the mechanism followed to extract entities and relationships. • The construction of the KG itself is not evaluated
[79]	Myopia Prevention	Automatic using python script		Schema-based	Baidu Encyclopedia, Chinese Wikipedia, and professional websites	#n: N/A #e: N/A	NA	• KG is not described in terms of mechanisms used to extract entities and relationships. • No proper evaluation is undertaken.
[93]	Depression disorder	XMedlan, Semantic Queries with regular expressions,		Hybrid	PubMed, Clinical Trials[5] DrugBank[26], DrugBook, Wikipedia, SIDER[27], and UMLS	#e: 8,892,722	Use cases	• Lack of proper evaluation, • insufficient use of other important medical repositories, • lack of discussion on both the methodology used for knowledge integration and KG statistics.
[91]	Autism spectrum disorder	MinHash lookup/UMLS	Skip-gram and kmeans++	Schema-free	PubMed[28] (autism spectrum disorder-related article abstracts)	#n: 6827 #e: 16,192	Hit@k	• Extracted relations are coarse-grained. • Difficult to distinguish semantically related relations, • Insufficient overall evaluation to the model

[24] https://downloads.thebiogrid.org/BioGRID.

[25] https://malacards.org/.

[26] https://www.drugbank.ca/.

[27] http://sideeffects.embl.de/.

[28] https://pubmed.ncbi.nlm.nih.gov/.

Table 2 (continued)

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[94]	Depression	SemRep[29],OpenIE and rule-based method		Schema-based	SemMedDB, PubMed	#n: 3,055 #e: 30	Jaccard	• Poor data quality • The utility of KG was not well-proven
[95]	Metabolism-depression associations	Manual curation and extraction by domain expert (traditional logical rules)		Schema-based	KEGG and scientific literature	#n: 3,724,526 #e: 5,725,821	Case study	• Ineffective inferences due to the incorporated traditional logical rules. • Automatic extraction methods are required to enrich the functional diversity of the depression KG.

[29] https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html.

lar software, namely BioDBLinker. This KG contains managed entities and relationships captured from at least five biomedical databases, such as UniProt, REACTOME, KEGG, DrugBank, SIDER, and Human Protein Atlas (HPA). He et al. [105] designed a KG for intestinal cells. First, the authors built an ontology as a conceptual model followed by extracting facts from the academic literature. Despite the problems with mechanisms used to construct the actual KG, the work presents an important attempt toward constructing KGs specifically to study the intestinal field, facilitating much easier observation of the processes of intestinal cytokines via various signalling channels. Constructing KGs to benefit generic biomedical domain was the subject of [106–111].

Microbiology KGs offer an excellent mechanism to conceptualise our understanding of microscopic organisms and their ecological traits. To this end, Joachimiak et al. [112] developed KG-Microbe, an integrated KG that contains prokaryotic data for phenotypic traits as well as supporting use cases in microbiology, biomedicine, and environmental science. Liu et al. [113] conceptualised gut microbiota using a semantically enriched KG, namely MiKG. MiKG integrates facts obtained from medical literature as well as other medical knowledge bases, thereby offering an interface for detection of possible connections between gut microbiota, neurotransmitters, and mental disorders. Authors of [114] developed a Microbe-Disease Knowledge Graph (MDKG) through an explorative study, thus identifying the associations between bacteria and diseases. MDKG is populated with entities and relations captured from textual content of Wikipedia as well as other semantic knowledge bases. Modelling Coronavirus using KG technology has recently attracted a lot of attention in the research community. For example, Zhang et al. [115] built a coronavirus KG by integrating entities captured from Analytical Graph and CORD-19 databases. The aim of the proposed KG is to provide a tool for the exploration of coronavirus on the entity level. Another attempt to help the biomedical research community comprehend the coronavirus using KGs is offered by Chen et al. [116]. The authors constructed a designated KG to discover any associated diseases, potentially effective drugs or treatments, and relevant genes and mutations. Using KG technology, modelling Coronavirus relevant information was also implemented and discussed in [57, 117, 118]. Further uses of KGs in microbiology are studied in [119]. Table 3 depicts a summary of KG construction approaches for the biomedical domain.

Miscellaneous healthcare

Constructing KGs from EMRs The ongoing efforts to leverage the proliferation of EMRs for multiple medical applications are well-documented in the scientific literature. Extracting valuable knowledge from such data silos has been made easier by KG technology. In this context, several studies attempted to construct medical KGs that can improve specific areas, for example, clinical decision support systems. One such attempt was undertaken by Li et al. [120], who followed a systematic approach consisting of eight steps to build a medical KG from EMRs obtained during the patients' visits. The authors constructed a quadruplet-based medical KG incorporating an additional item (properties) which includes a set of characteristics to rank the embedded entities. The main objective of this study is to ensure the robustness of facts in the KG related to the medical domain. Evaluating the robustness of a constructed KG in healthcare is of utmost significance to ensure the quality of the inferred knowledge. In this context [121], presented a meth-

odology to measure and evaluate the robustness of knowledge relating to diseases and symptoms, with data captured from existing health knowledge graphs as well as records of patient visits to the Beth Israel Deaconess Medical Center (BIDMC). Postiglione et al. [122] proposed an advanced entity recognition approach named PETER (Pattern-Exploiting Training for Named Entity Recognition), that integrates Pattern-Exploiting Training (PET) [123] to build an Italian-language KG for healthcare. EMRs represent a fertile source of information for healthcare KGs, hence their use for construction of KGs is becoming quite common, as exemplified in [124] and [125].

Query answering (QA) and question and answer (Q&A) Incorporating health KGs into a QA system was discussed by Sahu et al. [126]. The authors proposed a system that can be used to search for various health-based KGs and obtain a set of healthcare-related response sub-graphs. The possibility of using medical KG's to benefit QA applications was also discussed in [127]. Zhao et al. [128] made use of EMRs obtained from hospital patient records in Shanghai to build a medical knowledge graph based on the BILSTM-CRF model. Here, a KG is used as a part of a QA system to provide support for establishing medical diagnosis. Xie et al. [129] attempted to create a KG for Traditional Chinese Medicine (TCM), yet the KG they ended up with is very limited in terms of entities and relationship; thus, the applicability and utility of the graph is questionable. Another Chinese medical KG was proposed by [130]. The authors developed this KG from various structured, semi-structured, and unstructured resources and built a QA system that was not adequately validated due to irrelevant results. Also, Huang et al. [131] designed a QA system based on a constructed KG, with the information in the graph used to identify the question's intention. Deploying QA and Q&A systems based on medical and healthcare KGs is also a relevant topic in [132].

