Paper PP10

Introducing Graph Structure in Real-World Data Analyses

Kountouris Nikolaos, **Pfizer Inc.,** Athens, Greece
Panagiotakis Georgios, **Pfizer Inc.,** Athens, Greece
Puri Maria, **Pfizer Inc.,** Athens, Greece

ABSTRACT

In the realm of pharmaceuticals, real-world data (RWD) comprises an expansive and intricate web of information drawn from diverse sources. These include electronic health records, insurance claims, and patient registries. Traditional analytical approaches often fall short when attempting to fully harness the potential of this interconnected data landscape. Graph theory, with its emphasis on relational structures, emerges as a promising framework to unlock the full value of RWD. However, the practical applications of this theoretical approach in the pharmaceutical sector remain largely unexplored, presenting both opportunities and challenges for researchers and industry professionals alike. Graph theory’s strength lies in its unique ability to map out and examine complex relationships in health data. This approach can shed light on common health patterns and key intervention moments that often slip through the cracks of traditional analysis. By applying graph algorithms, researchers might uncover subtle trends within patient groups, potentially leading to more precise patient categorization for drug trials and tailored treatment plans. Moreover, viewing health data as an interconnected web, rather than isolated facts, could offer fresh perspectives on how diseases progress, how well treatments perform, and how healthcare systems truly function. This shift in viewpoint might spark innovative insights that were previously out of reach. Our poster serves as a primer on graph theory for those in the pharmaceutical and healthcare fields. We break down the concepts underpinning this approach and explore theoretical applications in the realm of real-world data. The goal is to illuminate how this powerful tool could reshape our understanding and use of complex health information.

1. INTRODUCTION

The healthcare landscape is undergoing a seismic shift, driven by an explosion of real-world data (RWD). RWD encompasses information about patient health and healthcare delivery collected from a variety of sources. This data deluge stems from the widespread adoption of digital technologies, including electronic health records (EHRs), claims information, wearable devices, social media, and other tech-driven services. The sheer volume and diversity of RWD sources have created unprecedented opportunities for gaining insights into patient health, treatment efficacy, and healthcare system performance. From claims databases to disease registries, these data streams paint a comprehensive picture of health trends and outcomes. This wealth of information promises to revolutionize healthcare delivery, drug development, and health policy. However, it’s not without its challenges. The complex, interconnected nature of RWD poses significant hurdles for traditional analytical methods, which were not designed to handle such vast and varied datasets. Current approaches to RWD analysis, primarily based on relational database models, struggle to capture and analyze the multidimensional relationships inherent in healthcare data. These methods often require complex joins and sub-queries to answer even relatively straightforward questions about patient journeys or treatment patterns, leading to performance issues and limiting the depth of insights that can be extracted. For example, tracking a patient’s journey through multiple care settings, treatments, and outcomes often requires navigating through numerous tables in a relational database, each join potentially degrading query performance. Moreover, identifying patterns across large patient populations or detecting subtle relationships between treatments and outcomes becomes increasingly challenging as the complexity of the data grows. Graph theory, a branch of mathematics that studies the properties of networks, offers a promising alternative approach to RWD analysis. By representing entities (such as patients, treatments, or healthcare providers) as nodes and their relationships as edges, graph models can naturally capture the complex interconnections within healthcare data. This representation allows for more intuitive querying and potentially more powerful analytical techniques. The objective of this poster is to provide an introduction into graph theory and explore the potential theoretical applications in enhancing RWD analysis within healthcare and pharmaceutical contexts.

2. THEORETICAL CONTEXT

2.1 Introduction to Graph Theory

Graph theory provides a powerful framework for representing and analyzing complex, interconnected data. A graph structure consists of sets of vertices/nodes representing entities and sets of edges representing relationships between the entities. In the context of RWD a graph could be visualized by having healthcare entities (e.g. patients, treatments, diagnoses) mapped as vertices and edges to be interactions within them (e.g., Patient -> "diagnosed with,"-> Diagnosis). There are different types of graphs with the most common to be:

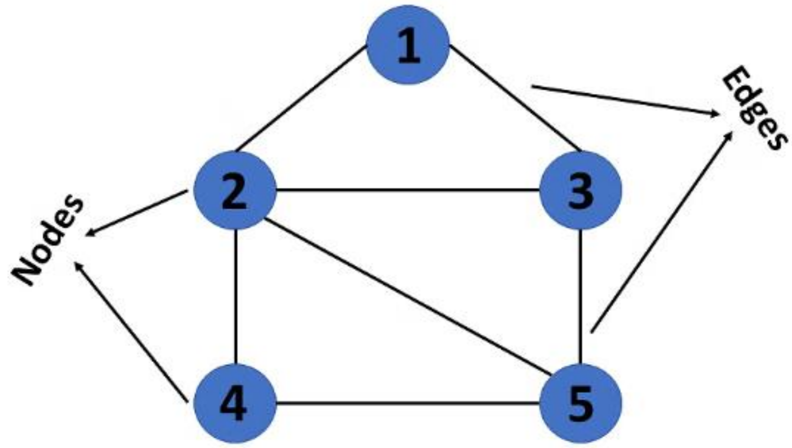

Figure 1: Example of a Graph with vertices and edges.
https://www.simplilearn.com/tutorials/data-structure-tutorial/graphs-in-data-structure

- **Undirected Graphs:** Edges do not have a direction.
- **Directed Graphs:** Edges having a uni-direction.
- **Weighted Graph:** Edges having associated weights or costs.
- **Directed Acyclic Graph (DAG):** A directed graph with no cycles. Meaning that always a node will exist that does not traverse to another node, resulting to not complete the graph.
- **Complete Graphs:** Every node is connected to every other node.

  When we are dealing with graphs there are several key concepts/ algorithms that often been used and could apply also in RWD with some of them to be:

-
- **Node centrality:** Measures of a node's importance in the network. There are several types of centralities:
  - **Degree centrality:** the number of a node's links to and from other nodes.
  - **Betweenness centrality:** which nodes are "bridges" between nodes in a network.
  - **Closeness centrality:** How close a node is to all other nodes in the network.
  - **Eigenvector centrality:** A measure of node influence within the network.
- **Path analysis:** Examination of sequences of connections between nodes. e.g., shortest path analysis.
- **Community detection:** Identification of clusters or subgroups within the network.
- **Modularity-based methods:** Algorithms that optimize modularity, a measure of the strength of division of a network into communities.

2.2 Characteristics of Real-World Data in Healthcare

Real-world data (RWD) exhibits distinct characteristics when compared to data gathered from clinical trials. The observational nature of RWD sets it apart, marked by significant heterogeneity due to its collection from a wide array of sources. These sources yield data with diverse structures, encompassing both structured and unstructured formats.
Types of structured data could be:

- **Electronic Health Records (EHRs):** Quantifiable data including laboratory results, vital signs, medical treatments, and diagnosed conditions.
- **Claims data:** Organized records detailing medical diagnoses, performed procedures, dispensed medications, and associated expenses.
- **Clinical trial data:** Systematic data gathered in carefully managed research environments.

Types of unstructured data could be as well:

- **Raw Text:** Collected from clinical notes often taken in conjunction with EHR.
- **Images:** MRI scans, X-ray, PET scans etc.
- **Wearable devices:** Data from wearable devices like smartwatches, or specialized medical tests such as polysomnography, falls into this category.

RWD is characterized by its massive volume compared to clinical trial data, a result of real-world variability. A single patient’s record may encompass thousands of data points across multiple dimensions, including demographics, diagnoses, medications, procedures, and lab results. Another key feature is the intricate web of relationships in RWD, such as links between patients and treatments, treatment outcomes, and provider-patient interactions. This data abundance, however, brings challenges. Incompleteness and noise are common issues, manifesting as missing information due to gaps in care or documentation, data entry errors, coding mistakes, and inconsistent clinical practices across providers. These factors present significant hurdles for conventional data analysis methods, especially those relying on relational database models.

2.3 Comparative Analysis: Graph vs Relational Models

In this section we will examine the differences between graphs and relational models. It is worth mentioning that neither option is perfect. Each approach has its advantages and is more suitable for certain situations rather than one being inherently better than the other. Exploring applications across scenarios is essential when considering the utilization of these models, in real world settings.

2.3.1 Data Representation

A significant difference between these two approaches is that relational models organize data into tables with predefined schema, making it challenging to incorporate new types of data or relationships, without significant schema modifications. On the other hand, graph theory represents data as nodes and edges offering a more flexible structure, that can accommodate new types of entities and relationships without requiring schema changes. This flexibility is particularly valuable in healthcare, where new treatments, diagnostic criteria, or types of data may merge rapidly. The representation of the relationships between these two methods varies significantly. In the relational models, relationships are implied with the use of Foreign Keys constraints and the use of joins between tables. In this way complex relationships most of the time require multiple joins which are more expensive computationally and adds an extra cost on demand databases. On the other hand, graph models explicitly represent relationships as edges between nodes and storing them as first-class citizens in the form of edges between nodes, meaning the relationships are directly accessible and do not need to be inferred through key matches. This allows for more natural representation of complex healthcare networks.

