Enhancing Healthcare Intelligence:

Knowledge Graph Construction for advanced Clinical Information Extraction

Table of Contents

Abstract

This poster investigates how unstructured clinical text can be transformed into knowledge graphs to reveal possible hidden medical relationships. By organizing entities within a graphical framework, we overcome traditional barriers to information utilization in healthcare settings. By employing Synthea, an open-source tool for synthetic healthcare data, we generate life-like clinical unstructured notes. We then apply traditional Natural Language Processing (NLP) methodologies that utilize rule-based systems along with Large Language Models (LLM) to better understand the relationships between entities. The resulting knowledge structures enable researchers to visualize patient journeys, identify unexpected treatment correlations, and integrate different data sources with remarkable efficiency, thereby supporting Real-World Evidence (RWE) generation. Such insights are crucial for decision-making, as health-practitioners, pharmaceutical companies and regulators explore how RWE can be leveraged to assess treatment effectiveness, monitor drug safety, and inform policy recommendations.

Introduction

Electronic Health Records (EHRs) are a valuable resource for healthcare analytics and Real-World Evidence research (RWE), with unstructured clinical notes capturing complex but detailed patient information that often escapes codified fields. These notes are written like stories and use different ways to express the same thing. Therefore, it is tough to analyze them with computers on a large scale or to connect them smoothly with other medical data systems. Recent approaches integrating Natural Language Processing (NLP), graph-based data modeling, and established biomedical ontologies have demonstrated potential for addressing this challenge. Synthetic datasets like Synthea provide an open platform for developing and testing extraction pipelines. Knowledge graphs (KGs), in turn, offer a powerful paradigm for representing the relationships and temporal flows assisting in patient journeys and clinical decision-making.[1][2][3] [4]

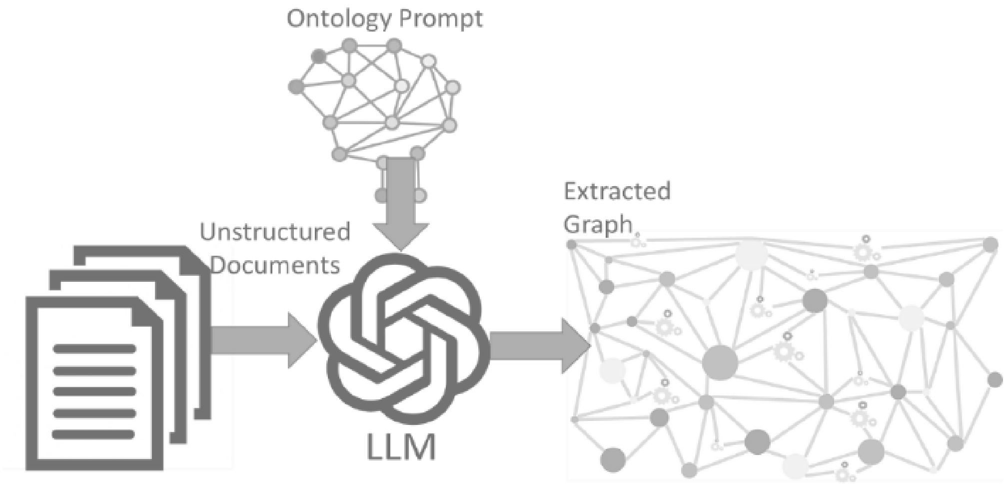

Figure 1. Process construction from raw Synthea clinical to informative Knowledge Graph

Methods and Steps

Synthea Data

Synthea is an open-source synthetic patient generator designed to create realistic medical records, including detailed medication profiles, without exposing any real patient data. Projects have enhanced Synthea by integrating open-source databases, such as the Medical Expenditure Panel Survey (MEPS) and RxNorm, to ensure the synthetic data mirrors authentic prescribing patterns and medication distributions in the United States. This progress supports broader applications of synthetic datasets for research, education, and technology development, offering large volumes of high-quality data with significantly reduced privacy risks. [16]