Healthcare Management: In the literature, it has been frequently suggested that a KG can be built to help with health management and to better address the most critical health-related issues and chronic disorders [133–136]. For example, Huang et al. [133], proposed a KG building approach that aids users who are seeking information about a healthy diet. Domain ontology was presented by the authors as the basic structure of a KG containing information about diet. Conditional Random Fields (CRF), Support Vector Machine (SVM), and Decision Tree (DT) methods were used to enrich the KG with entities harvested from a variety of healthcare websites. Haussmann et al. [134] developed an integrated KG (FoodKG) that brings together information about healthy food, recipes, and nutritional value. The authors used the RDF Nano publication to establish the reliability of their findings [137]. Chi et al. [135] developed an inclusive healthy diet KG by following a similar study path. In this case, the KG was comprised of five essential concepts: the meal, the dish, the nutritional aspect, the symptom, and the crowd. The proposed model was able to collect and import entities from a range of web resources and deployed multiple NLP and machine learning methods with a semi-automated extraction strategy. In addition, food domain-specific KGs were modelled in [138–140]. Another example of the use of KG-based technology to address difficulties in healthcare systems was discussed in [141–143].

Miscellaneous KGs In healthcare, addressing the timing factor in KG creation is critical. Ma et al. [144] developed a temporal KG that is useful for studying episodic memory

Table 3 A summary of KG construction approaches for the biomedical domain

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[100]	Generic biomedicine	Manual integration and mapping of entities and relationships		Schema- base	OMIM, DrugBank, PharmGKB, Therapeutic Target Database], SIDER, and HumanNet	#n: 7,603 #e: 500,958	Hits@N and Downstream tasks	• The quality and integrity of the metadata cannot be fully assured. • The final version of the constructed graph does not have large-scale of entities compared with state-of-the-art KGs. • No discussion is provided on the adopted ontology.
[101]	Generic biomedicine	PubTator[14] and manual annotation (EBC)	Stanford Dependency Parser[15]	Schema-free	Biomedical literature (Medline abstracts[16])	#n: N/A #e: 2,236,307	Benchmark comparison	• Heavily dependent on the co-occurrence of paths to map scarcer paths to themes, • Lack of handling complex relations • There is a potential of a parser error,
[102]	Translational biomedicine	Manually and automatically using Snakemake[17]		Schema- base	70 knowledge sources including SemMedDB, ChEMBL, etc.	#n: 6.4 m #e: 39.3 m	Benchmark comparison	• The automation process to construct the KG was not detailed. • The comparison with other KGs is not well discussed nor formulated.
[103]	Biomedical Causal Discovery	Manual and rule-based approach		Schema-free	PubMed	#n: N/A #e: N/A	Accuracy	The paper failed to extract implicit causality, The process to identify concepts and relationships between concepts is not detailed.
[82]	Marine Chinese medicine	Manual mapping between the ontology and the KG		Schema- base	Medical literature	#n: N/A #e: N/A	NA	• The paper inadequately described the construction and evaluation of the proposed KG.

[14] https://www.ncbi.nlm.nih.gov/research/pubtator/.

[15] https://nlp.stanford.edu/software/lex-parser.shtml.

[16] https://www.nlm.nih.gov/bsd/pmresources.html.

[17] https://snakemake.readthedocs.io/en/stable/.

Table 3 (continued)

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[104]	Generic biomedicine	BioDBLinker	Automatic mapping	Schema- free	UniProt[18], REACTOME[19], KEGG[20],DrugBank, SIDER, and d Human Protein Atlas (HPA)[21].	#n: N/A #e: N/A	Benchmark comparison	• Suffers from sparsity of data, • Train-test data leakage in case used without careful review
[105]	Intestinal cells	Manually based on the conceptual model		Schema- base	PubMed	#n: 2443 #e: 160,253	Case study	• Poor entity and relation extraction approaches. • Data source is static and limited to medical literature, yet medical facts of intestinal cells can be obtained from future experiments.
[112]	Microbiology	NER and NLP techniques		Schema- base	KG Hub – COVID19[22]	#n: 266,000 #e: 432,000	N/A	• Poor discussion on mechanisms followed to construct and validate the KG
[113]	Gut microbiota	Manual annotation and mapping		Schema- base	Google Scholar and PubMed, UMLS, MeSH, SNOMED CT, and KEGG	#f: 31,268,998	Case studies	• Poor extraction of entities and relations. • The correctness and completeness of extracted relations limit the semantic search's precision and reliability.
[114]	Microbe-Disease Associations	Kindred entity and relation classifier[23]		Schema- free	Wikidata, UMLS, NCBI	#n: 9,832 #e: 21,905	Hits@N	• KG can be expanded by means of a bacterial attribute mining tool, • Lacks a discussion on interactions between bacteria and antibiotics or viruses.
[115]	Coronavirus	Manual extraction and mapping		Schema- free	Analytical Graph (AG) and CORD-19[24]	#n: 588,820 #e: N/A	Case study	• Limited data sources, • Static KG

[18] https://www.uniprot.org/.

[19] https://reactome.org/.

[20] https://www.genome.jp/kegg/.

[21] https://www.proteinatlas.org/.

[22] https://github.com/Knowledge-Graph-Hub/kg-covid-19.

[23] https://kindred.stanford.edu/.

