Medical Knowledge Harmonization: A Graph-based,Entity-Selective Approach to Multi-source Diagnoses

Bianchi Andrea, Di Marco Antinisca

Department of Information Engineering, Computer Science and Mathematics
University Of L'Aquila / Italy

Bianchi Andrea

- **Education:**

Master's Degree in Computer Science, University of L'Aquila, 2019

- **Current Position:**

Final Year PhD Student, Specializing in Bioinformatics

- **Research Interests:**

My research focuses on the field of bioinformatics, with a particular emphasis on the integration of various omics data. This interdisciplinary approach allows for a more comprehensive understanding of biological systems, enhancing our ability to decipher complex biological questions and challenges.

Outline

- Setting the context
- Methodology
- Experimental Setting
- Results
- Conclusions

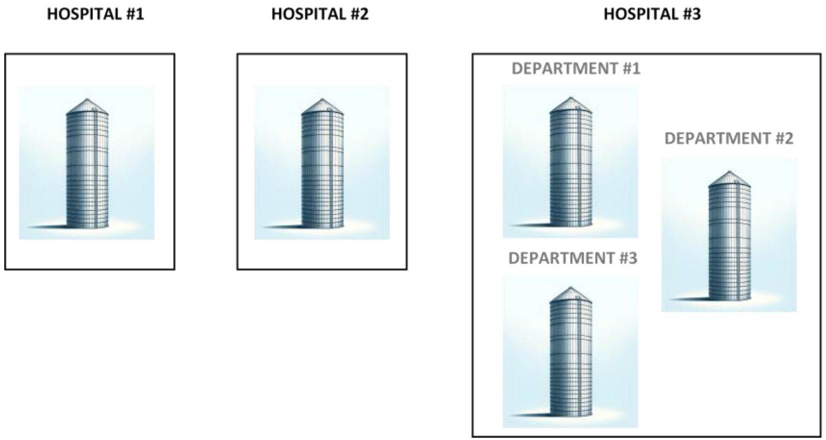
Setting the context

UNIVERSITÀ
DEGLI STUDI
DELL'AQUILA

DISIM
Dipartimento di Ingegneria
e Scienze dell'Informazione
e Matematica

Data Silos in the healthcare - 1

- Healthcare systems often operate in <u>silos</u>, with patient data spread across multiple platforms and departments.
- If several departments in a healthcare organization utilize distinct applications, data silos may potentially develop within the organization.
- Data about a patient who visits several healthcare providers will either be duplicated or hidden on the systems utilized by those providers.



Data Silos in the Healthcare - 2

- **Hindered Decision-Making:**
  Data silos obstruct the flow of comprehensive patient information, challenging healthcare providers in making well-informed, holistic decisions.

- **Increased Security Risks:**
  The risk of data breaches and privacy violations increases when data is dispersed throughout several systems, each with its own security measures.

- **Obstacles for Advanced Analytics:**
  The fragmentation of data hampers the effectiveness of advanced analytical tools like machine learning, leading to less accurate health insights.

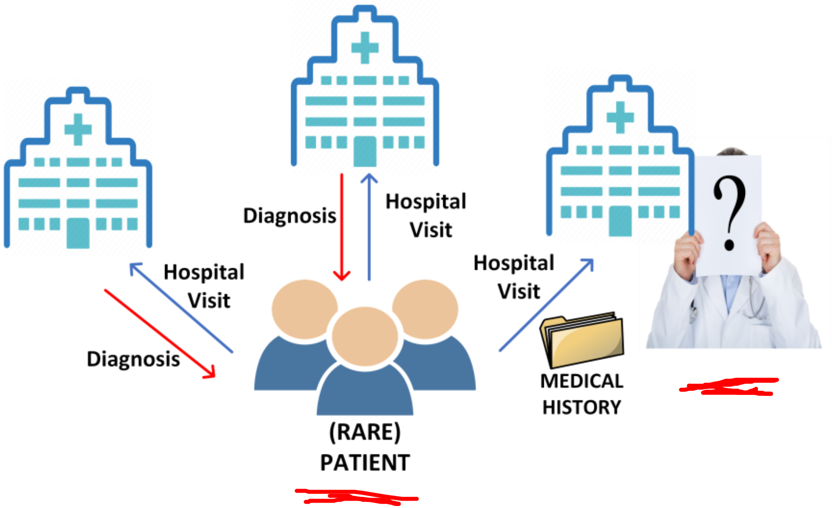
[1] P. S. Mathew and A. S. Pillai, "Big Data solutions in Healthcare: Problems and perspectives," *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2015, pp. 1-6, doi: 10.1109/ICIIECS.2015.7193211.

Scenario

- **Complex Patient Paths:**
  Patients, especially those with rare conditions, often visit multiple hospitals, receiving varied diagnoses. Each hospital visit generates its own set of records, leading to a scattered medical history.

- **The Challenge of Consolidation:**
  When these patients visit a new hospital, their entire medical history, often in extensive paper form, must be reviewed. This creates a daunting task for healthcare providers to assimilate and understand numerous documents, each with potentially crucial health information.

- **Impact on Diagnosis and Treatment:**
  This fragmented journey can lead to incomplete diagnoses and treatment plans, as vital information may be overlooked or underutilized due to the sheer volume and disorganization of the records.