2.3.2 Query Performance

For certain applications, the query building can become more complex on relational databases than in graphs databases. For example, in RWD for finding patients that received a specific sequence of treatments, in graph databases might be expressed as a simple path query due to the flexibility of the graph models to traverse relationships/edges more intuitively often resulting in simpler query expressions. In contrast, in relational databases these types of queries might involve multiple relationships that often require multiple joins and sub-queries which can be difficult to optimize [1]. In the book "Neo4j in Action" the authors reported the results from a comparison between a graph database (neo4j) and a relational database (MySQL) that for a two-join operation Neo4j was 60% faster than MySQL and for a depth four query Neo4J was 1,135 times faster than MySQL [16].

2.3.3 Scalability

When it comes to scale the database to fit more data, usually relational models scale their databases system vertically which means to increase their system resources while graph databases scale their systems horizontally (employing more machines) which can be advantageous for large datasets by using partitioning [3].

2.3.4 Schema Upgrading

Relational databases, characterized by their rigid schema, offer strong data integrity guarantees through predefined structures and constraints. However, this rigidity comes at a cost. Schema changes in relational systems can be complex, time-consuming, and potentially disruptive, especially for large databases. The process often involves careful planning, possible downtime, and consideration of backward compatibility. Graph databases, in contrast, offer a more flexible approach to schema management. Many graph databases are schema-optional or schema-flexible, allowing for easy addition of new types of nodes, relationships, or properties on the fly. This flexibility facilitates rapid adaptation to changing data requirements and can significantly reduce the complexity of schema evolution. However, this flexibility is a double-edged sword. Without careful management, it can lead to data inconsistencies and challenges in maintaining data integrity across the system.

2.3.5 Standardization and Tooling

In the terms of standardization relational databases have an advantage due to the existence and widely use SQL (Structured Query Language) which is a well-established, standardized language for querying relational databases, also an extensive amount of tools regarding relational databases exist compared to the graph databases, tools such as Crystal Reports for reporting that integrate well with relational data structures, and other tools for administration such as MySQL workbench. On the other hand, graph databases lack a universally query language, although there are popular languages such as **Cypher** (language used in Neo4j database) and **Gremlin**, but still their syntax is different. Several tools are being developed to support graph models and databases. These tools also cover analysis and visualization (e.g., Gephi, GraphXR etc.)

2.3.6 Integration

Integration capabilities significantly differ between relational and graph databases in the healthcare context. Relational databases benefit from well-established ETL (Extract, Transform, Load) processes and tools, along with standard interfaces such as JDBC for application connectivity.

3. METHODOLOGY

3.1 Graph Construction from RWD Sources

The transformation of relational healthcare data into graph structures represents a paradigm shift in how we conceptualize and analyze complex medical information. At its core, this process involves redefining healthcare entities as nodes within a vast, interconnected network. Patients, healthcare providers, diagnoses, medications, procedures, and facilities become distinct node types, each carrying its own set of properties that encapsulate relevant attributes. The relationships between these entities are represented as edges, capturing multifaceted interactions within the healthcare ecosystem. These edges not only connect patients to their diagnoses and treatments but also represent the temporal sequence of medical events, the associations between different medical concepts, and the interactions between various healthcare stakeholders. The mapping of relational data to this graph structure follows a set of principled transformations. Tables in the relational model typically translate to node types in the graph, with each row becoming a distinct node. Foreign key relationships in the relational model naturally evolve into edges in the graph, while table attributes become properties of nodes or edges [11]. This mapping process must also account for the temporal nature of healthcare data, incorporating time-stamped events as temporal edges and implementing versioning of node properties to track changes over time. The process of building healthcare graphs from strategies shares a crucial preliminary step: the careful design and conceptualization of how the relational structure will translate into a graph structure. This initial planning phase is essential for ensuring that the resulting graph accurately represents the complex relationships inherent in healthcare data.

ETL-based Construction

1. Extract data from relational sources.
2. Transform it to fit the graph model.
3. Load it into a graph database.
4. Consider data cleansing, standardization, entity resolution, and relationship inference.
5. Optimize for handling large volumes of healthcare data.
6. Employ incremental extraction for ongoing updates.

