Journal of Medical Systems (2023) 47:65
https://doi.org/10.1007/s10916-023-01951-2

COMMENT

From Data to Wisdom: Biomedical Knowledge Graphs for Real-World Data Insights

Katrin Hänsel[1] · Sarah N. Dudgeon[1] · Kei-Hoi Cheung[2,3] · Thomas J. S. Durant[1] · Wade L. Schulz[1,3]

Received: 22 December 2022 / Accepted: 15 April 2023 / Published online: 17 May 2023

Abstract
Graph data models are an emerging approach to structure clinical and biomedical information. These models offer intriguing opportunities for novel approaches in healthcare, such as disease phenotyping, risk prediction, and personalized precision care. The combination of data and information in a graph model to create knowledge graphs has rapidly expanded in biomedical research, but the integration of real-world data from the electronic health record has been limited. To broadly apply knowledge graphs to EHR and other real-world data, a deeper understanding of how to represent these data in a standardized graph model is needed. We provide an overview of the state-of-the-art research for clinical and biomedical data integration and summarize the potential to accelerate healthcare and precision medicine research through insight generation from integrated knowledge graphs.

Keywords Biomedical knowledge graph · Medical knowledge curation · Healthcare applications · Clinical outcome prediction

Introduction

The term Knowledge Graph (KG) was originally coined by Google [1] as a concept for structuring information into graphs to enhance web search. KGs are now ubiquitous around us and they power the modern web from tailored search results to personalized recommendations. KGs have also gained traction in biomedicine to represent public biomedical knowledge, integrate immunological research data, and advance drug discovery [2] and in healthcare, KGs have been used to support applications such as clinical decision support systems [3]. However, the use of KGs to model and drive discovery from real-world data (RWD), such as data from the electronic health record (EHR), has been limited. Among the emerging literature, findings suggest that there is a likely benefit from augmenting healthcare data with external knowledge, such as biomedical KGs, for applications such as disease risk prediction [4]. In this manuscript, we provide a brief overview of KGs and describe recent successes and future applications for healthcare data.

Technical background: Graphs to knowledge graphs

Biomedical data have an inherent graph structure, such as drug-disease interactions that capture data from multiple domains represented as bipartite networks, or protein-protein interactomes which can be represented in a unipartite network [5, 6]. However, traditional approaches to model and analyze these data are often reductionist, relying on constrained, tabular models without machine-readable context. Graphs provide the opportunity to model not just data, but also metadata and complex relationships between data elements. As opposed to the rows and columns used with more traditional approaches, graph-oriented models are represented with nodes, or vertices, and edges, or relationships. When data and information are joined in such a structure, the resulting data representation is referred to as a *knowledge graph*, which provides a computationally accessible (i.e.,

✉ Wade L. Schulz
wade.schulz@yale.edu

[1] Department of Laboratory Medicine, Yale School of Medicine, New Haven, CT, USA

[2] Section of Biomedical Informatics, Department of Emergency Medicine, Yale School of Medicine, 55 Park Street, PS 210, New Haven, CT 06510, USA

[3] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

machine-readable) representation of relationships between disparate biologic systems information. From these KGs, novel relationships can be identified and used to generate wisdom and actionable insights, as demonstrated in recent publications from the fields of computational biology and the life sciences [7-9]. In Fig. 1, we illustrate the application of the data-to-wisdom pyramid [10] to graph structures, which demonstrates how stepwise structuring, contextualization, and integration of graph-oriented data, information, knowledge, and wisdom can be used to drive insight generation.

Knowledge graph applications

The use of graph models and KGs has increased with access to more accessible graph database software. These technologies have shown efficient query performance, offer unique visualization tools, and have integrated specialized analytics packages for data science applications [11].

In recent years, biomedical researchers have increasingly adopted these technologies to better model the complexity of biological systems. In 2019, Bukhari et al. [12] used Neo4J – an enterprise graph database – to build a KG from multiple, variably formatted, publicly available data sources to support systems-based vaccinology. Related research demonstrates further insights derived from novel KG analysis [12]. For example, Youn et al. [9] describe the construction of an *Escherichia coli* antibiotic resistance KG that integrates 10 publicly available data sources. With this approach, the authors identified six novel *Escherichia coli* resistance genes that were identified via graph-based *in-silico* link predictions which were then validated biologically.