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

Table 3 (continued)

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[116]	Coronavirus	BioBERT		Schema- free	PubMed and CORD-19	#n: N/A #e: N/A	P, R, and F1-score	• KG can be expanded to other bio-medical datasets. • Further biomedical NLP models for NER, e.g., blueBERT can be attempted to verify the validy of the extracted knowledge.

in cognitive tasks. The Integrated Conflict Early Warning System (ICEWS) dataset and the Global Database of Events, Language, and Tone were used to create this temporal KG (GDELT). Their work was unique in that it involved four substantial static KGs embedding data to four-dimensional temporal/episodic KGs, which set them apart from other efforts in this direction. Two new RESCAL generalisations were also proposed and considered. Another important effort that integrated plausible reasoning with fine-grained biomedical ontologies to tackle the data incompleteness problem was undertaken by Mohammadhassanzadeh et al. [42]. The authors proposed a Semantics-based Data analytics (SeDan) framework that performs an exploratory analysis of the KG using the OWL extension and query rewriting algorithm. The framework incorporates data from various knowledge bases, including the DrugBank, Disease Ontology, and the large-scale semantic MEDLINE database (SemMedDB). Rastogi et al. [145] framed their personal health KG as a combination of context, personalization, and integration with other knowledge bases. Their study indicated that the literature on personalised health-related KGs is incomplete and lacks a unified standard representation to adequately describe the designated domain. To provide an overview of effective medications, side effects, and target populations relevant to COVID-19, the authors of [146] proposed a KG-based framework to support COVID-19 clinical research. This framework benefited from Stanford's Stanza toolbox to extract KG's entities and relationships that can be fed into a visualisation module for querying information. The application of KGs in healthcare and medical domains was documented in other relevant tasks including epidemic contact tracing [147], food waste detection [109], drug similarity [148], clinical decision support systems [120], and medical recommender systems [149, 150]. Table 4 shows a summary of KG construction approaches used in various miscellaneous healthcare applications.

Summary

This paper examines the most recent works related to KG construction methodologies in various healthcare domains, such as drugs (and their applications), diseases and disorders, biomedicine, etc. A closer look into these important domains reveals crucial research areas that benefit substantially from KG technology. This research demonstrates the popularity of using KGs to solve real-world healthcare-related problems and shows how KGs have proven to be an effective overall solution for reducing complexity, ensuring flexibility, and establishing a common-ground architecture where data from various sources can be readily incorporated. It is generally agreed that KG technology allows for semantic integration of data acquired from many sources, which may exist in different formats, and can then be fed into a single, coherent framework to be formally used to conceptualise the designated domain.

Findings, open issues, and opportunities

Despite the popularity of KG technology in the healthcare domain, this study reveals certain limitations that open new directions for future research.

- **KG data sources**: various previous studies have concentrated on knowledge curation and facts captured from a limited number of data sources. For example, certain KGs were constructed using only biomedical scientific publications (e.g. PubMed and SemMedDB) [94, 103, 105]. The extracted knowledge using such data sources

Table 4 A summary of KG construction approaches for miscellaneous healthcare

Ref.	KG Specific Functionality	Knowledge Extraction Techniques		Type of KB	KG Resource(s)	KG Stats	Evaluation Measure(s)	Shortcoming(s)
		Entity-level	Relation-Level					
[120]	A generic medical KG of patient visits.	BMM, BiLSTM-CRF and pattern recognizer	Nine predefined relations	Schema-free	Southwest Hospital in China: 16,217,270 de-identified visits of 3,767,198 patients	#n: 22,508 #e: 579,094	R, P, F1, and NDCG	• KG embedding was designed and limited to Bi-LTSM without considering other state-of-the-art techniques. • The evaluation was mainly conducted on the embedded components. • Besides the preliminary discussion on the applications, there is a lack of an overall evaluation of the KG.
[121]	KG of online EMR and emergency department	N/A	N/A	Schema-free	BIDMC dataset and EMRs from an emergency department	#n: N/A #e: N/A	F1 and the area under the precision-recall curve	• The provided statistics are on the sources of the KG; the stats on the KG in terms of entities and edges are missing. • There is no discussion on the mechanism followed to construct the KG in terms of entities and relations.
[133]	Smart Healthcare Management	CRF	Manual and classification-based algorithms	Schema-based	Chinese healthcare websites[14][15,16]	#n: 1,169 #e: 9,707	R, P, and F1	• The resultant KG can be consolidated with information about disease and drugs and link them with symptom entities.
[128]	Q&A	BILSTM-CRF	Manually	Schema-free	EMRs from a hospital in Shanghai	#n: 44,111 #e: 203,308	R, F1 and Accuracy	• Lack of comparative study of the model. • Limited practicability of the system • Limited size and pretreatment of the corpus
[129]	Q&A	BiLSTM + CRF		Schema-free	National Service Platform for Famous Old Chinese Medicine Experience[17]	#n: N/A #e: N/A	Case study and Hitration	• Poor KG with a minimal number of entities and relationships,
[42]	Q&A	Plausible reasoning		Schema-free	BioASQ, DrugBank, Disease Ontology, and SemMedDB	#n: N/A #e: N/A	Domain expert's verification	• Insufficient evaluation, • evaluating the performance of query rewriting algorithm does not exist
[130]	Q&A	Automatic mapping		Schema-free	Chinese medical websites	#n: 18,687 #e: 88,858	Case study	• Poor discussion on extraction of entities and relationships. • The QA system does not exhibit utility due to inapplicable results.
[131]	Q&A	Jieba[18]	Automatic mapping	Schema-free	A medical company (YiFeng Pharmacy[19])	#n: 34,788 #e: 601,475	Training and decision accuracy, cost, and time	• The construction of KG is not validated. • The system can answer one intention per question and cannot thus answer questions with multi-intensions.
[146]	COVID-19 Clinical Research	Stanza's NER[20]	Stanza's Bi-LSTM	Schema-free	Artificial Intelligence in Medicine	#n: N/A #e: N/A	Baseline comparison	• Lack of statistics on entities and relationships, • Poor KG validation method

[14] https://www.zhys.com/.

[15] http://app.huofar.com/grz/.

[16] http://www.cf555.com/.

[17] https://www.gjmlzy.com/.

[18] https://pypi.org/project/jieba/.

[19] http://www.yfdyf.cn/.