The need for Integration

- A comprehensive understanding of patient health is essential (for doctors) for **efficient treatment and care**, but it is hampered by this fragmentation.

Solution: Graphs as a Unifying (Visual) Tool
Graph technology becomes an important tool in solving this problem. Graphs provide vital relationships and patterns between seemingly unrelated bits of health data, offering insights that are essential for comprehending complicated medical situations.

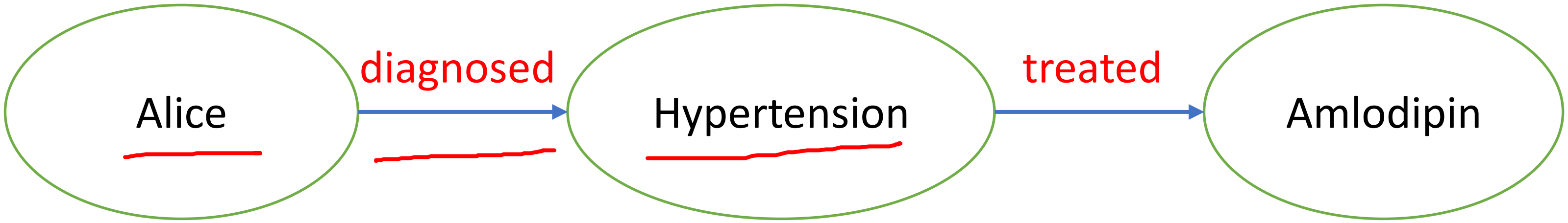
Knowledge Graphs (in Patient Data Management)

Addressing Data Fragmentation with Graphical Tools:

Traditional methods, reliant on sifting through extensive textual records, are often slow and error-prone.

What Are Knowledge Graphs?

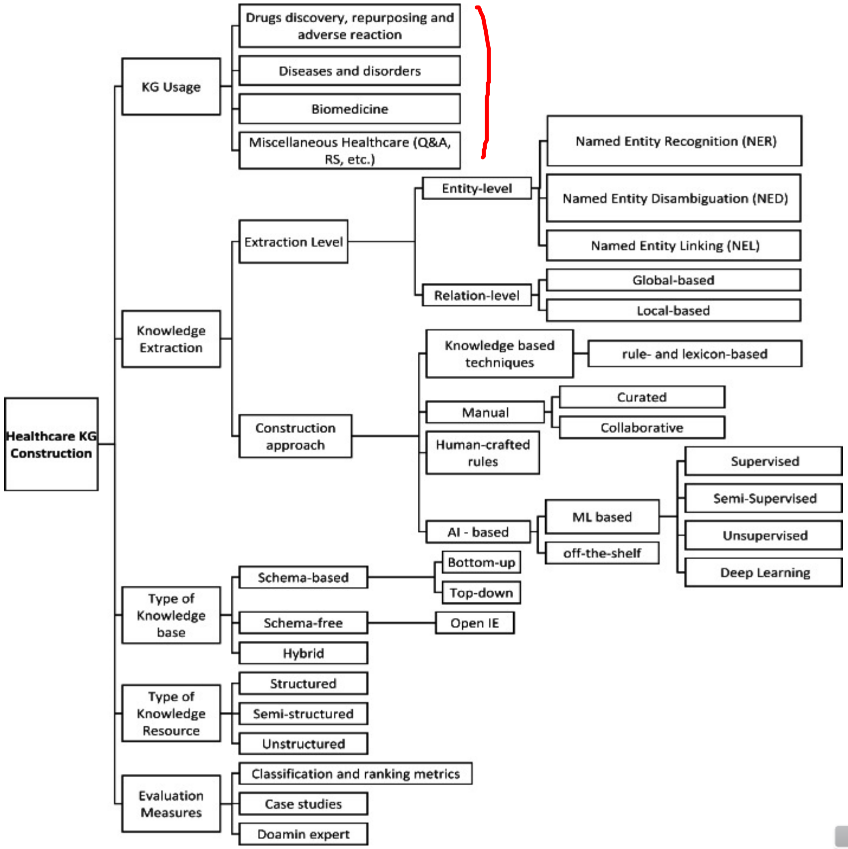
Knowledge graphs are advanced tools that represent data as interlinked nodes. These nodes are entities (like symptoms, diagnoses, treatments, …), and edges representing the relationships between them. This structure turns complex, disjointed information into a coherent, visually accessible network.



