

Mediverse: Multimodal Clinical Exploration and Search on a Single Graph

Andra Ionescu

KTH Royal Institute of Technology
Stockholm, Sweden
adio@kth.se

Sebastiaan Meijer

KTH Royal Institute of Technology
Stockholm, Sweden
smeijer@kth.se

Paris Carbone

KTH Royal Institute of Technology
Stockholm, Sweden
parisc@kth.se

Jayanth Raghothama

KTH Royal Institute of Technology
Stockholm, Sweden
jayanthr@kth.se

ABSTRACT

Mediverse aims to bridge clinical and medical research and combine medical knowledge into a single graph that can be used to perform deep exploration and predictions across diverse modalities (e.g., genomic, clinical studies, medical, etc.). In particular, Mediverse intends to establish trust in medical data systems and exploit the benefits of graph representation learning methods, which are capable of addressing multiple key challenges that harm digitalisation in healthcare and medicine sectors today, namely: multi-modal data integration across clinics, complex search, and uncertainty-aware predictive computing.

VLDB Workshop Reference Format:

Andra Ionescu, Paris Carbone, Sebastiaan Meijer, and Jayanth Raghothama. Mediverse: Multimodal Clinical Exploration and Search on a Single Graph. VLDB 2026 Workshop: Biomedical Data Management Systems (BioDMS).

1 INTRODUCTION

Interoperability in healthcare has been developed from the bottom-up, layer by layer. It started with the development of exchange mechanisms (e.g., HL7 v2 [10]), which enabled hospitals to share structured clinical events. Next, terminology systems such as SNOMED-CT [7], LOINC [12], and much older ICD [14] were developed, providing standardised vocabularies to encode clinical concepts. However, even with shared transport and shared vocabularies, the structure of clinical information remained inconsistent. Information models, such as openEHR [19], responded to this gap, promising to solve interoperability by separating the technical layer (i.e., how data are stored) from the clinical layer (i.e., what the data tell us).

The model-first philosophy and the strict standards required significant upfront resources for modelling and governance, which only a few health systems could afford [1]. Because each layer was standardised independently, no unified information model was agreed upon. Each implementation imposed its own structure at the intersection of these standards, and we could see that hospitals

adopting the same standards still produced semantically incompatible datasets [9, 25]. This fragmentation emerged from the architecture of the bottom-up approach, a consequence that no single standard can undo.

Despite these flaws, the bottom-up approach contributed to solutions and advancements that solved technical and syntactic interoperability, which are two of the four levels of interoperability in healthcare: technical, syntactic, semantic, and organisational [17]. We are currently addressing the challenges of semantic interoperability, with the goal of enabling and solving organisational interoperability.

As such, we support a top-down approach, where rather than standardising exchange and hoping semantic agreement follows, the approach begins by defining a shared information model and derives storage, terminology, and exchange from it. Initiatives such as the OMOP Common Data Model (CDM) [24] represent partial implementations, providing a target for interoperability through data harmonisation.

Data harmonisation resolves heterogeneity along three dimensions [6]: syntactic, structural and semantic. With the syntactic dimension, different data formats are reconciled. The structural dimension addresses the different representations of the content. For example, information about a patient over three days is represented as a single row with `start_date` and `end_date` columns spanning three days, or as three rows with a date column indicating the date for each day. In the last dimension, data with the same meaning have to be harmonised. For example, data on young adults can be collected for different age groups (e.g., 18-25 years vs. 18-30 years) or it can have different meanings (e.g., young adults, teenagers, etc.).

Currently, data harmonisation in healthcare is dominated by manual ETL pipelines that map the source data to a shared target schema and vocabulary, thus supporting the syntactic and structural dimensions. The persistent bottleneck is semantic alignment, the process of establishing correspondences between local terms and standardised vocabularies. This corresponds to the semantic dimension, which still requires human reviewers for validation.

In Mediverse, we embrace the semantic heterogeneity of healthcare, where clinical data vary between specialties, institutions, and regions. We address this diversity effectively using graph representation learning, leveraging medical ontologies and cross-ontology mappings to generate semantic links between disparate data fields, thus enabling a harmonised view over the fragmented datasets.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, ISSN 2150-8097.