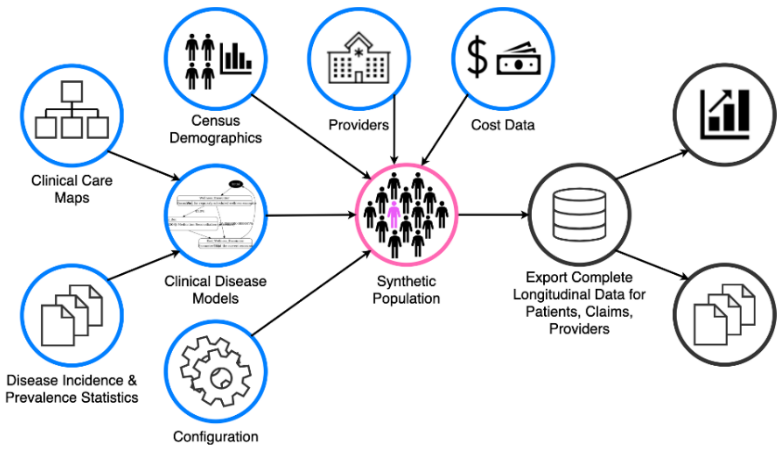

Figure 2. Synthetic Data Ingestions

Data Acquisition

Unstructured clinical notes generated by Synthea were collected, each representing individual patient encounters and containing free-text fields such as “Chief Complaint”, “Assessment,” and “Plan”. To prepare these notes for downstream processing, rigorous and multi-step preprocessing was conducted. This included text normalization, ensuring all characters conform to UTF-8 encoding standards and reducing extraneous whitespace to produce a clean and consistent text format. Next, the notes were segmented into meaningful sections by leveraging regular expressions and domain-specific heuristics aligned with clinical documentation conventions, accurately delineating narrative components and structured fields. Additional preprocessing steps annotated and removed artifacts typical of clinical records, such as date headers, sign-offs, and other non-informative boilerplate text — reducing noise and improving the quality of extracted information. Together, these preprocessing steps created a high-quality, normalized corpus optimized for accurate entity recognition, relationship extraction, and knowledge graph construction.[5] [10]

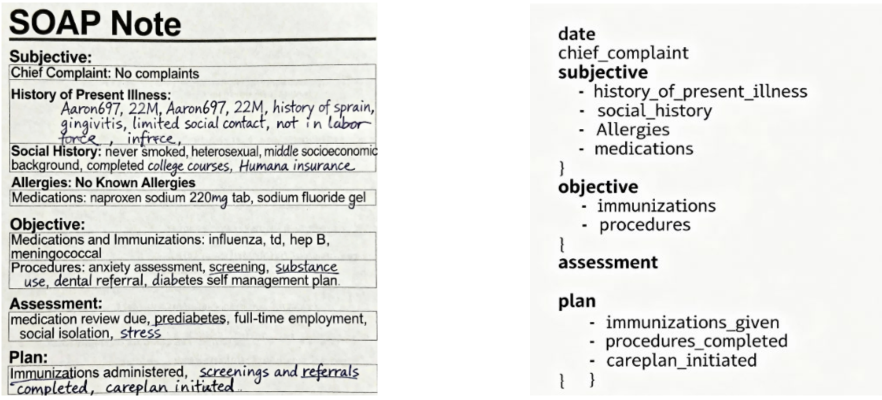

Figure 3. Example of unstructured Synthea clinical note (left side) versus structured JSON format

NLP Techniques for Unstructured Clinical Notes

Raw clinical notes were systematically transformed using a suite of natural language processing (NLP) techniques. These included rule-based segmentation to identify and separate canonical SOAP sections, advanced entity recognition algorithms to extract clinical concepts like diagnoses, procedures, medications, and social attributes, and normalization routines such as text cleaning, case standardization, and ambiguity resolution. Specialized tools matched extracted entities into a hierarchical JSON schema, ensuring consistency and interoperability for downstream modeling. These steps established a robust, machine-readable representation of patient records suited for advanced analytics and semantic retrieval.[17][18][19]

Prompt Engineering and Large Language Model

Once notes were normalized into JavaScript Object Notation (JSON) format, large language models—specifically GPT-4o—were leveraged using prompt engineering to enhance extraction of information and inference. Custom prompt templates were designed to supply the model with contextual instructions and historic patient information, retrieved from previous interactions or longitudinal records. This allowed the LLM to answer context-aware queries, identify relationships between clinical findings, and generate insights by considering both current and past inputs. The use of prompt-based retrieval in LLM workflows enriched, patient-specific outputs and improved the quality of downstream decision support and knowledge graph generation.