[20] https://github.com/stanfordnlp/stanza.

lacks completeness, leading to poor descriptiveness of the entities and potentially flawed relationships within a particular healthcare domain. This also limits the capacity of the graph to deliver useful facts or rules to power data-driven methods that can be used for making healthcare decisions [45]. To consolidate a healthcare KG and establish a cohesive viewpoint of the domain, alternative sources need to be incorporated and integrated including EMRs, PMRs, clinical trials, patient records, epidemiological surveillance, sensor data, disease registries, wearable devices, health workforce data, census data, implanted equipment, pill cameras, and all other relevant sources. However, full integration of such heterogeneous data sources can be a complicated and time-consuming task, especially when working with large-scale datasets where traditional data assimilation and aggregation techniques are not applicable. Therefore, there is still room for research to address the big data problem in healthcare KGs by developing advanced and sophisticated data collection and aggregation techniques.

- **Healthcare knowledge interoperability**: Linked Open Data (LOD) and Semantic Web technologies have made it possible to improve a variety of domain-specific applications [14, 151–153]. KGs represent an expansion of these efforts and are frequently connected with LOD initiatives because they improve data semantics by enhancing the conceptual representations of entities [154]. As a result, appropriate interlinking of entities gathered from different data sources facilitates information interoperability, resulting in multimodal KGs. However, some of the methodologies investigated in this study revealed difficulties in attaining the appropriate level of knowledge expandability and interoperability. In particular, semantic expansion strategies were underutilised, and their ability to take advantage of freely accessible vocabulary and semantic resources is mostly ignored. The expansion of healthcare knowledge with health records collected from different channels, such as hospital admissions, family physician visits, prescription drugs, pharmacy requests, laboratory blood analyses, and death certificates establishes a comprehensive individual health (or disease) profile [155]. This holistic view carries enormous implications for several research areas, such as epidemiology and precision medicine. Basic structure of KGs facilitates better data integration, unification, and information sharing. Semantic expansion adds context to the collected facts in the KGs and enhances the quality of the aggregated knowledge, eliminates redundant records, and detects missing entities. Based on success of existing healthcare semantic expansion initiatives such as the Centre for Health Record Linkage (CHeReL) in Australia [156] and Rochester Epidemiology Project in USA [157], more research in this direction should be conducted.
- **KG construction mechanisms**: The construction of the KG comprises several activities which might vary depending on the type of knowledge base (schema-based, schema-free, or hybrid), knowledge resources and their data types (structured or unstructured), knowledge extraction techniques (entity-level and relation-level), etc. Several of the examined studies failed to adequately disclose the internal mechanisms they used to build and implement the KGs. A shortcoming that was commonly observed was poor and/or limited discussion to explain either the overall construction methodology [48, 55, 66] or the essential construction tasks such as the ontology design [46, 100], entity and/or relation extraction [78, 83], and knowledge

integration [47, 93]. Furthermore, many of the KGs described in those papers are not publicly available for inspection. These drawbacks detract from knowledge sharing, translation, and reusing, and make the replication of the proposed approaches difficult. This is particularly problematic in the healthcare domain where knowledge replicability can assist in consolidating the facts about certain scientific tests and medical experiments [158]. Therefore, future studies must ensure that all steps of KG construction are well-explained, and the resultant KG must be publically shared with the community to reinforce FAIR principles (Findable, Accessible, Interoperable, Reusable)[13].

- **KG evaluation**: Despite the continuous propagation of KGs for the healthcare domain and its sub-domains, this survey reports evident problems with KG evaluation and/or case study implementation. Numerous KGs were constructed with no proper concern for evaluation of their quality [77–79, 82]. Additionally, there is only a limited utility in applying the constructed KGs to real-life applications. Instead of practical applications, the proposed KGs mainly attempted to provide an underlying conceptual structure of the domain utilising domain-specific entities, concepts, relationships, and events. For example, the authors of [76] attempted to build a KG for hepatocellular carcinoma with no verified utility in addressing the designated disease. Designing and implementing actionable healthcare analytics must be the essence of the KG construction philosophy, where relevant facts are obtained with the objective to conceptualise the correct context and address a domain problem, thereby achieving the hoped-for value. Future works must ensure that KGs are assessed using one or more appropriate evaluation and refinement methodologies such as (i) silver and gold standards [159]; (ii) theoretically proven computational measures such as precision and recall; and (iii) domain experts. In addition, the constructed KG must prove its utility and verify its applicability in real-life scenarios and for the execution of downstream tasks.
- **Data Quality and Privacy**Applying healthcare KGs to downstream tasks such as drug discovery, clinical decision support, and medical treatment relies profoundly on the high quality of the embedded facts. Although some of the examined works constructed their KGs using structural, verified and curated data sources [42, 94, 104], other KGs imported data from unstructured sources (such as scientific medical literature or social media), with little regard for applying data quality measures before incorporating the extracted information [56, 105]. Freely available texts such as scientific medical literature commonly comprise ambiguous data, abbreviations, and noisy data that includes words and phrases irrelevant to the designated context. EMRs also comprise a vital source of embedded clinical data that can be either mistakenly neglected or hard to collect due to confidentiality constraints. These challenges raise concerns about the quality and reliability of KGs generated from such data sources. Therefore, high-quality healthcare KGs should be constructed by selecting high-quality data sources and developing quality measurement techniques. Also, advanced NLP and deep learning algorithms that can efficiently and automatically identify high-quality entities and relations should be implemented wherever possible. Those tools should be used to improve data privacy, integrity,

[13] https://www.go-fair.org/fair-principles/.

and security, preventing malicious activities that attempt to abuse patients' sensitive medical information.

- **Recentness**: Most of the examined studies did not consider the temporal factor; their KGs are static in nature and often neglect the validity period of incorporated triples. A healthcare KG built based on just one snapshot of the knowledge landscape might not be a sustainable depiction of the designated domain, particularly with the emergence of wearable medical devices, sensors, health monitoring systems, and mobile applications [160] which make the construction of dynamic and frequently updated KGs a necessity. Ignoring the dynamic nature of healthcare knowledge degrades the quality and accuracy of facts embedded in KGs, consequently leading to poor data analytics and decision making.
- **Healthcare KG reasoning**: Reasoning of the KG aims to infer new facts and make new conclusions based on the existing data. KG reasoning allows for deriving new insights and enriches KGs with new relations. Several techniques have been proposed in the literature for KG reasoning, including ontology reasoning, logic rules, and random walk algorithm [161]. Recently, KG embedding approaches attracted a lot of attention in the research community due to their capacity to provide generalizations and infer new facts. KG embedding techniques aim to transform the KG into semantically-continuous low-dimensional space. The embedded KG can be then used for several downstream tasks including link prediction, knowledge discovery, etc [162]. This study reveals a relative lack of successful KG embedding strategies in the investigated papers.