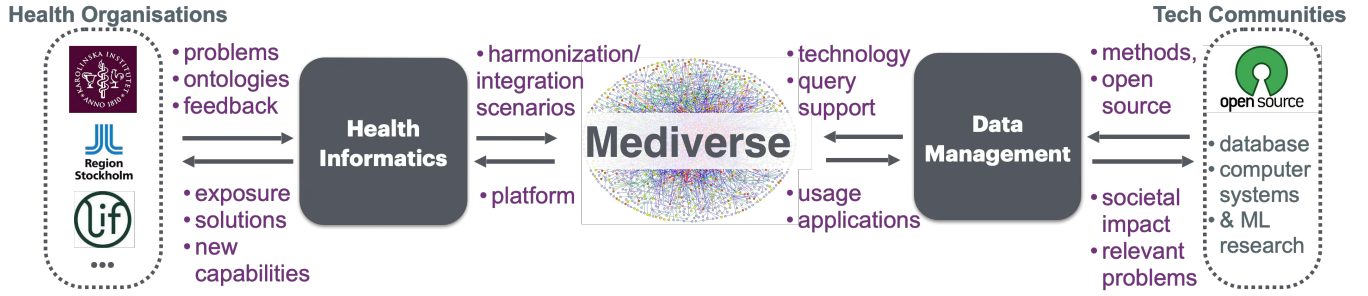


Figure 1: Bridging communities: actors, tasks, benefits and interactions

2 MEDIVERSE IN A NUTSHELL

Figure 1 encapsulates our vision of bridging two communities, health informatics and data management. Mediverse offers to health informatics a platform developed with data management support, exposure to solutions, technology, query systems, open-source methods, and databases. The data management community will be exposed to the problems and complexities of healthcare, get access to real datasets, and contribute to a direct societal impact by solving these problems. Moreover, clinicians contribute with domain knowledge and can offer the environment to deploy and test new solutions. As such, Mediverse aims to serve as a vehicle to bring the latest advancements in representation learning and knowledge graphs to the forefront of medical and clinical practice.

Our goal is to establish a reference platform for achieving data and metadata integration between healthcare providers and medical researchers. We leverage knowledge graphs for seamless exploration and interaction between semantically diverse elements, such as genomic sequences, clinical and medical imaging records. Graph neural networks (GNNs) extend this by capturing multi-hop reasoning paths that reflect causal and taxonomic relationships in clinical ontologies [11, 20]. With GNNs, we showcase the capabilities of predictive exploration in healthcare, revealing insights that would otherwise remain hidden. GNN-powered inference will equip clinicians with the ability to anticipate patient outcomes, outline potential treatment options, and accelerate medical discovery. Mediverse will further incorporate black-box uncertainty estimation methods in graph representation learning tasks, to address accuracy, which remains a crucial barrier to applying inference and neural networks in healthcare. This will ensure that clinicians and researchers can rank predicted outputs using statistically accurate uncertainty bounds.

Furthermore, by democratising the use of responsible AI in healthcare, we aim to attract the attention of clinicians and biomedical researchers to emerging technologies, eliminating data integration concerns that typically take years to solve, while enhancing digitalisation in healthcare, productivity, and trust by shifting focus to the healthcare problems that matter.

3 MULTI-MODAL INTEGRATION

There are many standards that facilitate both the organisation of healthcare data (openEHR, SNOMED-CT, LOINC), as well as their exchange (i2b2, HL7 FHIR) [9, 29]. However, the heterogeneity of

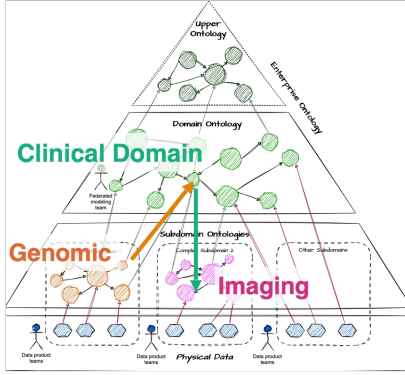
these standards and their partial coverage across sites and institutions [9] is one of the core challenges for data harmonisation. Figure 2 illustrates how common concepts across different domains, or across different ontologies of the same domain, can be linked with each other.

With this project, we aim to get a step closer to solving semantic interoperability, which is defined as “the ability of two or more systems or components to exchange information and to use the information that has been exchanged” [9]. As such, we work on the core component of semantic interoperability, semantic harmonisation, defined as “the process of combining multiple sources and representations of data into a form where items of data share meaning” [8]. Unlike conventional approaches, we investigate semantic harmonisation *without* transforming the data into a common format, since doing so is prone to information loss and requires a predefined target schema. Instead, to enable meaningful integration across diverse knowledge domains, representation formats, and data models, Mediverse leverages knowledge graphs to create flexible, standard-agnostic interoperability approaches.

