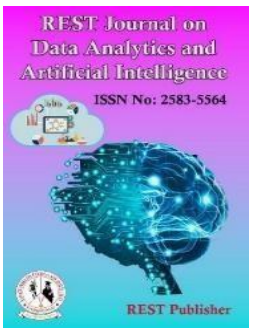
REST Journal on Data Analytics and Artificial Intelligence
Vol: 2(4), December 2023
REST Publisher; ISSN: 2583-5564

Website: http: //restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/2/4/1



Knowledge Graph Representation in Medical Era- An Analysis

[*1] Deepthi Rani S S, [2]Anub A Nair

[1]Christ Nagar College, Trivandrum,Kerala, India.
[2]Mahaguru Institute of Technology, Kayamkulam,Kerala, India.
*Corresponding Author Email: deepthisiva.ss@gmail.com

Abstract- *Healthcare knowledge graphs (HKGs) are considered as a tool for organizing medical knowledge in a structured and interpretable way, which provides a comprehensive view of medical concepts and their relationships. However, challenges such as data heterogeneity and limited coverage remain, emphasizing the need for further research in the field of HKGs.Here summarize the pipeline and key techniques for HKG construction, as well as the common utilization approaches (i.e., model free and model-based). To provide researchers with valuable resources, we organize existing HKGs1 based on the data types they capture and application domains, supplemented with pertinent statistical information. In the application section, we delve into the transformative impact of HKGs across various healthcare domains, spanning from fine-grained basic science research to high-level clinical decision support. Lastly, we shed light on the opportunities for creating comprehensive and accurate HKGs in the era of large language models, presenting the potential to revolutionize healthcare delivery and enhance the interpretability and reliability of clinical prediction*

1. INTRODUCTION

A knowledge graph (KG) is a data structure that captures the relationships between different entities and their attributes .KG models and integrates data from various sources, including structured and unstructured data, and has been studied to support a wide range of applications such as search engines, recommendation systems. Particularly for healthcare, KG facilitates an interpretable representation of medical concepts such as drugs and disease, which enables context aware insights and enhances clinical research, decision-making, and healthcare delivery On the data side, Healthcare knowledge graphs (HKGs) are usually built on the landscape from complex medical systems such as electronic health records, medical literature, clinical guidelines, and patient-generated data .However, these data resources are often heterogeneous and distributed, making it challenging to integrate and analyse them effectively .The data heterogeneity can also lead to incomplete or inconsistent data representations within HKGs, limiting their usefulness for downstream healthcare tasks. Additionally, the current use of domain-specific knowledge graphs may result in limited coverage and granularity of the knowledge captured across different levels, hindering the ability to identify correlations and relationships between medical concepts from multiple domains. These challenges underscore the need for continued research on HKGs to fully realize their potential.

A comprehensive and fine-grained healthcare knowledge graph holds the potential to revolutionize healthcare across various levels .At the micro-scientific level, HKGs can help researchers identify new phenotypic and genotypic correlations and understand the underlying mechanisms of disease ,leading to more targeted and effective treatments .At the clinical care level, HKGs can be used to develop clinical decision support systems that provide clinicians with relevant information, improving clinical workflows and patient outcomes. Therefore, conducting an extensive survey of the existing literature on healthcare knowledge graphs becomes an indispensable roadmap and invaluable resource

for constructing a comprehensive HKG that can drive transformative advancements in healthcare. To the best of our knowledge, this survey paper represents the first comprehensive overview of healthcare knowledge graphs (HKGs).

The content overview is depicted in Figure 1, providing a visual summary of the key aspects discussed. We delve into the construction pipelines of HKGs, including both building from scratch and integration approaches, and highlight the key techniques employed in HKG construction. Additionally, we explore two common utilization methods of HKGs, namely model-free and model-based approaches. Finally, we address the unique challenges associated with HKGs and discuss promising research directions, particularly in leveraging large language models to enhance their potential .This survey paper targets a wide range of audience, including researchers, practitioners, clinicians, and other experts in healthcare, medical informatics, data science, and artificial intelligence.A healthcare knowledge graph (HKG) is a domain-specific knowledge graph designed to capture medical concepts such as drugs, diseases, genes, phenotypes, and so on, and their relationships in a structured and semantic way.