Other work has demonstrated the reduction of complex, graph-structured information and knowledge into simpler mathematical representations, termed embeddings, which can be used to retain the structural information encoded in graphs to facilitate downstream processing and analytics [13]. Embeddings build the basis for many graph-based data science tasks, such as prediction of edges between

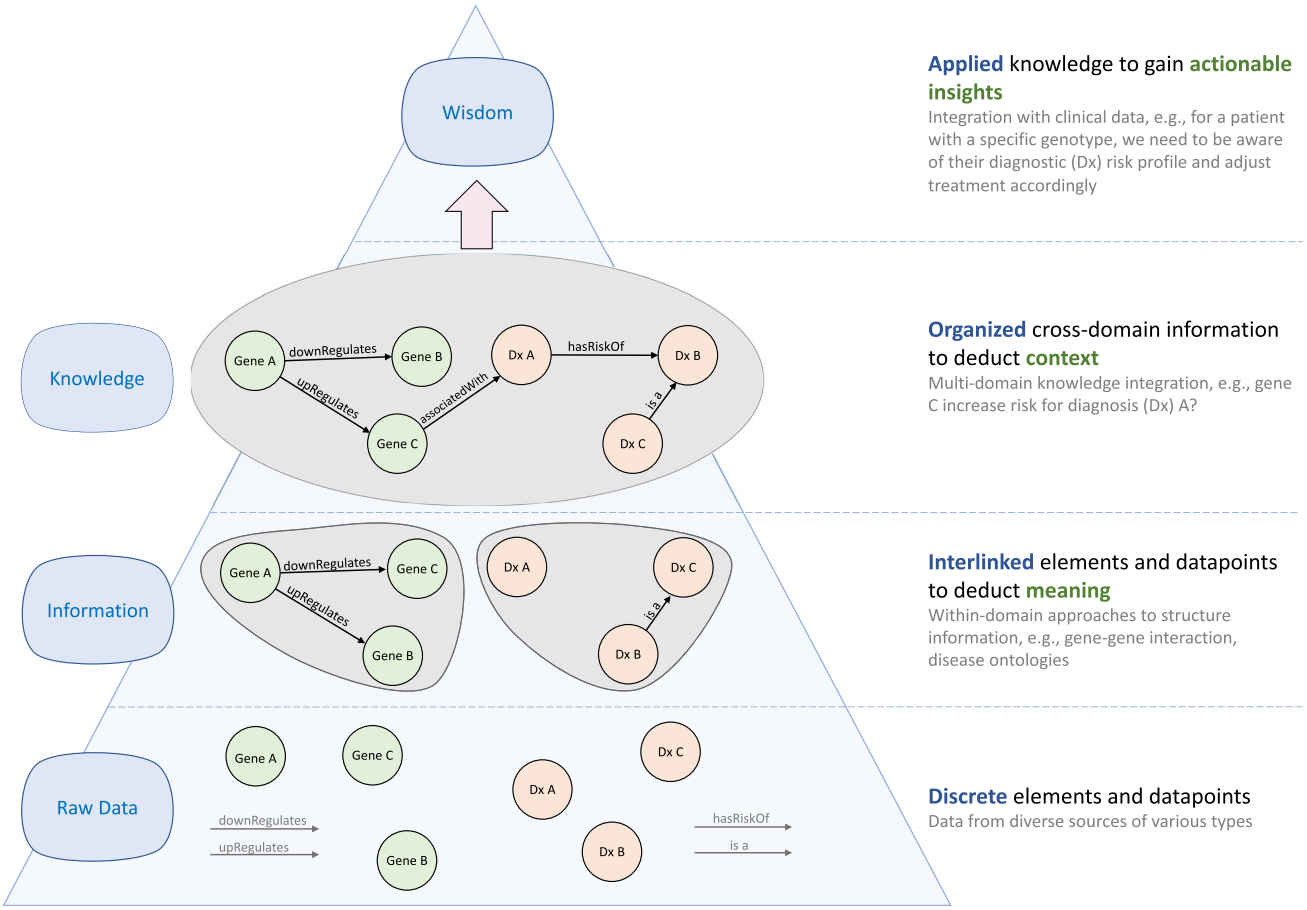

Fig. 1 Application of the Data-to-Wisdom Pyramid to biomedical graph data. Individual data elements can be connected within a graph to model information. The cross-domain aggregation of this information embeds knowledge directly within the data model, resulting in a knowledge graph (KG). These graphs can be used to generate insights or wisdom, such as clinical decision support or knowledge-enabled explanations, compared to a data model that may lack the more detailed context and relationships that are present in a graph model.

nodes, classification of node types, or clustering of related nodes [14]. Embeddings of large biomedical KGs based on the Unified Medical Language System (UMLS) have been used to predict similarity in meaning of medical concepts, a method that can be applied for natural language processing (NLP) of clinical texts [15].

Knowledge graph applications in healthcare

Graph networks have rapidly gained traction in basic and translational data sets and there is high potential for graph databases to be applied in healthcare [16]. However, less has been published on the use of graph representations to model patient data as shown in a recent systematic review by Schrodt et al. [3], which only identified 11 articles that used graphs to represent clinical data, such as laboratory results and comorbidities (Table 1).