Knowledge Extraction

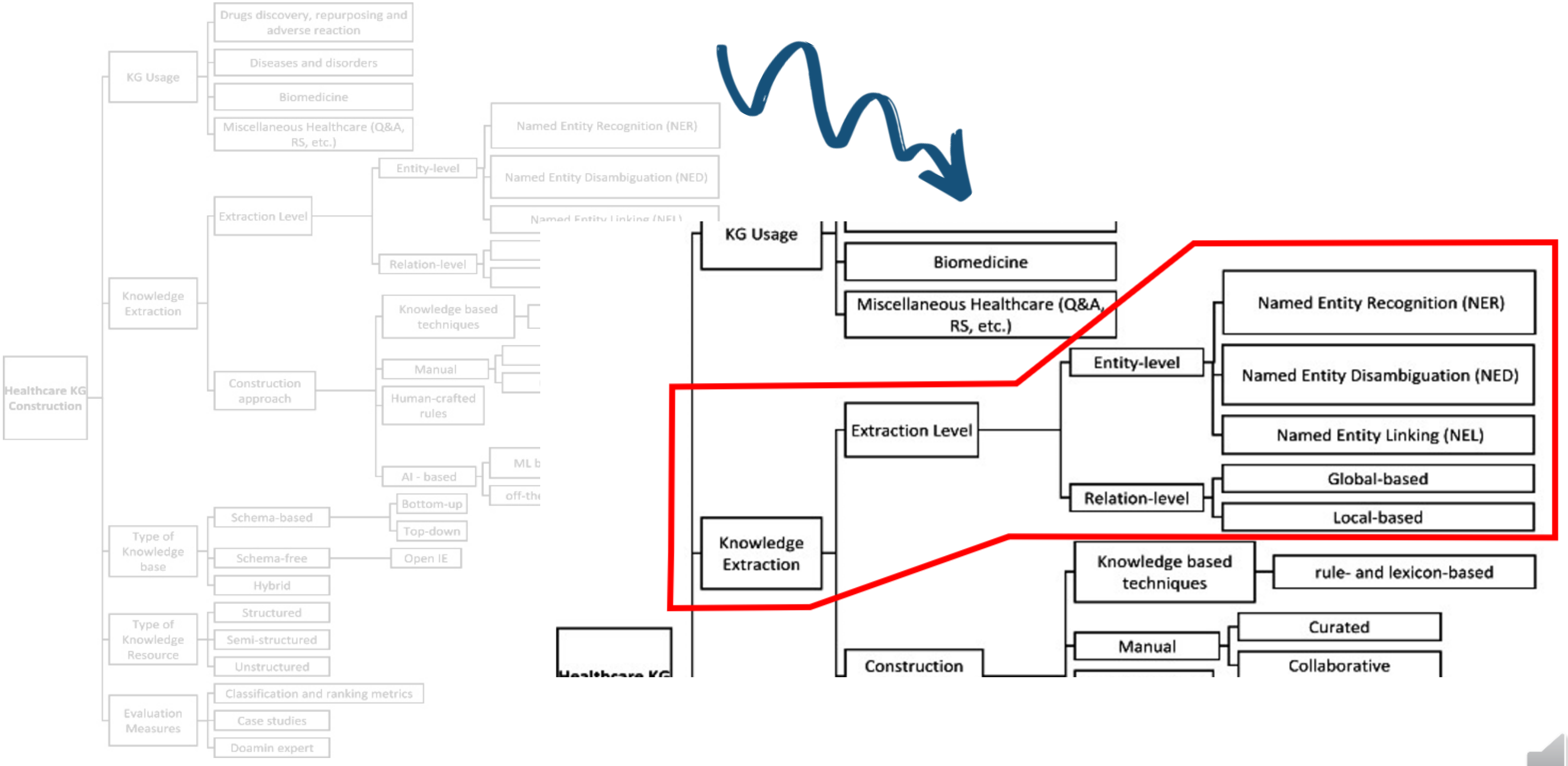
Extensive Use in Healthcare:

- Knowledge graphs have been adopted extensively across various domains in healthcare, from drug discovery to disease management.
- Focus on Knowledge Extraction, which involves the systematic identification and extraction of key information entities (symptoms, diagnoses, treatment responses, ..., along with their interrelationships)

