Network Analysis & ML with Graphs
CIS 4930 / CAI 5155

Project 1:
Uncovering Clinical Patterns with Network Analysis

1. Project Overview

In this project, you will use **Social Network Analysis** techniques to analyze patterns within the EHRShot dataset. Your task is to construct networks from patient health events (diagnoses, medications, procedures, labs, demographics, etc) and identify meaningful structures such as communities, progressions, and co-occurrence patterns.

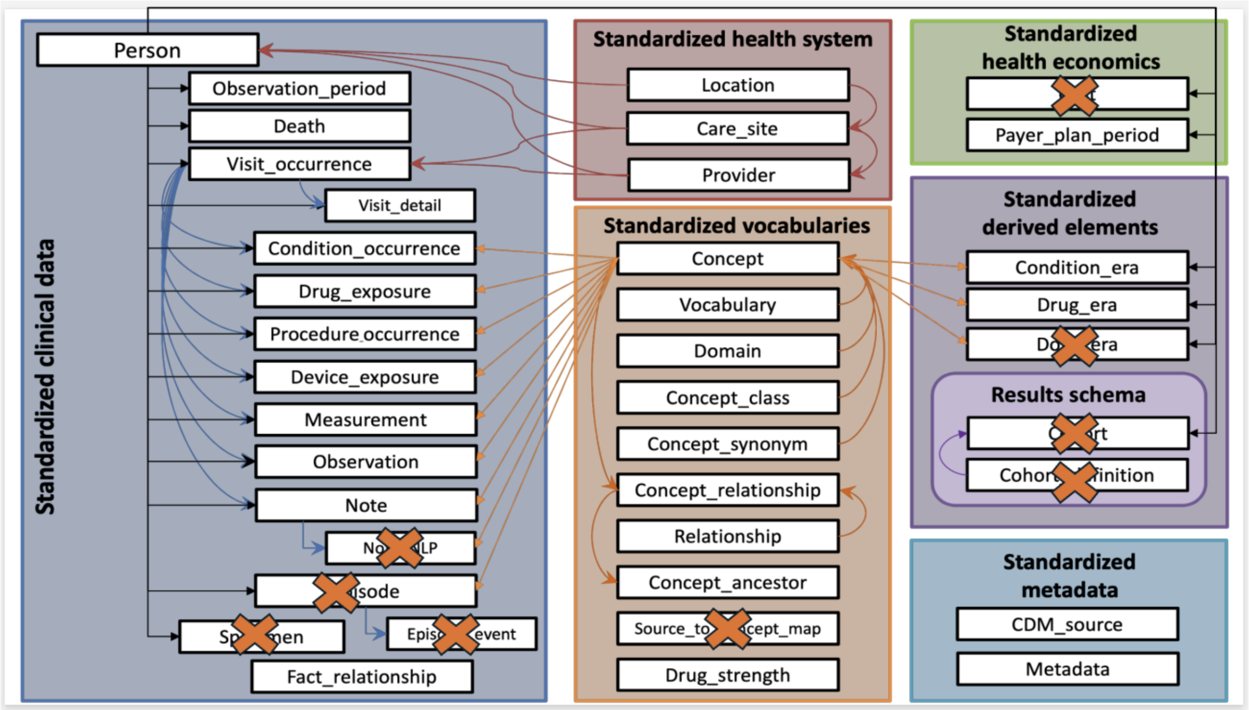
This project is designed to build intuition about the hidden structure of health data before we move into predictive modeling (e.g., Graph Neural Networks) later in the course.

2. Project Goal

- Represent EHR data as a network and understand its structural properties.
- Discover recurring patterns, groups, or sequences in medical data.
- Learn to interpret these structures as potential medical insights.
- Practice applying classical SNA tools (without machine learning).

3. Why Network Analysis on EHR Data?

- Medical records contain **complex interconnections**: diagnoses, treatments, labs, demographics.
- Networks provide a natural way to visualize and quantify these interconnections.
- SNA reveals hidden communities, care pathways, or demographic-specific trends.
- Structural analysis can uncover biases, data quality issues, and recurring clinical patterns.



The diagram above shows the relationship between tables in EHRShot dataset. For brief descriptions of tables, please refer to '**Lecture_03_Project dataset**' slides.
If you want to get names of attributes in each table, Please visit the website below:
https://stanford.redivis.com/datasets/53gc-8rhx41kgt/tables

4. Possible Types of Networks You Can Build

There are thousands of different ways to construct networks using EHR data. Below are some examples for your reference.

- **Co-occurrence networks**: Nodes = diagnoses/medications; Edges = appear together in same visit.
- **Temporal networks**: Directed edges from earlier → later events in patient history.
- **Bipartite networks**: Patients ↔ diagnoses, or diagnoses ↔ drugs.
- **Demographic-specific networks**: Separate graphs for subgroups (e.g., male/female, young/elderly).
- **Multilayer networks**: Combine different event types as multiple interconnected layers.

5. Example Research Questions

You are encouraged to frame your own research question. Be **CREATIVE**!
Examples include:

- What communities of diagnoses or medications frequently co-occur?
- How do disease progressions differ by age or gender?
- Which diagnoses act as bridges between otherwise separate clusters?
- Are there distinct sub-communities for chronic vs. acute conditions?
- What are the most central or influential events in the network?