We achieve this by developing a single graph representation of (medical) ontologies. These ontologies are open-source, many are available as graphs, and nearly all of them have bindings (i.e., links between concepts across ontologies) with each other. These representations and bindings are imported into a graph database to provide a single knowledge base of medical/clinical concepts. As illustrated in Figure 2, we want to query data using a single query language (e.g., Cypher) and the expressiveness of graph databases.

Our first use case connects Electronic Health Records (EHR) data with ontologies. We propose a semi-automatic approach for concept normalisation, a sub-problem of semantic alignment, which links individual data values from EHR to concepts in a standardised vocabulary. EHR data comes in multiple formats, the most common being free text (i.e., medical notes) and tabular values (i.e., tables with patient records, admissions, prescriptions, etc). Semantic alignment in EHR data requires integrating contextual, structural, and domain-specific signals to ensure correct interpretation across heterogeneous data sources. Thus, the type of EHR data determines the approaches we need for integration. Free text carries richer context, thus we apply biomedical entity linking solutions. Tabular values carry column/row context, thus we apply standard schema matching techniques [21] alongside solutions from entity reconciliation. Once the entity boundaries are known, we perform concept normalisation.

Ontological Data Harmonisation



```

MATCH (patient:Patient)-[:HAS_GENE]
->(gene:GenomicFeature { ID:'MGMT'})
MATCH (gene)-[:ASSOCIATED_WITH]->(effect:DrugResistance)
MATCH (patient)-[:RECEIVED_TREATMENT]
->(therapy:Treatment { name:'Targeted Therapy'})
MATCH (patient)-[:UNDERWENT_IMAGING]->(mri:MRI_Scan)
WHERE mri.result = 'Tumor Shrinkage'

```

Figure 2: Harmonisation using a single graph of multi-modal data and a sample Cypher query

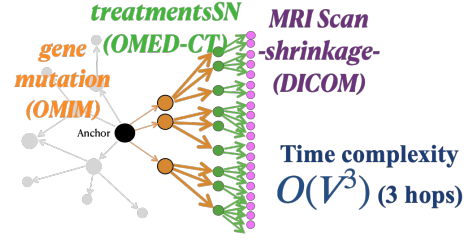
4 UNCERTAINTY AWARE SEARCH AND EXPLORATION

Graphs are highly generalised models for data and their relationships. In the medical sector, a graph can represent drug and patient vertices connected with respective edges representing interactions (drug-drug) or reactions (patient-drug). Other graphs can connect diseases, genetic mutations, or patients with clinical trials and outcomes. When graphs grow, they tend to follow certain trends such as becoming sparse (i.e., with missing connections) and following a typical power law distribution of connections. In graph theory, these are also known as natural graphs.

Traversing natural graphs to detect certain patterns is problematic due to the disproportional amount of computing power necessary to reach abundant information, a corresponding search space also known to follow “non-Euclidean” geometry. To address these issues, the field of graph representation learning studies techniques that can efficiently compress graph information into fixed, lower-dimensional Euclidean spaces such as embeddings [13]. More particularly, GNNs have already showcased breakthroughs in fields such as protein interaction [18] and physics simulations [28]. Embeddings are also shown to be capable of learning computational relations within graphs.

We reframe semantic harmonisation as a graph alignment problem rather than a schema transformation one. We will experiment with both supervised and unsupervised methods, such as EMBERT (Entity-rich Medical BERT) [3], FINAL (Fast Attributed Network Alignment) [30], SEU (Simple but Effective Unsupervised entity alignment) [23], CONE-Align [5], investigate and showcase the domain-efficacy of these models with a focus on multimodal medical/clinical data.

Multi-modal Graph Query Search



Predictive Search via Representation-Learning (Query2Box)

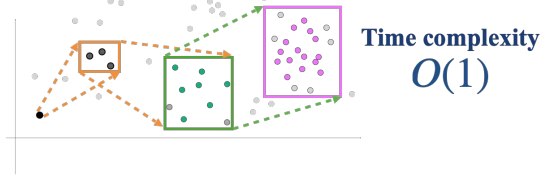


Figure 3: Graph query search and predictive execution

Figure 2 shows a 3-hop Cypher query that searches through genomic, pharmacological, and clinical ontologies. On a medical dataset, such as AIMedGraph [26] which has millions of connections, this computation could take up to a minute, since the algorithmic graph search has $O(V \times h)$ complexity (where V is the number of vertices, and h represents the steps/hops). In contrast, using Query2Box [27] on fixed dimensions reduces this to $O(1)$, drastically speeding up the process to second, or sub-second latency, while also detecting hidden links in the data. Figure 3 illustrates the search complexity using algorithmic graph traversals compared to representation learning.