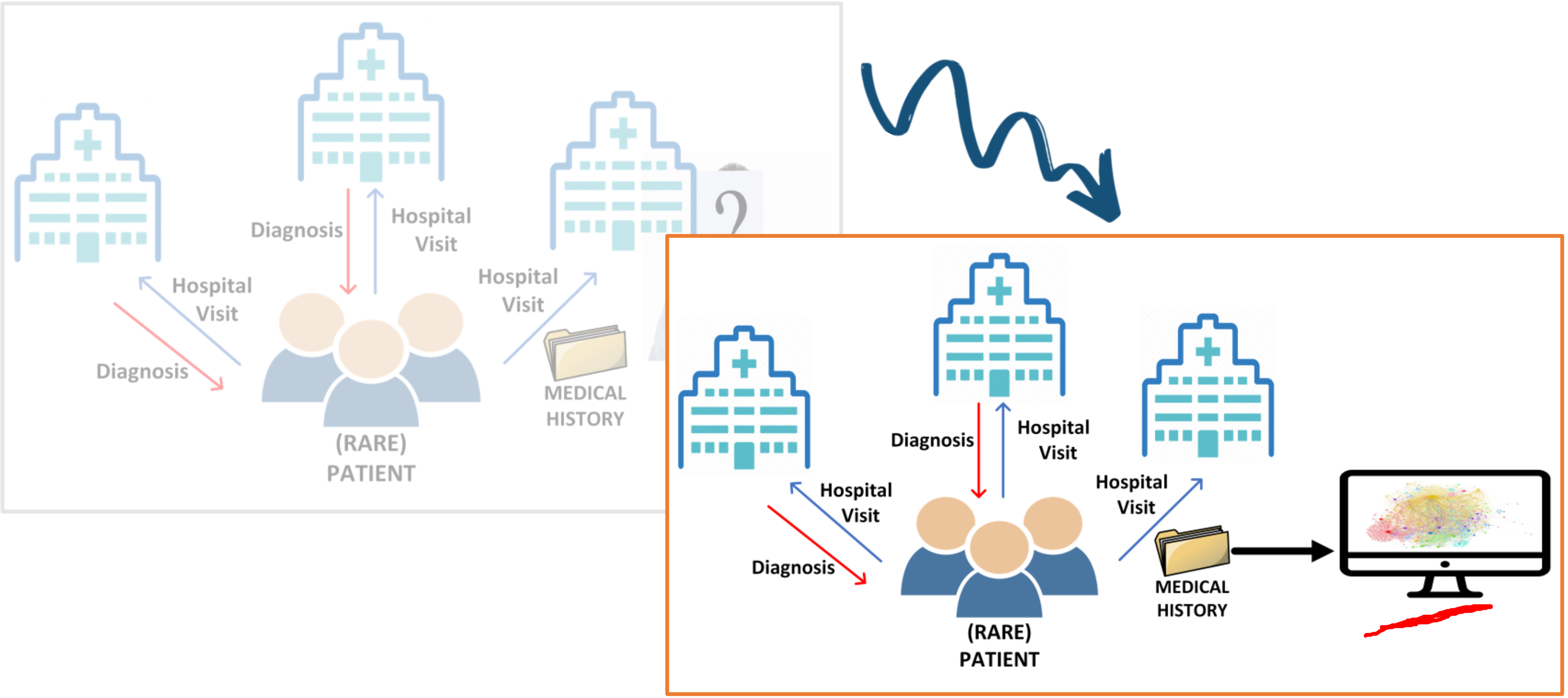


[2] Abu-Salih, Bilal, et al. "Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities." *Journal of Big Data* 10.1 (2023): 81.

Knowledge Graphs for Knowledge Extraction



Transition to a Graph-Based Visualization tool



Methodology

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DISIM
Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

From Reports to Knowledge Graphs

1. **Textual Medical Reports as a Starting Point**:

   Initiating with raw textual medical reports, densely packed with data but intricate for manual doctor analysis.

2. **Processing Data with NLP Techniques**:

   Employing Natural Language Processing (NLP), we performed Named Entity Recognition (NER) and Relation Extraction (RE), to meticulously identify and link crucial biomedical entities within the reports. In particular, we focused on nine specific entities: Gene/Protein, Disease, Drug/Chemical, Mutations, Species, Cell Lines, Cell Type, DNA, and RNA.

3. **Generating Knowledge Graphs from Reports**:

   Post extraction, individual knowledge graphs are generated for each report, visualizing the extracted entities and their interrelations.

4. **Creating a Unified Visualization Tool**:

   These individual knowledge graphs are then amalgamated to form a unified visualization tool. This integration illuminates common entities and their interactions across various reports, crafting a comprehensive dashboard that encapsulates the patient's medical history.

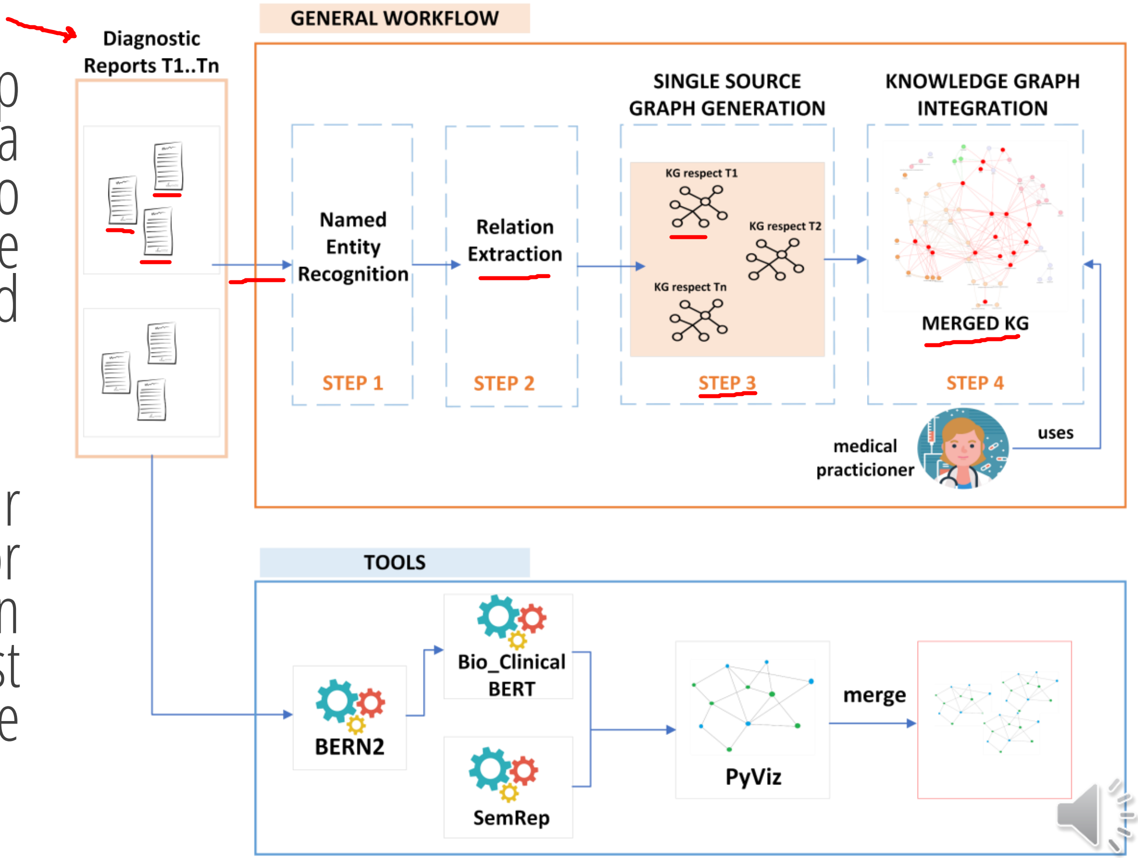
5. **Outcome**:

   The final product is a holistic and interactive knowledge graph that provides healthcare professionals with a dynamic tool for diagnosis and treatment planning.

Two running options:

1. **Manual Entry**: Simply drop medical reports into a designated folder to automatically generate individual and merged knowledge graphs.

2. **Data Ingestion**: Utilize our pre-compiled dataset for structured experimentation and analysis, yielding robust and comprehensive knowledge graphs.



Experimental Setting

UNIVERSITÀ DEGLI STUDI DELL'AQUILA
DISIM
Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

Data Ingestion Mode: Dataset

Database | Credentialed Access

MIMIC-IV-Note: Deidentified free-text clinical notes

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, Roger Mark

Published: Jan. 6, 2023. Version: 2.2

- The MIMIC (Medical Information Mart for Intensive Care) dataset is a large, freely available database comprising de-identified health-related data about patients, sourced from the electronic health records of the Beth Israel Deaconess Medical Center.
- Initial dataset: **145,915 patients** from **MIMIC-IV-Note** discharge summaries. Total of de-identified discharge summaries: **331,794**

[3] Johnson, Alistair, et al. "MIMIC-IV-Note: Deidentified free-text clinical notes." (2023).

Data Pre-Processing

- Preprocessing steps: Selection, extraction, and transformation focused on **History of Present Illness**.
- Refined dataset: Narrowed down to **59,051 unique patients** with multiple hospitalizations.
- Aim: To map out detailed patient health journeys and construct a comprehensive knowledge graph.
- Outcome: A focused dataset enabling an in-depth study of complex medical histories.