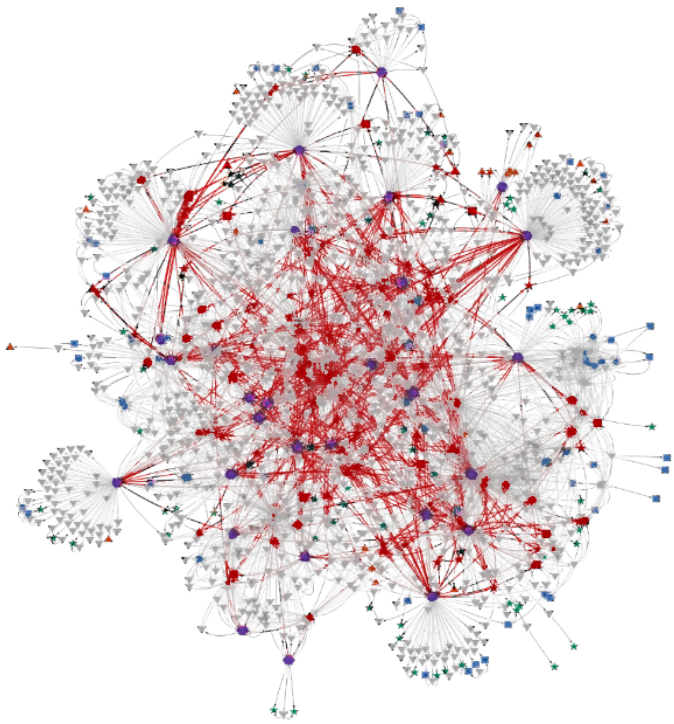

Figure 4. Biomedical Knowledge Graph

Relationship and Temporal Extraction

Explicit and implicit relationships, e.g. patient-diagnosis, diagnosis-medication, and event timelines—were derived using dependency parsing and regular expression-based heuristics. Temporal event extraction included date normalization and sequencing to support longitudinal queries.[7][12]

Entities and relationships extracted from clinical notes were transformed into graph structures, where entities (e.g., patients, diagnoses, and encounters) were represented as nodes, and relationships (e.g., have diagnosis, treated with) are captured as edges between these nodes. This structuring enabled semantic querying, data integration, and advanced analytics within a knowledge graph (KG).[4] [6]

High Level Knowledge Graph Schema

Type	Name	Description
Node	**Patient**	**The subject of receiving healthcare**
Node	**Encounter**	**An interaction/visit between patient and provider**
Node	**Diagnosis**	**Coded/annotated clinical findings or disease**
Edge	**Has diagnosis**	**Links a patient or encounter to a diagnosis**
Edge	**Treated with**	**Connects a diagnosis to prescribed intervention**
Edge	**precedes**	**Captures sequence/order of clinical events**

Table 1: Some node types are core entities representing clinical concepts, while edge types define how these concepts are interlinked within the knowledge graph structure.

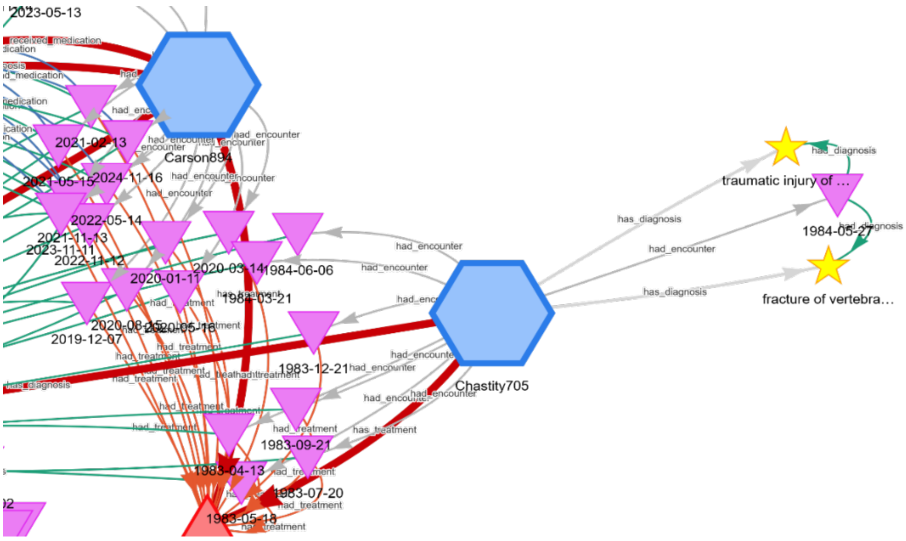

Figure 5. Visualizing Patient Data - Entities and their Relationships

Intuitive mapping for each clinical entity or event in graph Schema

Symbol	KG Entity/Relationship	Meaning in Knowledge Graph
⬢ Hexagon	Patient	The individual receiving healthcare
☆ Star	Diagnosis	Medical condition or annotated clinical finding
▲ Triangle Up	Treatment	Therapeutic intervention or clinical procedure
■ Square	Medication	Drug that is prescribed or administered
◯ Ellipse	Immunization	Event or record indicating vaccination status
▼ Triangle Down	Encounter/Date	Clinical visit or temporal marker
Red highlight	Common event	Event shared by two or more patients
Red arrow	Relationship (edge)	Connection/flow to common event
Pale Color	Single patient event	Event unique to one patient only