In-memory construction

1. Extract data from relational sources and store them in-memory of the programming interface.
2. Transform the relational data to graph structure: To do this the most common is to use already implemented programming libraries that leverage parallel processing techniques and optimized memory management strategies to build and analyze graphs quickly.

Healthcare data presents unique challenges that must be addressed in the graph construction process. The integration of multi-modal data, including structured data, unstructured text from clinical notes, genomics data, and medical imaging data, requires sophisticated approaches to data harmonization and linking. Equally crucial is handling temporal data in healthcare, necessitating methods for accurately representing and analyzing event sequences, interval-based relationships, and complex temporal patterns. Another challenge is moving from SQL to graph query languages where optimizing queries for graph databases often require different strategies than for relational databases, as well the lack of a standardized graph query language makes it more challenging to interact with different graph databases. Finally, ensuring that the graph models comply with data governance policies and regulations requires careful planning, as well implementing access controls and audit trails may require new approaches

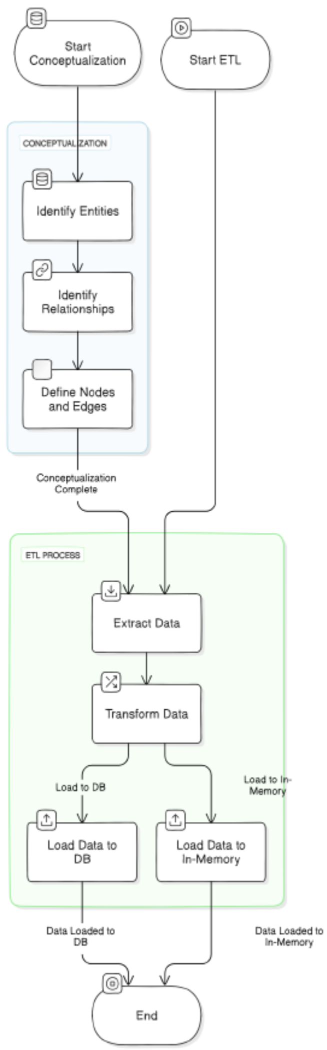

Figure 2: Representation of the process on how to build a healthcare graph from relational data.

4. THEORETICAL APPLICATIONS

Graph models may offer advantages over relational databases in several theoretical healthcare scenarios, including:

4. 1 Patient Journey Mapping

A patient's journey can be visualized as paths through a healthcare graph, offering a comprehensive view of the patient's interactions of healthcare system over time. This view can be useful on problems such as finding **common treatment sequences** e.g., discovering typical treatment pathways for specific conditions, for instance patients with type 2 Diabetes often follow a sequence of oral medication before transitioning to insulin therapy by applying subgraph mining algorithms to extract common patterns.