Results

UNIVERSITÀ
DEGLI STUDI
DELL'AQUILA

DISIM
Dipartimento di Ingegneria
e Scienze dell'Informazione
e Matematica

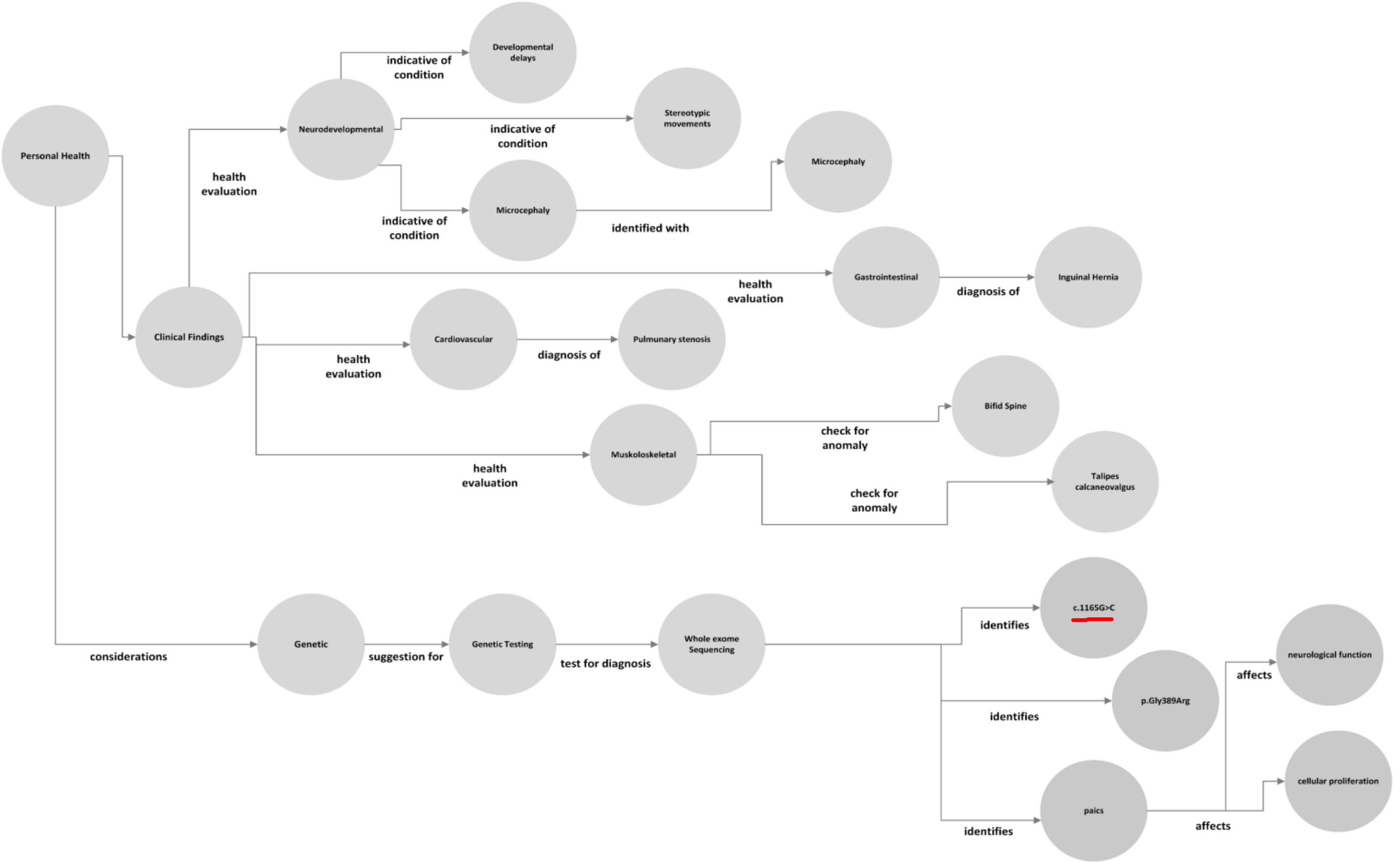
Example Case – Medical Report #1

Age at last evaluation: ___ year and ___ months. **Anthropometric Data**: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. **Clinical Findings**: Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. **Genetic Considerations**: Genetic Testing Results: Gene: PAICS (NM_001079525.1), Mutation Identified: c.1165G>C ; p.Gly389Arg. The identified variant in the PAICS gene may potentially explain some of the clinical findings observed in this patient. The PAICS gene encodes a bifunctional enzyme involved in de novo purine biosynthesis. Mutations in this gene could potentially affect cellular proliferation and neurological function, although the exact clinical significance of the identified variant (c.1165G>C ; p.Gly389Arg) needs further evaluation.