Predictability in query processing was explored in the context of Approximate Query Processing (AQP) [4], an early attempt based on data sampling to speed up query responses in SQL databases at the expense of inaccurate output. BlinkDB [2] is a known system that featured error bounds but could not achieve widespread adoption due to its inability to generate new missing connections in the data and diminishing returns in speed over high accuracy penalties. More recent research in data management explores the use of ML [22], yet, their focus is on optimisation purposes, such as learning indices, and not to enrich data exploration.

Within an interoperability pipeline, conformalised link prediction can offer a concrete mechanism to quantify and user-bound uncertainties. Systems such as OrbDB [15] and confidence calibrated vector search engines (e.g. Conann [16]) can use conformal scores to recalibrate similarity results and predictions so that the reported confidence matches observed correctness. Thus, we will explore the use of black-box uncertainty estimators (e.g., conformal and Venn predictors) to estimate uncertainties with statistical bounds on graph representation learning models. The system will estimate and display the uncertainty of the computed outputs, along with the average uncertainty of the entire query.

5 CONCLUSION

Mediverse addresses a long-standing barrier in semantic interoperability by reframing harmonisation as a graph alignment problem rather than a schema transformation one. Through a unified graph representation of medical ontologies and open clinical datasets, combined with uncertainty-aware graph learning, the project lays the groundwork for interoperability approaches that are flexible, standard-agnostic, and robust under ambiguity. Ultimately, Mediverse aspires to make heterogeneous clinical data meaningfully usable across institutions, unlocking new opportunities for both primary care and secondary research.

6 AUTHORS

Andra Ionescu is a Postdoctoral Researcher at KTH Royal Institute of Technology, in the Department of Biomedical Engineering and Health Systems. Her current work focuses on graph-based clinical data harmonisation and semantic interoperability challenges.

Paris Carbone is Associate Professor at KTH Royal Institute of Technology and director of the Data Systems Lab. He is working on scalable data processing systems, generative computing, data streaming, and graph databases. He is a co-recipient of the ACM SIGMOD Systems Award 2023 (Apache Flink).

Sebastian Meijer is Professor of Health Care Logistics at KTH Royal Institute of Technology. He is working on healthcare, health prevention and promotion systems, and is the coordinator of the European Digital Innovation Hub Health Data Sweden.

Jayanth Raghothama is Associate Professor at KTH Royal Institute of Technology, in the Department of Biomedical Engineering and Health Systems. His research interests include complex systems, health informatics, policy design, and applied ML.

ACKNOWLEDGMENTS

This work was partially funded by Digital Futures.