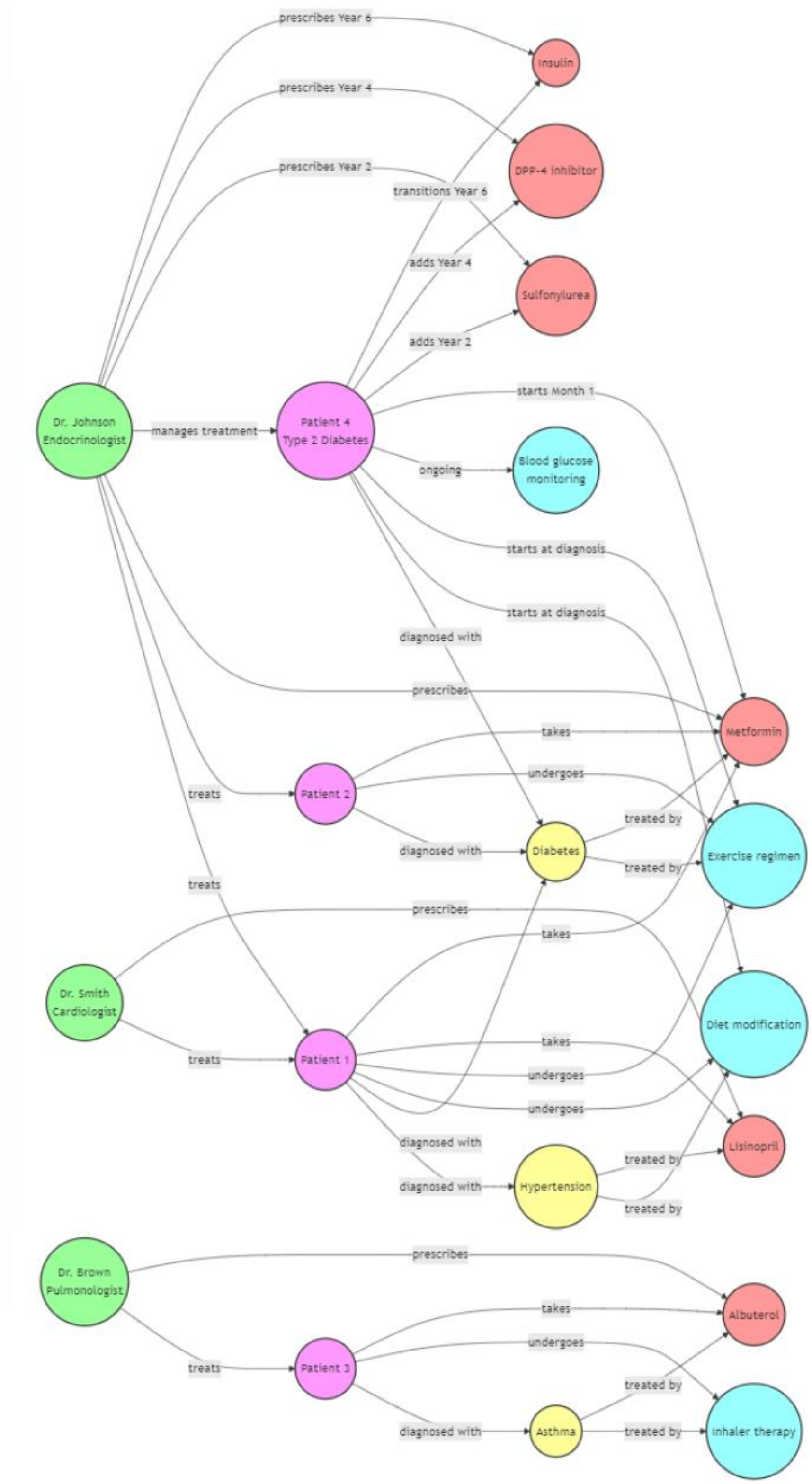

Figure 3: This graph illustrates the typical treatment pathway for a T2DM patient over time, as well as patients with other conditions. The circular nodes represent health states, medications, and ongoing management practices. Arrows (edges) indicate the progression of the disease (**T2DM**) and treatment intensification, with temporal labels showing the typical timing of each step.

4.2 Modeling Disease Progression Networks

Disease progression can be represented as a network of health states connected by transition probabilities, enabling the identification of key factors that influence certain disease progressions e.g., lifestyle factors or medications that have a significant influence on the progression of chronic diseases such as T2DM (Type 2 Diabetes Mellitus). Using centrality analysis could be helpful to identify biomarkers or factors that frequently appear on paths leading to complications.