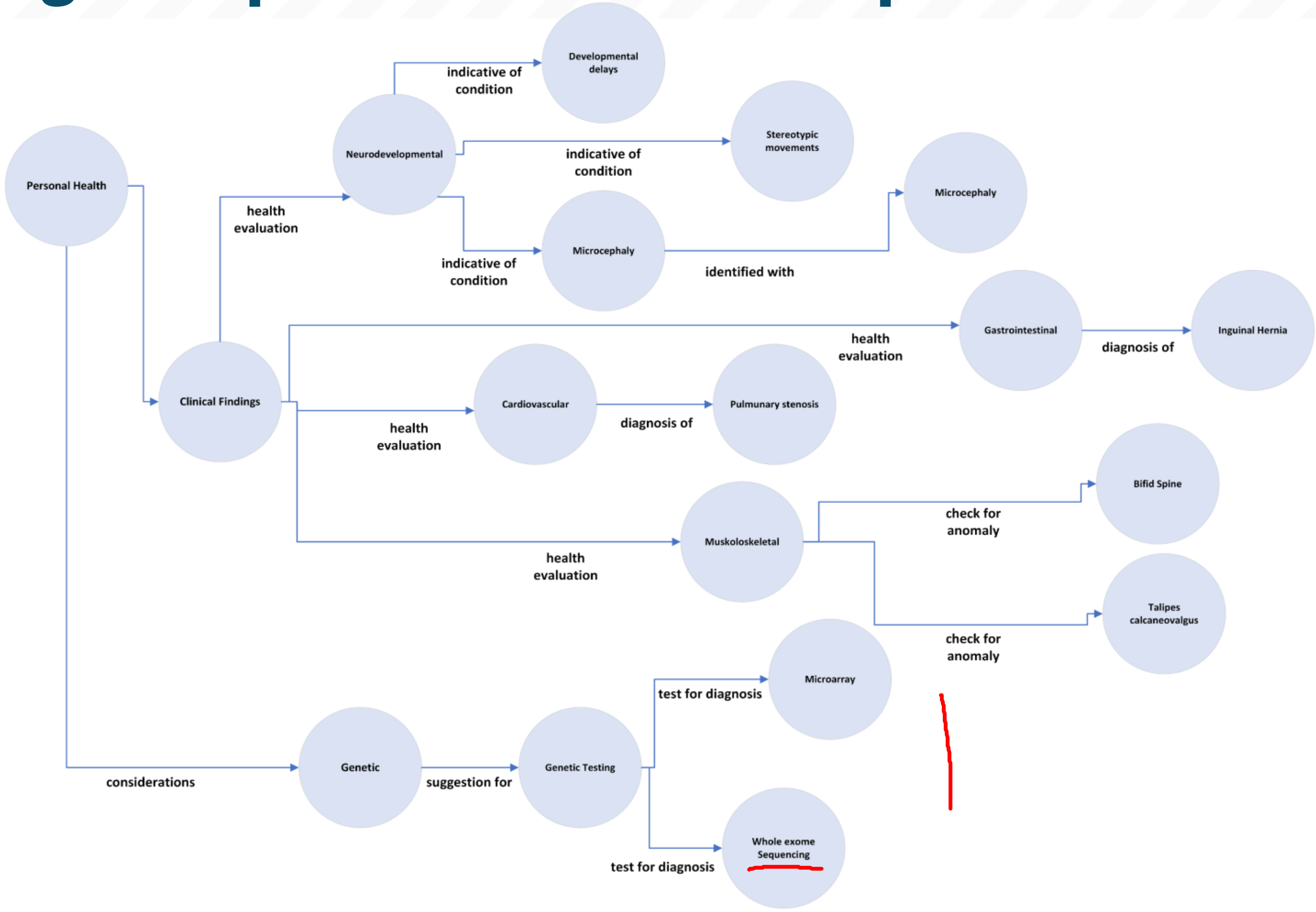
Example Case – Medical Report #2

Age at last evaluation: ___ year and ___ months. **Anthropometric Data**: At Birth: Weight: ___ grams, Length: 46.5 cm, Occipitofrontal circumference (OFC): 32 cm. At ___ Year: OFC: 43 cm (P<3), indicative of Microcephaly. **Clinical Findings**: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus (a deformity involving the ankle and heel) present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. **Genetic Considerations**: The combination of microcephaly, developmental delay, and other physical anomalies may suggest a possible genetic syndrome. Genetic testing, including chromosomal microarray and/or whole exome sequencing, may be indicated to identify any underlying genetic etiologies.