Conclusion

The vast volume of healthcare data that is collected in a variety of formats and pertains to a wide range of subject matters presents a critical challenge for analysts. Knowledge Graphs (KGs) offer an effective answer to this challenge and open new possibilities for machines to understand meanings, closing the semantic gap between them and people. As a result, domain-specific knowledge graphs have been developed and applied to various real-world problems. Healthcare industry has greatly benefited from this technology, with numerous KGs created specifically to address different healthcare issues. However, the deficiencies and limitations of the current KG construction techniques stand in the way of obtaining the hoped-for value from this technology.

This paper offers a bird's eye view of the healthcare KG domain and tries to define a relevant construction paradigm. A critical review of the current construction approaches is conducted considering the methods used for knowledge extraction, types of knowledge bases and sources, and the adopted evaluation metrics. Finally, in conjunction with a summary of limitations and deficiencies, it provides pointers for potential future research that we hope will inspire scholars in this field.

Acknowledgements
Not applicable.

Authors' contributions
BAS: Conceptualization, Methodology, Visualization, Writing original draft. MQ: Writing, Review & Editing. BAH, MAW and MAS: Validation, Review & Editing. RAF, HS, and MJ: Supervision, Review & Editing. All authors read and approved the final manuscript.

Funding
Not applicable.

Data Availability
Not applicable.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate
Not applicable.

Consent for publication
We give the publisher the permission of the authors to publish the work.

Received: 28 June 2022 / Accepted: 17 May 2023
Published online: 28 May 2023

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–2.
2. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Informatics. 2018;114:57–65.
3. Rehman A, Naz S, Razzak I. *Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities*. Multimedia Syst, 2021: p. 1–33.
4. Hubauer T et al. *Use Cases of the Industrial Knowledge Graph at Siemens*. in *International Semantic Web Conference (P&D/Industry/BlueSky)*. 2018.
5. Abu-Salih B. Domain-specific knowledge graphs: a survey. J Netw Comput Appl. 2021;185:103076.
6. Wolffe TA, et al. A survey of systematic evidence mapping practice and the case for knowledge graphs in environmental health and toxicology. Toxicol Sci. 2020;175(1):35–49.
7. Min W, et al. Applications of knowledge graphs for food science and industry. Patterns. 2022;3(5):100484.
8. Chaoyu Z, Lei L. A review of medical decision supports based on knowledge graph. Data Anal Knowl Discovery. 2020;4(12):26–32.
9. Moher D, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9.
10. Gruber TR. *A translation approach to portable ontology specification*. in *Knowledge Acquisition*. 1993.
11. Stevens R. *What is an Ontology?* 2001 [cited 2017 3rd March]; Available from: http://www.cs.man.ac.uk/~stevensr/onto/node3.html
12. Cobb BL, et al. Genes and Sjögren's syndrome. Rheum Dis Clin North Am. 2008;34(4):847–vii.
13. Ehrlinger L, Wöß W. *Towards a Definition of Knowledge Graphs* SEMANTiCS (Posters, Demos, SuCCESS), 2016. 48: p. 1–4.
14. Abu-Salih B, et al. Social Big Data Analytics. Springer; 2021.
15. Zhang Y, et al. HKGB: an Inclusive, Extensible, Intelligent, semi-auto-constructed Knowledge Graph Framework for Healthcare with Clinicians' Expertise Incorporated. Volume 57. Information Processing & Management; 2020. p. 102324. 6.
16. Chen P et al. *An automatic knowledge graph construction system for K-12 education*. in *Proceedings of the fifth annual ACM conference on learning at scale*. 2018.
17. Schindler D, Zapilko B, Krüger F. *Investigating software usage in the social sciences: A knowledge graph approach*. in *European Semantic Web Conference*. 2020. Springer.
18. Tchechmedjiev A et al. *ClaimsKG: a knowledge graph of fact-checked claims*. in *International Semantic Web Conference*. 2019. Springer.
19. Yadav V, Bethard S. *A survey on recent advances in named entity recognition from deep learning models* arXiv preprint arXiv:1910.11470, 2019.
20. Al-Moslmi T, et al. Named entity extraction for knowledge graphs: a literature overview. IEEE Access. 2020;8:32862–81.
21. Abu-Salih B, et al. Time-aware domain-based social influence prediction. J Big Data. 2020;7(1):1–37.
22. Abu-Salih B. *Applying Vector Space Model (VSM) Techniques in Information Retrieval for Arabic Language* arXiv preprint arXiv:1801.03627, 2018.
23. Wongthontham P, Abu-Salih B. *Ontology-based approach for semantic data extraction from social big data: state-of-the-art and research directions* arXiv preprint arXiv:1801.01624, 2018.
24. Kim K, et al. GREG: A global level relation extraction with knowledge graph embedding. Appl Sci. 2020;10(3):1181.
25. Smirnova A, Cudré-Mauroux P. Relation extraction using distant supervision: a survey. ACM Comput Surv (CSUR). 2018;51(5):1–35.
26. Nickel M et al. *A review of relational machine learning for knowledge graphs* Proceedings of the IEEE, 2015. 104(1): p. 11–33.
27. Dong X et al. *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014.
28. Heist N. *Towards Knowledge Graph Construction from Entity Co-occurrence*. in *EKAW (Doctoral Consortium)*. 2018.
29. Kuhn P et al. Type inference on Wikipedia list pages. Informatik 2016, 2016.
30. Suchanek FM, Kasneci G, Weikum G. *Yago: a core of semantic knowledge*. in *Proceedings of the 16th international conference on World Wide Web*. 2007.
31. Wu W et al. *Probase: A probabilistic taxonomy for text understanding*. in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 2012.