REFERENCES

- [1] Kola Adegoke, Abimbola Adegoke, Deborah Dawodu, Akorede Adekoya, Ayoola Bayowa, Temitope Kayode, and Mallika Singh. 2025. Interoperability as a Catalyst for Digital Health and Therapeutics: A Scoping Review of Emerging Technologies and Standards (2015–2025). *International Journal of Environmental Research and Public Health* 22, 10 (2025), 1535.
- [2] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. 2013. BlinkDB: queries with bounded errors and bounded response times on very large data. In *Proceedings of the 8th ACM European conference on computer systems*. 29–42.
- [3] Zerui Cai, Taolin Zhang, Chengyu Wang, and Xiaofeng He. 2021. Embert: A pre-trained language model for chinese medical text mining. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 242–257.
- [4] Surajit Chaudhuri, Bolin Ding, and Srikanth Kandula. 2017. Approximate Query Processing: No Silver Bullet. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) (SIGMOD '17). Association for Computing Machinery, New York, NY, USA, 511–519.
- [5] Xiuyan Chen, Mark Heimann, Fatemeh Vahedian, and Danai Koutra. 2020. Cone-align: Consistent network alignment with proximity-preserving node embedding. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1985–1988.
- [6] Cindy Cheng, Luca Messerschmidt, Isaac Bravo, Marco Waldbauer, Rohan Bhavikatti, Caress Schenk, Vanja Grujic, Tim Model, Robert Kubinec, and Joan Barceló. 2024. A general primer for data harmonization. *Scientific data* 11, 1 (2024), 152.
- [7] Roger A Cote and Stanley Robboy. 1980. Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED). *Jama* 243, 8 (1980), 756–762.
- [8] James A. Cunningham, Michel Van Speybroeck, Dipak Kalra, and Rudi Verbeek. 2017. Nine Principles of Semantic Harmonization. *AMIA Annual Symposium Proceedings* 2016 (Feb. 2017), 451–459.
- [9] Blanda Helena De Mello, Sandro José Rigo, Cristiano André Da Costa, Rodrigo da Rosa Righi, Bruna Donida, Marta Rosecler Bez, and Luana Carina Schunke. 2022. Semantic interoperability in health records standards: a systematic literature review. *Health and technology* 12, 2 (2022), 255–272.
- [10] Robert H Dolin, Liora Alschuler, Calvin Beebe, Paul V Biron, Sandra Lee Boyer, Daniel Essin, Elliot Kimber, Tom Lincoln, and John E Mattison. 2001. The HL7 clinical document architecture. *Journal of the American Medical Informatics Association* 8, 6 (2001), 552–569.
- [11] Zhengxiao Du, Chang Zhou, Jiangchao Yao, Teng Tu, Letian Cheng, Hongxia Yang, Jingren Zhou, and Jie Tang. 2021. CogKR: Cognitive graph for multi-hop knowledge reasoning. *IEEE Transactions on Knowledge and Data Engineering* 35, 2 (2021), 1283–1295.
- [12] Arden W Forrey, Clement J Mcdonald, Georges DeMoor, Stanley M Huff, Dennis Leavelle, Diane Leland, Tom Fiers, Linda Charles, Brian Griffin, Frank Stalling, et al. 1996. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical chemistry* 42, 1 (1996), 81–90.
- [13] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).
- [14] JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. 2016. ICD-10: history and context. *American Journal of Neuroradiology* 37, 4 (2016), 596–599.
- [15] Sonia Horchidan and Paris Carbone. 2023. ORB: Empowering Graph Queries through Inference.. In *ESWC Workshops*.
- [16] Sonia Horchidan, Fabian Zeiher, Henrik Boström, and Paris Carbone. 2026. Co-nANN: Conformal Approximate Nearest Neighbor Search. *Proceedings of the VLDB Endowment* 19 (2026).
- [17] The Healthcare Information and Management Systems Society (HIMSS). 2026. *Interoperability in Healthcare*. <https://nationalcapitalarea.himss.org/resources/interoperability-healthcare>
- [18] Kanchan Jha, Sriparna Saha, and Hiteshi Singh. 2022. Prediction of protein-protein interaction using graph neural networks. *Scientific Reports* 12, 1 (2022), 8360.
- [19] Dipak Kalra, Thomas Beale, and Sam Heard. 2005. The openEHR foundation. *Studies in health technology and informatics* 115 (2005), 153–173.
- [20] Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. 2025. Biohopr: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. In *Findings of the Association for Computational Linguistics: ACL 2025*. 12894–12908.
- [21] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. 2021. Valentine: Evaluating matching techniques for dataset discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 468–479.
- [22] Tim Kraska, Mohammad Alizadeh, Alex Beutel, Ed H Chi, Jialin Ding, Ani Kristo, Guillaume Leclerc, Samuel Madden, Hongzi Mao, and Vikram Nathan. 2019. SageDB: A learned database system. In *9th Biennial Conference on Innovative Data Systems Research, CIDR 2019*.
- [23] Xinnian Mao, Wen Ting Wang, Yuanbin Wu, and Man Lan. 2021. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2843–2853.
- [24] OHDSI Community. 2024. Standardized Data: The OMOP Common Data Model. <https://www.ohdsi.org/data-standardization/>. Accessed 2025-12-03.
- [25] Sari Palojoki, Lasse Lehtonen, and Riikka Vuokko. 2024. Semantic interoperability of electronic health records: Systematic review of alternative approaches for enhancing patient information availability. *JMIR medical informatics* 12, 1 (2024), e53535.
- [26] Xueping Quan, Weijing Cai, Chenghang Xi, Chunxiao Wang, and Linghua Yan. 2023. AIMedGraph: a comprehensive multi-relational knowledge graph for precision medicine. *Database* 2023 (2023), baad006.
- [27] H Ren, W Hu, and J Leskovec. 2020. Query2box: Reasoning Over Knowledge Graphs In Vector Space Using Box Embeddings. In *International Conference on Learning Representations (ICLR)*.
- [28] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. Learning to simulate complex physics with graph networks. In *International conference on machine learning*. PMLR, 8459–8468.
- [29] Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, and Michael Zakharyashev. 2018. Ontology-Based Data Access: A Survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 5511–5519.
- [30] Si Zhang and Hanghang Tong. 2016. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD*. 1345–1354.