The longitudinal efforts of Baranzini and colleagues have demonstrated how biomedical KGs, such as Hetionet [5] and its successor, the Scalable Precision Medicine Open Knowledge Engine (SPOKE), can be used to integrate information such as biological processes, molecular functions, complex diseases, as well as macro-cellular structures, proteins, and pathways for use in a variety of biomedical and healthcare applications. In 2019, Nelson et al. [20] described the connection of SPOKE with clinical electronic health record (EHR) data by leveraging shared concepts between the KG and EHR. Compared to using EHR data alone, the enrichment of clinical data with KG embeddings from SPOKE improved the performance of a downstream machine learning-based algorithm to predict the future diagnosis of multiple sclerosis [4]. It was hypothesized that integrating the KG data compensated for missing and/or incomplete data in the EHR. While this study showed promising results in the detection of the prodromal phase of multiple sclerosis, the generalizability of their embedding approach for other EHR data sets and predictive tasks remains unknown.

Limitations to knowledge graph implementation

While KGs are an intriguing solution to layer biomedical knowledge on clinical data sets, there remain challenges related to implementation and generalizable application. Firstly, biomedicine is an ever-evolving field with over one million new publications added to PubMed each year – nearly one publication per minute [21]. This poses the challenge of how to efficiently and accurately integrate this newly generated knowledge into a graph model. Secondly, the selection of appropriate node types and vocabularies to model data from diverse data sources can be a time-consuming and imperfect process, but one that is necessary to connect typically disparate data sets. Thirdly, while much has been done to create relational common data models for real-world data, no such standards exist today for graph-based data models. There is a need to integrate knowledge graphs with existing ontologies and linked data that have been implemented using semantic web technologies such as the Resource Description Framework (RDF). Finally, additional studies assessing the performance of graph-based clinical machine learning and artificial intelligence are needed to demonstrate the usefulness of knowledge graph models for these applications.

Table 1 Overview of biomedical and clinical use cases that can be addressed using graph and knowledge graph-based approaches

Healthcare Use Case	Graph and Knowledge Graph Mechanics: Explanation and examples
Drug repurposing, Comorbid risk prediction	**Link prediction:** Prediction of the likelihood of an existing edge between two nodes based on the entirety of the knowledge. For example, prediction of edge likelihood of drug compound and patient to predict personal risk of adverse reaction or links between two diseases posing a high comorbidity risk [7, 17].
Disease subtyping	**Community detection/graph clustering:** Identification of highly connected regions within a real-world data graph that can identify patients with a high similarity, e.g., patients with a certain disease subtype [8].
Outcome, status, and risk prediction	**Node classification:** Prediction of the likelihood of a patient node being assigned a label based on the entirety of their medical data. For example, patient node gets assigned a disease risk label [4].
Visual insights	**Graph layout and visualization:** There have been several studies into the visualization and lay outing of graph-structured data, e.g., biomedical or healthcare data, for aiding human interpretability and pattern recognition [18].
Complex patient data queries	**Graph traversal:** The inherent connected representation in graphs allows for the easy traversal of the graph to identify pieces of information that are separated by several nodes. When combining patient data with terminological knowledge, this allows for complex queries, e.g., identification of all patients based on a medical condition and its subtypes [19].

Conclusions

The integration of EHR data into KGs represents a promising approach to enhance clinical and translational research. However, these efforts are still in the early stages of development and require more in-depth testing and translation before they will be routinely placed into practice. The enrichment of EHR and other real-world data with broad biomedical knowledge bases is a lofty, but intriguing and alluring goal. As a scientific community, through the accumulation and contextualization of vast amounts of information and knowledge, we have the opportunity to create next-generation data models that can embed knowledge to provide greater context for analytics and machine learning applications, drive applications that provide actionable insights, and advance the field of real-world evidence generation. But to make these data models accessible and generalizable, further research is needed to understand best practices regarding how clinical data can be transformed into graph models to support downstream analytic tasks. Achieving this will allow us to better model complex, multi-modal information borne out from discoveries of the past and years to come.

Acknowledgements SD received funding from the Immunohematology Transfusion Medicine Research Training Grant (NHLBI), award number 2T32HL007974.

Author contributions All authors conceptualized the study. K.H., S.N.D., and T.J.S.D. drafted the manuscript. K-H.C. and W.L.S. provided significant writing, reviewing, and editing of all versions.

Data Availability As a comment article, this manuscript did not analyze any specific data sets.