32. Fader A, Soderland S, Etzioni O. *Identifying relations for open information extraction*. in *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011.
33. Carlson A, et al. *Toward an architecture for never-ending language learning*. In *Aaai*. Atlanta; 2010.
34. Chen H et al. *A Practical Framework for Evaluating the Quality of Knowledge Graph*. in *China Conference on Knowledge Graph and Semantic Computing*. 2019. Springer.
35. Gao J et al. *Efficient Knowledge Graph Accuracy Evaluation (Technical Report Version)*
36. Al-Okaily M, Al-Okaily A. An empirical assessment of enterprise information systems success in a developing country: the jordanian experience. The TQM Journal; 2022.
37. Al-Okaily M, et al. The effect of digital accounting systems on the decision-making quality in the banking industry sector: a mediated-moderated model. Global Knowledge, Memory and Communication; 2022.
38. Al-Okaily M et al. *Examining the critical factors of computer-assisted audit tools and techniques adoption in the post-COVID-19 period: internal auditors perspective*. VINE J Inform Knowl Manage Syst, 2022(ahead-of-print).
39. Kejriwal M. Domain-specific knowledge graph construction. Springer; 2019.
40. Pezeshkpour P, Tian Y, Singh S. Revisiting evaluation of knowledge base completion models. Automated Knowledge Base Construction; 2020.
41. Zaveri A, et al. Quality assessment for linked data: a survey. Semantic Web. 2016;7(1):63–93.
42. Mohammadhassanzadeh H et al. Investigating plausible reasoning over knowledge graphs for semantics-based health data analytics. in 2018 *IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. 2018. IEEE.
43. Cui L et al. *DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation*. in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020.
44. Zhang Y, et al. HKGB: an Inclusive, Extensible, Intelligent, semi-auto-constructed Knowledge Graph Framework for Healthcare with Clinicians' Expertise Incorporated. Information Processing & Management; 2020. p. 102324.
45. Malik KM, et al. Automated domain-specific healthcare knowledge graph curation framework: subarachnoid hemorrhage as phenotype. Expert Syst Appl. 2020;145:113120.
46. Mann M et al. Open Drug Knowledge Graph. 2021.
47. Che M, et al. Knowledge-graph-based drug repositioning against COVID-19 by Graph Convolutional Network with attention mechanism. Future Internet. 2021;13(1):13.
48. Zhang S, Lin X, Zhang X. *Discovering DTI and DDI by Knowledge Graph with MHRW and Improved Neural Network*. in 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021. IEEE.
49. Ye Q, et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun. 2021;12(1):1–12.
50. Zeng X, et al. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol. 2022;72:114–26.
51. Wang J et al. *SynLethDB 2.0: A web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery* Database, 2022. 2022.
52. Su C, Hou Y, Wang F. *GNN-based Biomedical Knowledge Graph Mining in Drug Development*, in *graph neural networks: foundations, frontiers, and applications*. Springer; 2022. pp. 517–40.
53. Dobreva J, Jovanovik M, Trajanov D. *DD-RDL: Drug-Disease Relation Discovery and Labeling*. in *International Conference on ICT Innovations*. 2022. Springer.
54. Zhou Y, et al. Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health; 2020.
55. Richardson P, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet (London England). 2020;395(10223):e30.
56. Wang Q et al. *COVID-19 literature knowledge graph construction and drug repurposing report generation* arXiv preprint arXiv:2007.00576, 2020.
57. Domingo-Fernández D, et al. COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics. 2021;37(9):1332–4.
58. Zhu Y et al. *Knowledge-driven drug repurposing using a comprehensive drug knowledge graph*. Health Inf J, 2020: p. 1460458220937101.
59. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinformatics. 2020;36(2):603–10.
60. Yan VK et al. *Drug Repurposing for the Treatment of COVID-19: A Knowledge Graph Approach* Advanced Therapeutics, 2021: p. 2100055.
61. Yang F et al. *Machine learning applications in drug repurposing* Interdisciplinary Sciences: Computational Life Sciences, 2022: p. 1–7.
62. Manian V, et al. Detection of Target genes for Drug Repurposing to treat skeletal muscle atrophy in mice flown in spaceflight. Genes. 2022;13(3):473.
63. Pan X, et al. Deep learning for drug repurposing: methods, databases, and applications. Wiley Interdisciplinary Reviews: Computational Molecular Science; 2022. p. e1597.
64. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. The lancet. 2000;356(9237):1255–9.
65. Walter SR, et al. The impact of serious adverse drug reactions: a population-based study of a decade of hospital admissions in New South Wales, Australia. Br J Clin Pharmacol. 2017;83(2):416–26.
66. Bean DM, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep. 2017;7(1):1–11.
67. Zhang F, et al. Prediction of adverse drug reactions based on knowledge graph embedding. BMC Med Inf Decis Mak. 2021;21(1):1–11.
68. Wang M, et al. Adverse drug reaction Discovery using a tumor-biomarker knowledge graph. Front Genet. 2020;11:1737.
69. Abdelaziz I, et al. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. J Web Semant. 2017;44:104–17.
70. Stanovsky G, Gruhl D, Mendes P. *Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models*. in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017.