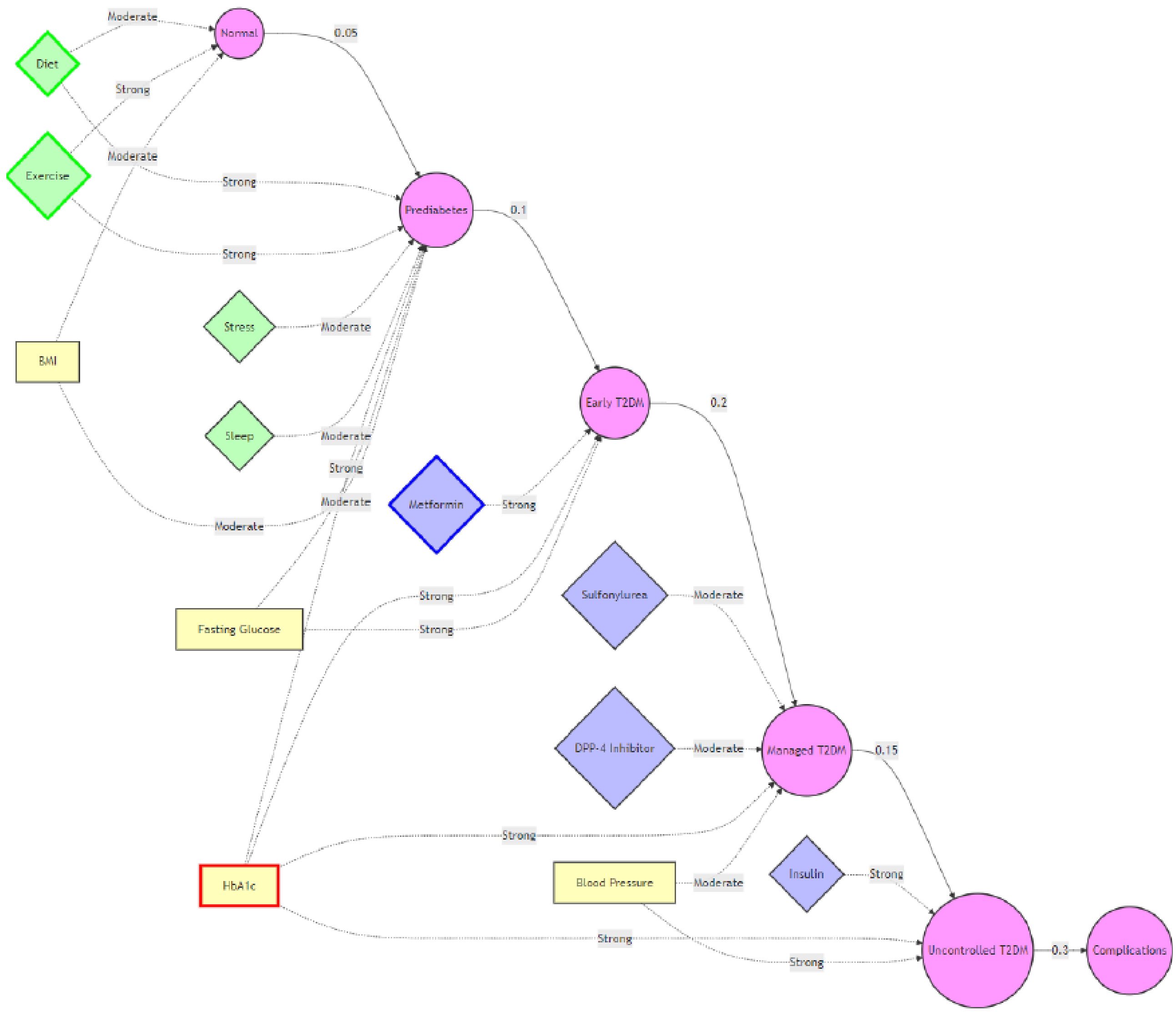

Figure 4: This graph illustrates the complex interplay between health states (pink circles), lifestyle factors (green hexagons), medications (blue hexagons), and biomarkers (yellow rectangles) in T2DM management. Solid lines between health states represent disease progression with **hypothetical** transition probabilities. Dotted lines show the influence of lifestyle factors and medications on various stages of T2DM, while dash-dot lines indicate the relevance of biomarkers. Highlighted nodes (HbA1c, Diet, Exercise, and Metformin) represent factors with high centrality in the network, potentially key in influencing disease trajectories.

4.3 Identifying Co-treatments and Comorbidities

Community detection algorithms are able to find clusters of patients who share similar diagnosis and treatment patterns. These clusters might reveal frequently co-occurring conditions and treatments that might not be apparent in traditional statistical analyses. For instance, in a study of chronic pain patients, graph analysis might reveal a significant cluster of patients with both fibromyalgia and depression, treated with a combination of pregabalin and SSRIs. Measures of centrality can help pinpoint diagnoses or treatments that often appear alongside other conditions or therapies Furthermore, by incorporating temporal data and using dynamic graph analysis, we can track how comorbidities and co-treatments evolve over time, potentially uncovering causal associations or treatment synergies.