Declarations

Competing interests T.J.S.D. was a consultant for Roche, a diagnostics company (fees); was a consultant for Instrumentation Laboratories, a diagnostics company (fees).W.L.S. was a technical consultant to HugoHealth, a personal health information platform (equity, fees); is a cofounder of Refactor Health, an AI-augmented data management platform for healthcare (equity); was a consultant for Abbott, a diagnostics company (fees); received a speaker honorarium from Instrumentation Laboratories, a diagnostics company. K.H. reports no competing financial or non-financial interests. S.N.D. reports no competing financial or non-financial interests. K-H.C. reports no competing financial or non-financial interests. T.J.S.D. was a consultant for Roche, a diagnostics company (fees); was a consultant for Instrumentation Laboratories, a diagnostics company (fees). W.L.S. was a technical consultant to HugoHealth, a personal health information platform (equity, fees); is a cofounder of Refactor Health, an AI-augmented data management platform for healthcare (equity); was a consultant for Abbott, a diagnostics company (fees); received a speaker honorarium from Instrumentation Laboratories, a diagnostics company.

References

1. Singhal A (2012) Introducing the knowledge graph: things, not strings. In: Official google blog. https://blog.google/products/search/introducing-knowledge-graph-things-not/. Accessed 2 Dec 2022
2. Nicholson DN, Greene CS (2020) Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal 18:1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017
3. Schrodt J, Dudchenko A, Knaup-Gregori P, Ganzinger M (2020) Graph-Representation of Patient Data: a Systematic Literature Review. Journal of Medical Systems 44:86. https://doi.org/10.1007/s10916-020-1538-4
4. Nelson CA, Bove R, Butte AJ, Baranzini SE (2022) Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. Journal of the American Medical Informatics Association 29:424–434. https://doi.org/10.1093/jamia/ocab270
5. Himmelstein DS, Lizee A, Hessler C, et al (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife 6:e26726. https://doi.org/10.7554/eLife.26726
6. Luck K, Kim D-K, Lambourne L, et al (2020) A reference map of the human binary protein interactome. Nature 580:402–408. https://doi.org/10.1038/s41586-020-2188-x
7. Bean DM, Wu H, Iqbal E, et al (2017) Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Scientific Reports 7:16416. https://doi.org/10.1038/s41598-017-16674-x
8. Liu C, Cao W, Wu S, et al (2022) Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19:1193–1202. https://doi.org/10.1109/TCBB.2020.3010509
9. Youn J, Rai N, Tagkopoulos I (2022) Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature Communications 13:2360. https://doi.org/10.1038/s41467-022-29993-z
10. Ackoff, Russel L. (1989) From data to wisdom. Journal of Applied Systems Analysis 16:
11. Miller JJ (2013) Graph database applications and concepts with neo4j. In: Proceedings of the Southern Association for Information Systems Conference
12. Bukhari SAC, Pawar S, Mandell J, et al (2021) LinkedImm: a linked data graph database for integrating immunological data. BMC Bioinformatics 22:105. https://doi.org/10.1186/s12859-021-04031-9
13. Grover A, Leskovec J (2016) Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, pp 855–864
14. Cai H, Zheng VW, Chang KC-C (2018) A comprehensive survey of graph embedding: Problems, techniques, and applications.

IEEE Transactions on Knowledge and Data Engineering 30:1616–1637. https://doi.org/10.1109/TKDE.2018.2807452
15. Mao Y, Fung KW (2020) Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. Journal of the American Medical Informatics Association 27:1538–1546. https://doi.org/10.1093/jamia/ocaa136
16. Stothers JAM, Nguyen A (2020) Can Neo4j Replace PostgreSQL in Healthcare? AMIA Jt Summits Transl Sci Proc 2020:646–653
17. Fernandes da Silva C, Abraham KJ, Seron Ruiz EE (2019) Comorbidity prediction and validation using a disease gene graph and public health data. In: 2019 8th brazilian conference on intelligent systems (BRACIS). pp 860–865
18. Dabek F, Chen J, Garbarino A, Caban JJ (2015) Visualization of longitudinal clinical trajectories using a graph-based approach. In: Proceedings of the 2015 workshop on visual analytics in healthcare. Association for Computing Machinery, New York, NY, USA
19. Campbell WS, Pedersen J, McClay JC, et al (2015) An alternative database approach for management of SNOMED CT and improved patient data queries. Journal of Biomedical Informatics 57:350–357. https://doi.org/10.1016/j.jbi.2015.08.016
20. Nelson CA, Butte AJ, Baranzini SE (2019) Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nature Communications 10:3045. https://doi.org/10.1038/s41467-019-11069-0
21. Landhuis E (2016) Scientific literature: Information overload. Nature 535:457–458. https://doi.org/10.1038/nj7612-457a