71. Muñoz E, Nováček V, Vandenbussche P-Y. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. Brief Bioinform. 2019;20(1):190–202.
72. Su X, et al. Attention-based knowledge graph representation learning for Predicting Drug-drug interactions. Brief Bioinform. 2022;23(3):bbac140.
73. Hao X et al. *Enhancing drug–drug interaction prediction by three-way decision and knowledge graph embedding*. Granul Comput, 2022: p. 1–10.
74. Han X, et al. SmileGNN: drug–drug Interaction Prediction based on the SMILES and graph neural network. Life. 2022;12(2):319.
75. Zhao J et al. Construction of knowledge graph of drug-action mechanism. in 2021 *3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC)*. 2021. IEEE.
76. Li N, et al. KGHC: a knowledge graph for hepatocellular carcinoma. BMC Med Inf Decis Mak. 2020;20(3):1–11.
77. Yin Y et al. Study on construction and application of knowledge graph of TCM diagnosis and treatment of viral hepatitis B. in 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021. IEEE.
78. Choi W, Lee H. Identifying disease-gene associations using a convolutional neural network-based model by embedding a biological knowledge graph with entity descriptions. PLoS ONE. 2021;16(10):e0258626.
79. Sun Y et al. *An Intelligent Question-Answering System for Myopia Prevention and Control based on Knowledge Graph*. in *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences*. 2021.
80. Yang X, et al. Mining a stroke knowledge graph from literature. BMC Bioinformatics. 2021;22(10):1–19.
81. Lin C et al. *Construction of Knowledge Graph of Stroke Based on Meta-analysis Literature*. in *2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*. 2020. IEEE.
82. Cheng B et al. *Research on medical knowledge graph for stroke* Journal of Healthcare Engineering, 2021. 2021.
83. Huang L et al. COVID-19 Knowledge Graph for Drug and Vaccine Development. in 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021. IEEE.
84. Ma X, et al. InDISP: an interpretable model for dynamic illness severity prediction. Cham: Springer International Publishing: Cham; 2022. pp. 631–8.
85. Yang JJ, et al. Knowledge graph analytics platform with LINCS and IDG for Parkinson's disease target illumination. BMC Bioinformatics. 2022;23(1):1–15.
86. Hao F, Zheng K. *Online Disease Identification and Diagnosis and Treatment Based on Machine Learning Technology* Journal of Healthcare Engineering, 2022. 2022.
87. Lan W, et al. KGANCDA: predicting circRNA-disease associations based on knowledge graph attention network. Brief Bioinform. 2022;23(1):bbab494.
88. Yu S, et al. A knowledge-driven network for fine-grained relationship detection between miRNA and disease. Brief Bioinform. 2022;23(3):bbac058.
89. Yu G, et al. Improving chronic disease management for children with knowledge graphs and artificial intelligence. Expert Syst Appl. 2022;201:117026.
90. Yang R et al. *Decision-Making System for the Diagnosis of Syndrome Based on Traditional Chinese Medicine Knowledge Graph* Evidence-Based Complementary and Alternative Medicine, 2022. 2022.
91. Yuan J, et al. Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowl Inf Syst. 2020;62(1):317–36.
92. Broder AZ. *On the resemblance and containment of documents*. in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. 1997. IEEE.
93. Huang Z et al. *Constructing knowledge graphs of depression*. in *International Conference on Health Information Science*. 2017. Springer.
94. Li Z et al. Construction of Depression Knowledge Graph Based on Biomedical Literature. in 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021. IEEE.
95. Fu C et al. *MDepressionKG: a knowledge graph for metabolism-depression associations*. in *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences*. 2021.
96. Gunjan VK et al. *Machine Learning and Cloud-Based Knowledge Graphs to Recognize Suicidal Mental Tendencies* Computational Intelligence and Neuroscience, 2022. 2022.
97. Zhu Y et al. *A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection*. Knowl Based Syst, 2022: p. 108952.
98. Kim J, Sohn M. Graph representation learning-based early Depression Detection Framework in Smart Home environments. Sensors. 2022;22(4):1545.
99. Yang K, Zhang T, Ananiadou S. A mental state knowledge–aware and Contrastive Network for early stress and depression detection on social media. Inf Process Manag. 2022;59(4):102961.
100. Zheng S, et al. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in Bioinformatics; 2020.
101. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics. 2018;34(15):2614–24.
102. Wood E et al. *RTX-KG2: a system for building a semantically standardized knowledge graph for translational biomedicine*. bioRxiv, 2021.
103. Zhang Y et al. *Causal Discovery and Knowledge Linkage in Scientific Literature: A Case Study in Biomedicine*. in *International Conference on Information*. 2022. Springer.
104. Walsh B, Mohamed SK, Nováček V. *Biokg: A knowledge graph for relational learning on biological data*. in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020.
105. He F, et al. Research on construction of knowledge graph of intestinal cells. J Artif Intell Med Sci. 2020;1(1–2):15–22.
106. Sun Y et al. *KGBReF: A Knowledge Graph based Biomedical Relation Extraction Framework*. in *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences*. 2021.
107. Fei H, et al. Enriching contextualized language model from knowledge graph for biomedical information extraction. Brief Bioinform. 2021;22(3):bbaa110.
108. Varma S et al. Link Prediction Using Semi-Automated Ontology and Knowledge Graph in Medical Sphere. in 2020 *IEEE 17th India Council International Conference (INDICON)*. 2020. IEEE.