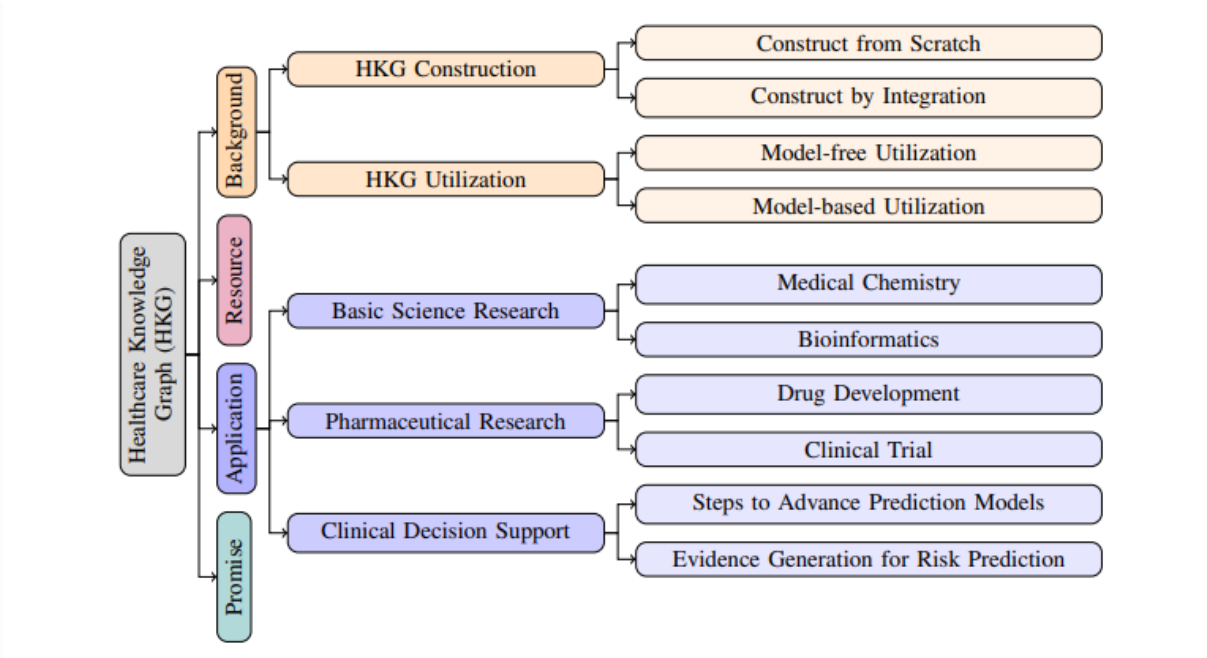


FIGURE 1. Overview of healthcare knowledge graph

2. HEALTHCARE KNOWLEDGE GRAPH CONSTRUCTION

Techniques for HKG Constructions. Traditionally, each step of HKG construction involves one specially designated model. For instance, Hidden Markov Models and Recurrent Neural Networks are widely used for healthcare named entity recognition, relation extraction, and other sequence tagging tasks, while Translational Models and Graph Neural Networks are used for HKG completion and conflicts resolution tasks. Recently, large language models (LLMs) have shown great utility to serve as a uniform tool for constructing KGs .Several key steps of constructing KGs, such as named entity recognition, relation extraction, entity, and KG completion, have been successfully tackled by these large foundation models. Early explorations of construction HKG with large foundation models show that healthcare entity normalization, healthcare entity recognition, healthcare entity linking, and healthcare knowledge fusion can also be performed, without extensive training on expensive healthcare annotated corpus. On the other hand, researchers start to construct KGs under the open-world assumption, thus getting rid of the dependency on pre-defined schemas and exhaustive entity relation normalization. Although open-world KGs greatly increase the coverage, ensuring the quality of extracted knowledge is still an open research challenge, especially for explainable and trustworthy HKGs.

3. HEALTHCARE KNOWLEDGE GRAPH UTILIZATION

Various query languages can be used for KGs, such as SPARQL, Cypher, and. These query languages allow users to query healthcare KGs using a standardized syntax, thus enabling users to retrieve, manipulate, and analyze data in a structured and consistent way. More complex applications can be further supported by graph queries.

For instance, automatic healthcare question answering can be tackled by Natural Language Question-to-Query (NLQ2Query) approach (Kim et al., 2022), where natural language questions are first translated into executable graph queries and then answered by the query responses. HKGs can also be utilized as an up-to-date and trustworthy augmentation to large language models (LLMs) for many applications. Some pioneering studies show that retrieved knowledge triples can improve the reliability of LLMs in various knowledge-intensive tasks, by addressing the nonsensical or unfaithful generation. Moreover, KGs can be a useful tool for fact-checking as they provide a structured representation of information that can be used to quickly and efficiently verify the accuracy of claims. Researchers have explored the utility of HKGs in identifying ingredient substitutions of food (Shirai et al., 2021), COVID-19 fact-checking, etc.