Table 2: Shapes and highlights visually differentiate node and edge types within the patient-centric knowledge graph

Discussion

The results support recent findings that high-precision entity-linking and graph extraction are achievable from heterogeneous, synthetic clinical note corpora. The utility of KGs is especially evident in supporting cohort discovery, risk prediction, and hypothesis-driven clinical research at the point of care. Temporal extensions further enrich the analytic possibilities, advancing beyond cross-sectional analyses.[8] [14][5][7][13][2][3]

Key strengths

- End-to-end automation from raw notes to KG which can be queried. [3]
- Use of open, reproducible Synthea (synthetic) data.[16]

Limitations and Future direction

- Synthetic nature of Synthea limits some generalizability to 'real' EHR narrative styles; transfer learning suggested.[4]
- Use of open, reproducible Synthea (synthetic) data.

Conclusion

This work illustrates the feasibility of transforming unstructured Synthea clinical notes into a knowledge graph for the Real-World Evidence field but also opens avenues for real-world applications in clinical decision support, predictive analytics, and interoperability across healthcare systems. The approach leverages advanced NLP and prompt engineering with large language models, robust graph-based data modeling, producing a resource fit for diverse secondary research, clinical informatics, and data-driven healthcare innovation. [1]

The resulting knowledge graph captures a wide array of entities—including patients, diagnoses, symptoms, medications, vaccinations, procedures, and temporal markers—and connects them through meaningful relationships allowing for modelling of individual patient journeys as directed graphs where clinical events are temporally and causally linked, enabling longitudinal analysis and identification of intervention points.

Moreover, the graph supports cross-patient insights by linking individuals with shared diagnoses or symptom clusters, facilitating cohort-level studies and comparative analytics.

Lastly, this methodology contributes to the evolution of intelligent, data-driven healthcare systems that support evidence-based decision-making and improved patient outcomes.

References

1. Wiles, J. "Knowledge Graphs: Principles, Techniques, and Applications." Wiley Interdisciplinary Reviews: Cognitive Science. https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.1549

2. SciELO. "Intelligent Data Management and Analytics." SciELO South Africa. http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2224-78902023000300010

3. Tang, Y. et al. "Efficient Knowledge Graph Construction from Clinical Data." arXiv preprint. https://arxiv.org/html/2306.04802v4

4. Goldstein, M. et al. "Real-World Data on Healthcare Knowledge Graphs." PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC10760044/

5. Subramanian, S. et al. "Mining Knowledge Graphs from Raw Primary Care Physician Notes." PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC10785900/

6. Wang, Q. et al. "PubMed knowledge graph 2.0: Connecting papers, patents, clinical trials, datasets and more." Nature Scientific Data. https://www.nature.com/articles/s41597-025-05343-8

7. Senzing Inc. "Temporal Knowledge Graphs: GPH3 Overview." Senzing Blog. https://senzing.com/gph3-temporal-knowledge-graphs/

8. Han, Y. et al. "A Knowledge Graph-Based Approach for Data Harmonization in Biomedical Research." ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2001037020302804

9. The Synthea Project. "Synthea: An Open-Source Synthetic Patient Generator." https://synthea.mitre.org

10. Sharma, A. et al. "Knowledge Graph-Driven Clinical Data Models." PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC7821471/

11. Zhu, X. et al. "Healthcare knowledge graph construction: A systematic review of the state of the art." PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC10225120/

12. Wang, Q. et al. "A knowledge graph of clinical trials (CTKG)." Nature Scientific Reports. https://www.nature.com/articles/s41598-017-05778-z

13. The Synthea Project. https://synthea.mitre.org

14. Clinical concept normalization with a hybrid natural language processing approach. https://pmc.ncbi.nlm.nih.gov/articles/PMC7647369/

15. Clinical Term Normalization Using Learned Edit Patterns and XML-Based Routing. https://medinform.jmir.org/2021/1/e23104/

16. Text Normalization in NLP. https://www.getgalaxy.io/learn/glossary/text-normalization-for-nlp
17. Harnessing Generative AI for Patient-Centric Clinical Note Generation. https://arxiv.org/pdf/2405.18346.pdf