109. Sun H, et al. Medical knowledge graph to Enhance Fraud, Waste, and abuse detection on claim data: Model Development and performance evaluation. JMIR Med Inf. 2020;8(7):e17653.
110. Li J et al. *PlagueKG: A plague knowledge graph based on biomedical literature mining* 2022.
111. Li Z, et al. DeepKG: an end-to-end deep learning-based workflow for biomedical knowledge graph extraction, optimization and applications. Bioinformatics. 2022;38(5):1477–9.
112. Joachimiak MP et al. *KG-Microbe: A Reference Knowledge-Graph and Platform for Harmonized Microbial Information* 2021.
113. Liu T, et al. Exploring the microbiota-gut-brain axis for mental disorders with knowledge graphs. J Artif Intell Med Sci. 2020;1(3–4):30–42.
114. Fu C et al. *An Integrated Knowledge Graph for Microbe-Disease Associations*. in *International Conference on Health Information Science*. 2020. Springer.
115. Zhang P, et al. Toward a Coronavirus Knowledge Graph. Genes. 2021;12(7):998.
116. Chen C et al. *Coronavirus knowledge graph: A case study* arXiv preprint arXiv:2007.10287, 2020.
117. Wu J. Construct a knowledge graph for China Coronavirus (COVID-19) Patient Information Tracking. Risk Manage Healthc Policy. 2021;14:4321.
118. Krämer A, et al. The Coronavirus Network Explorer: mining a large-scale knowledge graph for effects of SARS-CoV-2 on host cell function. BMC Bioinformatics. 2021;22(1):1–20.
119. Santos A et al. *A knowledge graph to interpret clinical proteomics data*. Nat Biotechnol, 2022: p. 1–11.
120. Li L, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med. 2020;103:101817.
121. Chen IY et al. Robustly Extracting Medical Knowledge from EHRs: a case study of learning a Health Knowledge Graph. Pac Symp Biocomput. 2020. World Scientific.
122. Postiglione M. *Towards an Italian Healthcare Knowledge Graph*. in *International Conference on Similarity Search and Applications*. 2021. Springer.
123. Schick T, Schütze H. *Exploiting cloze questions for few shot text classification and natural language inference* arXiv preprint arXiv:2001.07676, 2020.
124. Schindler D, et al. The role of software in science: a knowledge graph-based analysis of software mentions in PubMed Central. PeerJ Comput Sci. 2022;8:e835–5.
125. Koshman V, Funkner A, Kovalchuk S. An Unsupervised Approach to structuring and analyzing repetitive semantic structures in Free text of Electronic Medical Records. J Pers Med. 2022;12(1):25.
126. Sahu A, et al. Method and system for content processing to query multiple healthcare-related knowledge graphs. Google Patents; 2018.
127. Sheng M et al. *DSQA: A Domain Specific QA System for Smart Health Based on Knowledge Graph*. in *International Conference on Web Information Systems and Applications*. 2020. Springer.
128. Zhao J et al. *Design and Construction of Knowledge Graph of Electronic Medical Record Based on BiLSTM-CRF*. in *2021 4th International Conference on Big Data Technologies*. 2021.
129. Xie Y. *A TCM Question and Answer System Based on Medical Records Knowledge Graph*. in 2020 *International Conference on Computing and Data Science (CDS)*. 2020. IEEE.
130. Shuai Q et al. *Research on Intelligent Question Answering System Based on Medical Knowledge Graph*. in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. 2019. IEEE.
131. Huang X, et al. A knowledge graph based question answering method for medical domain. PeerJ Comput Sci. 2021;7:e667.
132. Yin Y et al. *Question Answering System Based on Knowledge Graph in Traditional Chinese Medicine Diagnosis and Treatment of Viral Hepatitis B* BioMed Research International, 2022. 2022.
133. Huang L et al. *Towards smart healthcare management based on knowledge graph technology*. in *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. 2019.
134. Haussmann S et al. *FoodKG: a semantics-driven knowledge graph for food recommendation*. in *International Semantic Web Conference*. 2019. Springer.
135. Chi Y, et al. Knowledge Management in Healthcare sustainability: a Smart Healthy Diet Assistant in Traditional Chinese Medicine Culture. Sustainability. 2018;10(11):4197.
136. Zhang X et al. *Research and Analysis on the Field of Food Additive by Knowledge Graph Construction*. in *Journal of Physics: Conference Series*. 2019. IOP Publishing.
137. Groth P, Gibson A, Velterop J. The anatomy of a nanopublication. Inform Serv Use. 2010;30(1–2):51–6.
138. Yu H et al. *A domain knowledge graph construction method based on Wikipedia*. J Inform Sci, 2020: p. 0165551520932510.
139. Zeng J, Nakano YI. *Exploiting a Large-scale Knowledge Graph for Question Generation in Food Preference Interview Systems*. in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*. 2020.
140. Salehian H, et al. Food knowledge graph for a health tracking system. Google Patents; 2019.
141. Wang F et al. *Recent Advances on Graph Analytics and Its Applications in Healthcare*. in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020.
142. Tao X, et al. Mining health knowledge graph for health risk prediction. World Wide Web; 2020. pp. 1–22.
143. Goodwin TR, Harabagiu SM. Knowledge representations and inference techniques for medical question answering. ACM Trans Intell Syst Technol (TIST). 2017;9(2):1–26.
144. Ma Y, Tresp V, Daxberger EA. Embedding models for episodic knowledge graphs. J Web Semant. 2019;59:100490.
145. Rastogi N, Zaki MJ. *Personal Health Knowledge Graphs for Patients* arXiv preprint arXiv:2004.00071, 2020.
146. Su J et al. *An Interactive Knowledge Graph Based Platform for COVID-19 Clinical Research*. in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 2022.
147. Chen T, et al. A knowledge graph-based method for epidemic contact tracing in public transportation. Transp Res Part C: Emerg Technol. 2022;137:103587.
148. Shen Y, et al. KGDDS: a system for drug-drug similarity measure in therapeutic substitution based on knowledge graph curation. J Med Syst. 2019;43(4):92.
149. Sadeghi A, Lehmann J. *Linking physicians to medical research results via knowledge graph embeddings and Twitter*. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2019. Springer.
150. Lin Z, Yang D, Yin X. Patient similarity via Joint Embeddings of Medical Knowledge Graph and Medical Entity descriptions. IEEE Access; 2020.

151. Abu-Salih B, Wongthongtham P, Chan KY. *Twitter mining for ontology-based domain discovery incorporating machine learning*. J Knowl Manage, 2018.
152. Wongthongtham P, et al. State-of-the-art ontology annotation for personalised teaching and learning and prospects for smart learning recommender based on multiple intelligence and fuzzy ontology. Int J Fuzzy Syst. 2018;20(4):1357–72.
153. Wongthongtham P, Salih BA. Ontology-based approach for identifying the credibility domain in social Big Data. J Organizational Comput Electron Commer. 2018;28(4):354–77.
154. Ehrlinger L, Wöß W. *Towards a definition of knowledge graphs* SEMANTiCS (Posters, Demos, SuCCESS), 2016. 48(1–4): p. 2.
155. Montecucco F, et al. Implementation strategies of systems medicine in clinical research and home care for cardiovascular disease patients. Eur J Intern Med. 2014;25(9):785–94.
156. Irvine K, Hall R, Taylor L. *A profile of the Centre for Health Record linkage*. Int J Popul Data Sci, 2019. 4(2).
157. Rocca WA, et al. Data resource profile: expansion of the Rochester Epidemiology Project medical records-linkage system (E-REP). Int J Epidemiol. 2018;47(2):368–368j.
158. Coiera E, et al. Does health informatics have a replication crisis? J Am Med Inform Assoc. 2018;25(8):963–8.
159. Paulheim H. Knowledge graph refinement: a survey of approaches and evaluation methods. Semantic web. 2017;8(3):489–508.
160. Ed-daoudy A, Maalmi K. A new internet of things architecture for real-time prediction of various diseases using machine learning on big data environment. J Big Data. 2019;6(1):1–25.
161. Chen X, Jia S, Xiang Y. A review: knowledge reasoning over knowledge graph. Expert Syst Appl. 2020;141:112948.
162. Abu-Salih B, et al. Relational learning analysis of social politics using knowledge graph embedding. Data Min Knowl Disc. 2021;35(4):1497–536.

Publisher's Note