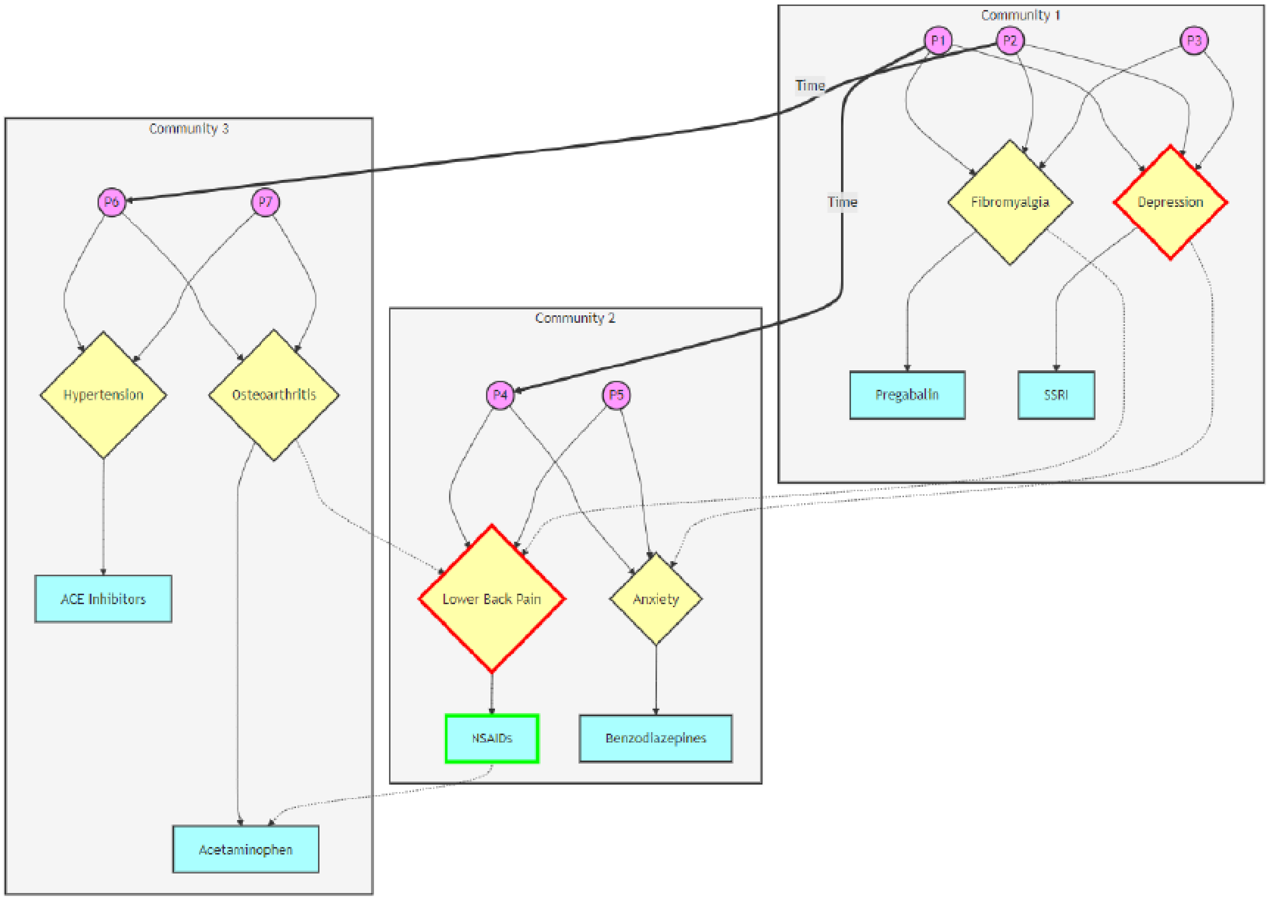

Figure 5: The graph illustrates three **hypothetical** distinct communities, each representing a cluster of patients with similar conditions and treatments. Within these clusters, we see patients grouped by shared diagnoses and therapeutic approaches. The solid black arrows originating from Patient 1 (P1) and Patient 2 (P2) in the Fibromyalgia and Depression community demonstrate how patients' conditions can evolve over time. These arrows show that P1 and P2, who initially presented with fibromyalgia and depression, later developed different conditions that resulted in their transition to separate communities. This temporal progression highlights the dynamic nature of chronic pain conditions and how patient profiles can shift between different symptom clusters as their health status changes over time. Last but not least, the highlighted red borders on the **lower back pain**, and **depression** entities highlight the high centrality of these nodes that frequently co-occur with other conditions across different communities.

It's important to recognize that these hypothetical instances highlight scenarios that just touch upon the possibilities of utilizing graph-based methods. There are uses waiting to be explored, each holding the promise of uncovering valuable insights that could go unnoticed when using traditional analytical techniques.