Model-based Utilization. Utilizing HKGs in complex reasoning tasks often involves utilizing machine learning models. HKG embeddings have shown great potential to tackle these tasks. In particular, HKG embedding models are a class of machine learning models that aim to learn low-dimensional vector representations, or embeddings, of the entities and relations in a knowledge graph. After obtaining HGK embeddings, they can be plugged into any kind of deep neural network and further fine-tuned toward downstream objectives. On the other hand, symbolic logic models represent another prominent approach for KG reasoning due to their interpretability. More specifically, symbolic reasoning models first mine logical rules from existing knowledge by inductive logic programming, association rule mining, or Markov logic networks. These minded rules are used to infer new facts, make logical deductions and answer complex queries. Recently, researchers start to explore combining logical rules into KG embedding to further improve the generalization and performance of HKG reasoning.

4. APPLICATIONS

Several previous biological terms can also be considered as knowledge graphs such as ontology (e.g., gene ontology, cell ontology, disease ontology), network (e.g., gene regulatory network), etc. We use the original biological terms as they are more popular according to historical reasons.

Drug-drug interactions (DDIs): It refer to changes in the actions, or side effects, of drugs when they are taken at the same time or successively (Giacomini et al., 2007). In general, DDIs are a significant contributor to life-threatening adverse events (Su et al., 2022; Pang et al., 2022; Yu et al., 2023), and their identification is one of the key tasks in public health and drug development. The existence of diverse datasets on drug-drug interactions (DDIs) and biomedical KGs has enabled the development of machine learning models that can accurately predict DDIs (Zhong et al., 2023). Yu et al. (2021) develop SumGNN, a model that includes a subgraph extraction module to efficiently extract relevant subgraphs from a KG, a self-attention-based summarization scheme to generate reasoning paths within the subgraph, and a multichannel module for integrating knowledge and data, resulting in significantly improved predictions of multi-typed DDIs. Su et al. (2022) propose DDKG, an attention-based KG representation learning framework that involves an encoder-decoder layer to learn the initial embeddings of drug nodes from their attributes in the KG. Karim et al. (2019) compare various techniques for generating KG embeddings with different settings and conclude that a combined convolutional neural network and LSTM yield the highest accuracy when predicting drug-drug interactions (DDIs). Dai et al. (2021) propose a new KG embedding framework by introducing adversarial autoencoders based on Wasserstein distances and Gumbel-Softmax relaxation for DDI tasks. Lin et al. (2020) develop KGNN that resolves the DDI prediction by capturing drug and its potential neighbourhoods by mining their associated relations in KG.

Drug-target interactions (DTIs): It is just as important as DDIs (Chen et al., 2016). Machine learning models can leverage knowledge graphs constructed from various types of interactions, such as drug-drug, drug-disease, proteindisease, and protein-protein interactions, to aid in the prediction of DTIs. For instance, Li et al. (2023) utilize the KG transfer probability matrix to redefine the drugdrug and target-target similarity matrix, thus constructing the final graph adjacent matrix to learn node representations by VGAE and augmenting them by utilizing dual Wasserstein Generative Adversarial Network with gradient penalty. Zhang et al. (2021c) propose a new hybrid method for DTI prediction by first constructing DTI-related KGs and then employing graph representation learning model to obtain feature vectors of the KG. Wang et al. (2022b) construct a knowledge graph of 29,607 positive drug-target pairs by

DistMult embedding strategy, and propose a ConvConv module to extract features of drug-target pairs. Ye et al. (2021b) learn a low-dimensional representation for various entities in the KG, and then integrate the multimodal information via neural factorization machine.

5. CONCLUSION

In conclusion, healthcare knowledge graphs (HKGs) offer a promising approach to capturing and organizing medical knowledge in a structured and interpretable way, providing a comprehensive and fine-grained view of medical concepts and relationships. Despite challenges like data heterogeneity and limited coverage, recent technical advancements have enabled the creation of comprehensive and precise HKGs. This survey provides a comprehensive overview of the current state of HKGs, covering their construction, utilization models, and applications in healthcare. We also discuss potential future developments, emphasizing the importance of HKGs in facilitating efficient and effective healthcare delivery

REFERENCES

[1]. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. Large language models are few-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1998–2022, 2022.

[2]. Alshahrani, M., Khan, M. A., Maddouri, O., Kinjo, A. R., Queralt-Rosinach, N., and Hoehndorf, R. Neurosymbolic representation learning on biological knowledge graphs. Bioinformatics, 33(17):2723–2730, 2017.

[3]. Antikainen, E., Linnosmaa, J., Umer, A., Oksala, N., Eskola, M., van Gils, M., Hernesniemi, J., and Gabbouj, M. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. Scientific Reports, 13(1):3517, 2023