6. Network Analysis Techniques to Use

- Degree and weighted degree
- Centrality: betweenness, closeness, eigenvector
- Community detection (e.g., Louvain, modularity maximization)
- Density and clustering coefficient
- Motif or triad analysis (optional)
- Subgroup comparison across demographic-specific networks

7. Project Tasks

1. **Explore Dataset**
2. **Brainstorm with your teammates and Define a Research Question**
3. **Select Data Subset**: Choose a manageable portion of EHRShot data (a specific disease, patients older than 90 years old, etc). Document your selection and preprocessing steps.
4. **Define Nodes and Edges**: Decide on a network representation. Examples include:
    a. Co-occurrence networks (diagnoses or medications appearing together).
    b. Temporal networks (directed edges from earlier to later events).
    c. Bipartite networks (patients ↔ diagnoses, or diagnoses ↔ drugs).
    d. Demographic-specific networks (separate graphs by gender, age group, etc.).
5. **Build the Network**: Construct the graph using a library such as PyG, NetworkX, or igraph. Ensure edge weights and node labels are meaningful.
6. **Analyze Structure**: Apply core social network analysis techniques such as degree distribution, centrality, clustering coefficients, and community detection.

7. **Interpret Results**: Identify and describe patterns such as co-occurring clusters, common progressions, or demographic-specific trends. Relate your findings to real-world medical understanding.
8. **Visualize and Summarize**: Provide clear, readable network visualizations and summarize findings in figures or tables.

8. Deliverables

- **Graph files**: Node and edge lists.
- **Code**: A Jupyter notebook or Python script with clear documentation.
- **Project Report**
    - **Section 1. Introduction**
        - Motivate why your research question matters in the context of healthcare and EHR data.
        - Find and describe related works.
        - Provide background on network analysis in healthcare (e.g., co-occurrence of diagnoses, progression of diseases, patient-drug networks).
        - Clearly state your research question(s) or objectives. Examples:
            - "We aim to identify comorbidity clusters among chronic diseases."
            - "We analyze temporal progression of diagnoses in older adults."
        - End with a summary of contributions: what your analysis reveals and why it is valuable.

    - **Section 2. Network Construction Details**
        - Dataset: Describe what subset of EHRShot you used.
        - Preprocessing: Explain how you cleaned, filtered, and standardized data (e.g., merged rare codes, removed missing values).
        - Node definition: What entities are represented (diagnoses, medications, procedures, demographics)?
        - Edge definition: How are relationships formed (co-occurrence, temporal sequence, bipartite links)?
        - Edge weights: Whether you used raw counts, frequencies, or binary connections.

    - Provide enough detail for another student to replicate your network construction.

- **Section 3. Metrics and Analysis Methods**
    - Explain which network metrics you applied and why.
    - Justify why each metric is relevant to your research question.
    - If you compare across subgroups (e.g., male vs female), describe how you built and compared those networks.
    - Reference standard definitions briefly (formulas not required unless you use nonstandard variants).

- **Section 4. Visualizations and Tables**
    - Include network diagrams to illustrate your findings (keep them readable — avoid excessive clutter).
    - Provide tables or plots of key statistics
    - Each figure/table should have a clear caption and be referenced in the text.
    - Focus on clarity and interpretability, not artistic visuals.

- **Section 5. Interpretations and Conclusions**
    - Interpret structural patterns in the context of healthcare.
    - Discuss what your findings reveal about patient care or medical relationships.
    - Acknowledge limitations (e.g., data sparsity, coding bias, lack of temporal granularity).
    - Suggest future directions (e.g., adding more event types, applying GNNs later in the course).

- Writing Guidelines
    - Length: 4–5 pages in double-column (~2,000 words).
    - Style: Clear, concise, scientific writing.
    - Use citations where appropriate (e.g., to network science or healthcare informatics papers).
    - Each section should flow logically and connect back to your research question.

9. Grading Rubric (100 points)

Category	Points	Low Score (0–50%)	High Score (80–100%)
Research Question	10	Research question is vague, trivial, or not connected to healthcare/EHR context. Little to no motivation provided.	Research question is clearly defined, original, and relevant to EHR/healthcare. Strong motivation showing why it matters.
Data Preparation	5	Minimal description of how data was selected or cleaned. Choices not justified or inconsistent.	Clear and concise explanation of subset selection, cleaning, and filtering. Choices are justified and reproducible.
Network Construction	20	Nodes/edges poorly defined or misaligned with research question. Incomplete or confusing explanation of construction steps.	Nodes/edges well-defined and directly tied to research question. Construction is systematic, detailed, and reproducible.
Network Analysis	25	Few or irrelevant metrics applied. No justification for metrics. Results are incomplete or misinterpreted.	Appropriate and diverse set of metrics applied. Strong justification and correct interpretation of each metric.
Interpretation of Findings	20	Results are reported but not connected to medical or healthcare context. Limited or no insights, just descriptive.	Results interpreted in meaningful healthcare context. Provides insightful conclusions and acknowledges limitations.
Report Quality	20	Poor organization, unclear writing, missing figures/tables. Sections incomplete or underdeveloped.	Well-structured, professional scientific writing. Figures/tables support text. All sections complete, concise, and coherent. Proper formatting and citations.