5. CONCLUSION & DISCUSSION

This poster explored the potential of graph theory in revolutionizing the analysis of real-world data (RWD) in healthcare and pharmaceutical research. We have presented a theoretical introduction about graph theory and how graph-based approaches can extract insights from complex, interconnected healthcare datasets. We examined their comparative advantages over traditional relational models and discussed potential applications across various domains of healthcare analytics. The key take away is the natural representation of complex relationships that graph models provide, with as well the improved performance for relationship-centric queries that involve traversing multiple relationships which is common in the RWD space. The adoption of graph-based approaches in healthcare RWD analysis has the potential to drive significant advancements in the healthcare field in topics such as precision medicine by enabling more sophisticated patient stratification and treatment response prediction, clinical decision support by providing a holistic view of patient data and treatment options that could support more informed clinical decision-making. Despite the promising potential graph-models offer, there are several challenges to be addressed, such as data quality and standardization where the heterogeneity and potential incompleteness of healthcare data sources pose significant challenges for constructing accurate and comprehensive healthcare graphs. Limitations in privacy where the interconnected nature of graph data raises important questions about data privacy, consent, and the potential for re-identification and many other limitations that may show up. The application of graph theory to healthcare RWD analysis represents a promising frontier in biomedical informatics and data science. Realizing its full potential will require a concerted effort from researchers across multiple disciplines, including computer science, statistics, medicine, and ethics. In conclusion, graph-based approaches offer a powerful new paradigm for analyzing and deriving insights from healthcare RWD. As we continue to generate larger and more complex healthcare datasets, graph theory may prove to be an essential tool in our quest to improve human health and healthcare delivery.

REFERENCES

[1] ALMABDY, S. Comparative analysis of relational and graph databases for social networks. In 2018 1st International Conference on Computer Applications Information Security (ICCAIS) (2018), pp. 1–4.

[2] ANDREW, D. Social network analysis 101: centrality measures explained. Accessed: 2024-09-24.

[3] AWS. What’s the difference between a graph database and a relational database? Accessed:2024-09-24.

[4] DAGLIATI, A., SACCHI, L., ZAMBELLI, A., TIBOLLO, V., PAVESI, L., HOLMES, J., ANDBELLAZZI, R. Temporal electronic phenotyping by mining care flows of breast cancer patients. Journal of Biomedical Informatics 66 (2017), 136–147.

[5] HANSEL, K., DUDGEON, S. N., CHEUNG, K.-H., DURANT, T. J., AND SCHULZ, W. L. From data to wisdom: Biomedical knowledge graphs for real-world data insights. Journal of Medical Systems 47, 1 (May 2023), 65.

[6] IAN ROBINSON, JIM WEBBER, E. E. Graph Databases, 2nd Edition. O’reilly, 2015.

[7] JENSEN, A. B., MOSELEY, P. L., OPREA, T. I., ELLESOE, S. G., ERIKSSON, R., SCHMOCK, H., JENSEN, P. B., JENSEN, L. J., AND BRUNAK, S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications5 (2014), 4022.

[8] KOTIRANTA, P., JUNKKARI, M., AND NUMMENMAA, J. Performance of graph and relational databases in complex queries. Applied Sciences 12 (06 2022), 6490.

[9] LIU, F., AND PANAGIOTAKOS, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology 22 (2022), 287.

[10] LIU, Z., WAN, G., PRAKASH, B. A., LAU, M. S. Y., AND JIN, W. A review of graph neural networks in epidemic modeling, 2024.

[11] NEO4J. Modeling: relational to graph. Accessed: 2024-09-24.

[12] PAVLOPOULOS, G. A., SECRIER, M., MOSCHOPOULOS, C. N., SOLDATOS, T. G., KOSSIDA, S., AERTS, J., SCHNEIDER, R., AND BAGOS, P. G. Using graph theory to analyze biological networks. Bio Data Mining 4 (2011), 10.

[13] SAJID, H. Knowledge graphs vs. relational databases: Everything you need to know, 2023.Accessed: 2024-09-24.

[14] SINGH, B. Graph data model and graph native storage processing. Accessed: 2024-08-10.

[15] VICKNAIR, C., MACIAS, M., ZHAO, Z., NAN, X., CHEN, Y., AND WILKINS, D. A comparison of a graph database and a relational database: A data provenance perspective. vol. 10, p. 42.

[16] VUKOTIC, A., NICKI WATT WITH TAREQ ABEDRABBO, D. F., AND PARTNER, J. Neo4j in Action. Mannning, 2014.

[17] YAOWEN, C. Comparison of Graph Databases and Relational Databases When Handling Large-Scale Social Data. Master thesis, University of Saskatchewan, 2016.

[18] ZHANG, Z., HE, P., YAO, H., JING, R., SUN, W., LU, P., XUE, Y., QI, J., CUI, B., CAO,M., AND NING, G. A network-based study reveals multimorbidity patterns in people with type 2 diabetes. iScience 26, 10 (Sep 2023), 107979.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Panagiotakis Georgios
Email: (Georgios.Panagiotakis@pfizer.com)

Co-Author Name: Kountouris Nikolaos
Email: (Nikolaos.Kountouris@pfizer.com)

Co-Author Name: Puri Maria
Email: (Maria.Puri@pfizer.com)