Knowledge Graphs Generation - Report #1

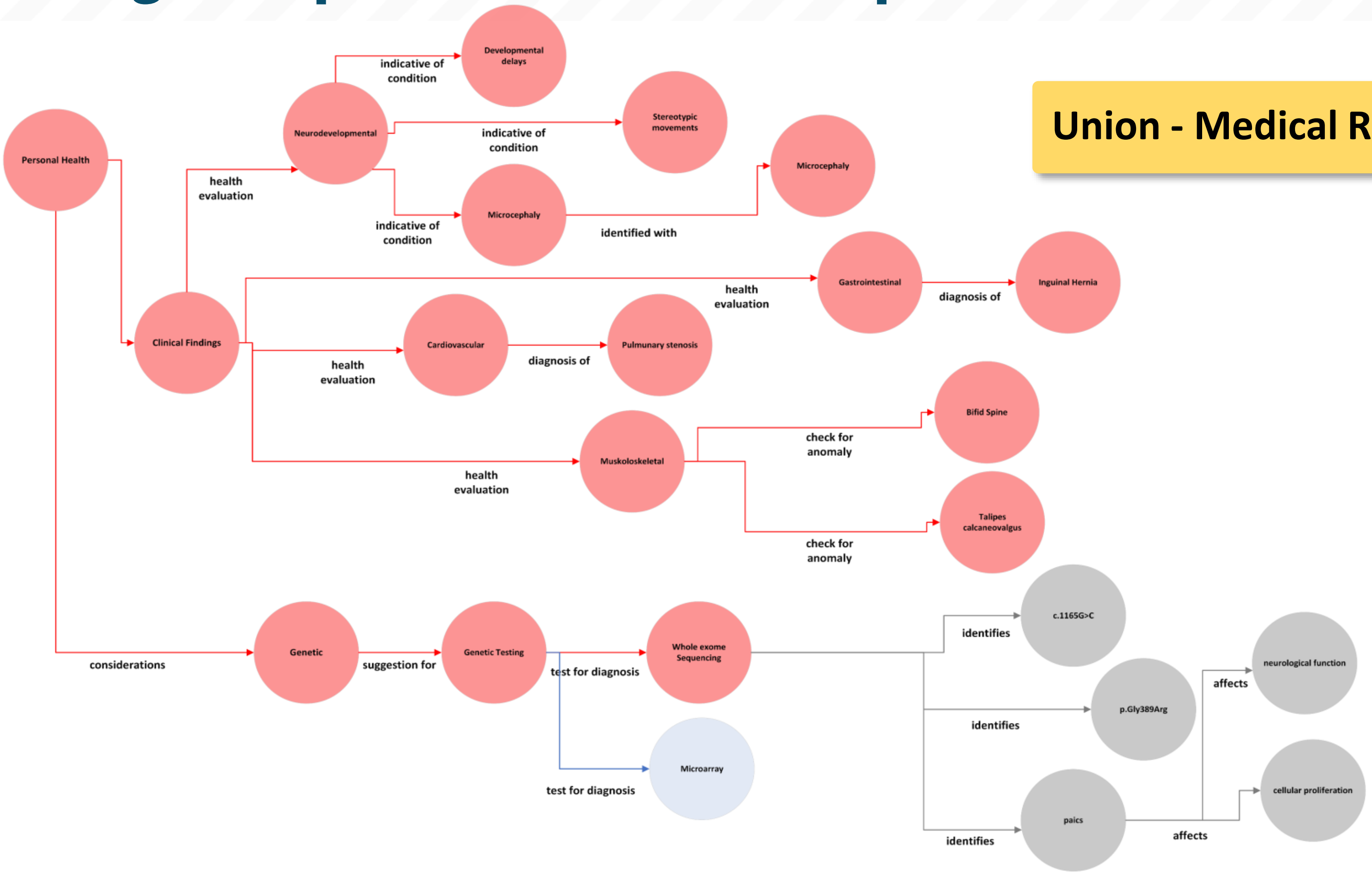


Knowledge Graphs Generation - Report #2



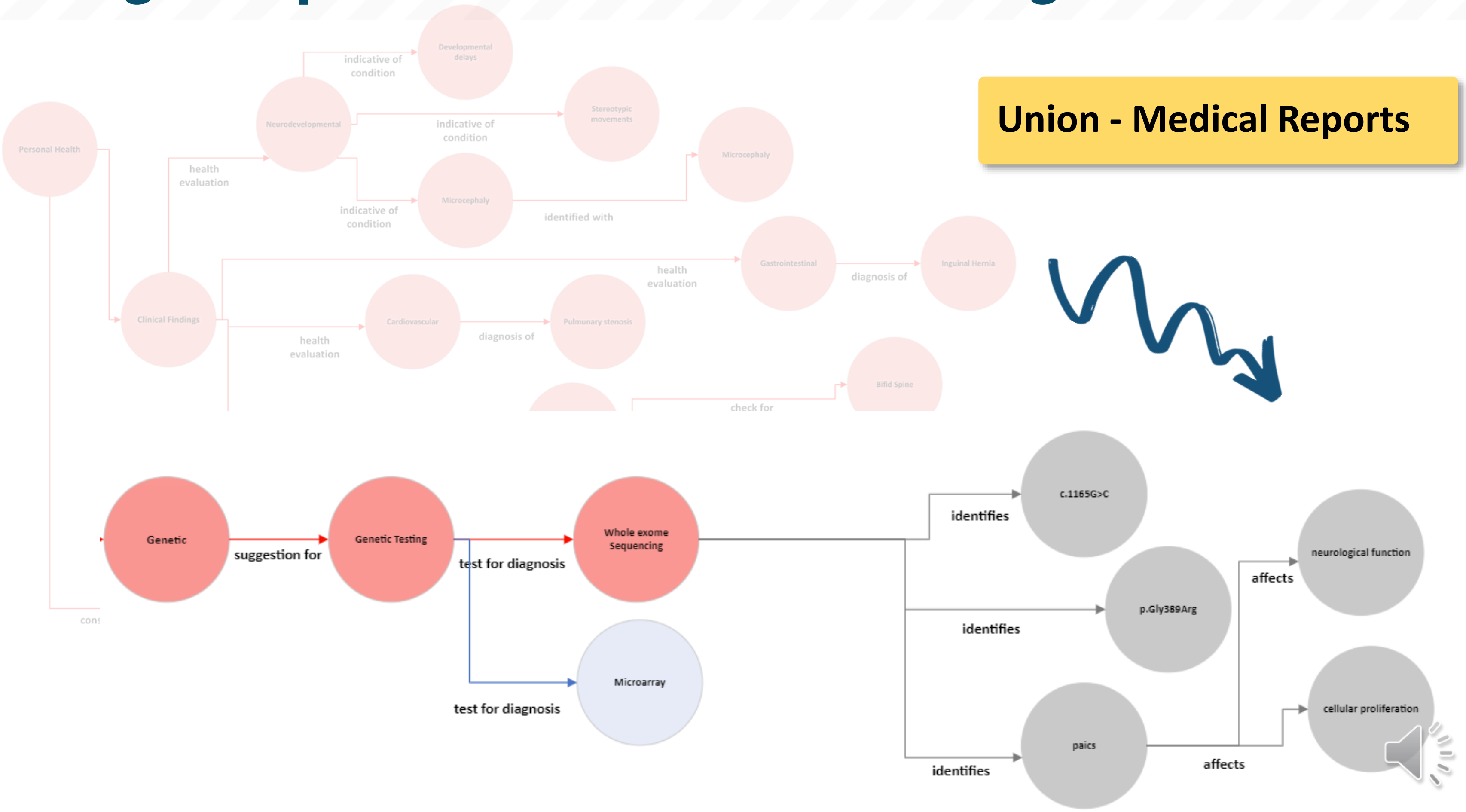
Knowledge Graphs Generation - Report #1 and #2

Union - Medical Reports



Knowledge Graphs Generation – Focus on genetic

Union - Medical Reports



Validation

Data Source Verification:

- Cross-referenced medical data with authoritative sources.
- Validated patient information with (extended) electronic health records.

Conclusions

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DISIM
Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

Discussion

- **Autonomous System Proficiency:** Our study presents a system that autonomously creates and integrates knowledge graphs from vast medical texts, effectively mapping patients' health journeys.
- **Strategic Entity Focus:** By concentrating on select biomedical entities, our system showcased proficiency in isolating and interlinking key data points, yielding comprehensive patient narratives.
- **Healthcare Information Complexity:** Our approach addresses the intricate nature of healthcare information, offering a visual representation that enhances diagnostic processes and therapeutic strategies.
- **Societal Advantages:** Our system facilitates the management of medical data, empowering both patients and practitioners in understanding complex health records, which is especially beneficial for those facing difficulties in recounting detailed medical histories.

Future Directions

- **Multilingual Model Integration:** Recognizing the need for inclusivity, we aim to extend our system's capabilities to encompass diverse languages, thereby catering to a global demographic.
- **Collaborative Validation:** To further validate our system's clinical relevance, we will engage in collaboration with medical experts, aiming to integrate expert feedback and real-world applicability into our model validation processes.
- **Advanced Pattern Recognition:** We plan to implement pattern recognition techniques to scrutinize the knowledge graphs for recurring diagnostics patterns, which could revolutionize early disease detection and patient care strategies.

Thanks for your attention !!

UNIVERSITÀ
DEGLI STUDI
DELL'AQUILA

DISIM
Dipartimento di Ingegneria
e Scienze dell'Informazione
e Matematica