

ENHANCING HEALTHCARE INTELLIGENCE

ADVANCING CLINICAL INFORMATION EXTRACTION THROUGH KNOWLEDGE GRAPH CONSTRUCTION

AUTHORS

Maria.Puri@pfizer.com - George.Panagiotakis@pfizer.com - Nikolaos.Kountouris@pfizer.com



INTRODUCTION

This work explores how unstructured clinical text can be transformed into knowledge graphs (KG) to uncover hidden medical relationships. Using Synthea synthetic records, we apply Natural Language Processing (NLP) techniques and Large Language Models (LLMs) to extract entities and relationships from clinical notes (SOAP) structured. These KG enable visualization of patient journeys, discover treatment associations ultimately strengthening Real-World Evidence (RWE) generation for healthcare.

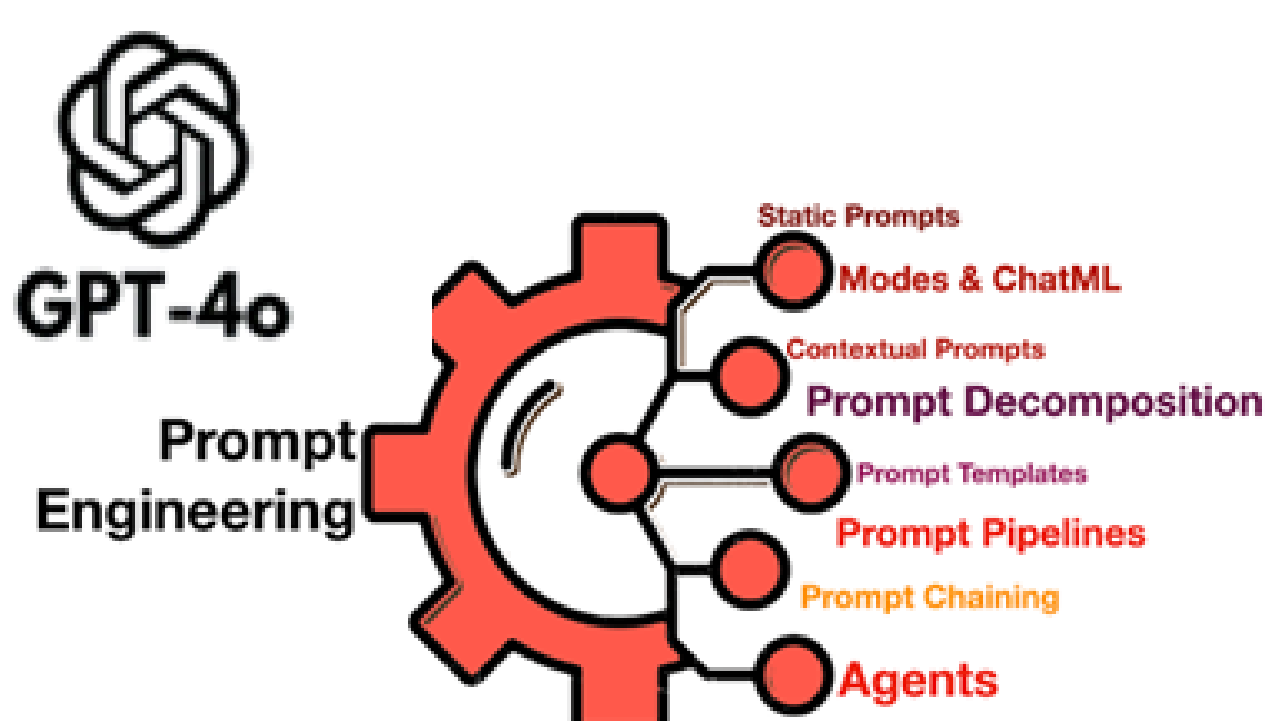
OBJECTIVE

Doctors' notes are long, unstructured stories about patient's health - hard to analyze at scale. The objective of this work is to demonstrate how unstructured clinical narratives, such as those generated by a tool called Synthea, can be systematically transformed into structured knowledge graphs. By converting free-text notes into a format which is easy to analyze, we aim to show how complex patient histories can be traced, patterns of care identified, and RWE more reliably generated.

SYNTHEA UNSTRUCTURED CLINICAL NARRATIVES

2024-04-24
Chief Complaint No complaints.
History of Present Illness Buddy254 is a 40-year-old non-Hispanic white male. History: part-time employment, medication review due, primary dental caries, reports of violence in environment, full-time employment, drug overdose, gingivitis, stress.
Social History Married; prior opioid addiction, quit smoking at age 17. Low socioeconomic background, incomplete high school, Medicaid insured.
Allergies No Known Allergies.
Medications - Vancomycin 150 ml, 5 mg/ml IV - Piperacillin 2000 mg / tazobactam 250 mg IV - Sodium fluoride oral gel
Assessment and Plan Medication review due, part-time employment.
Plan Immunizations: influenza (seasonal), Hep A (adult). Procedures: assessment of health/social care needs, domestic abuse screening, depression screening (twice).

LLM EXTRACTION RULES



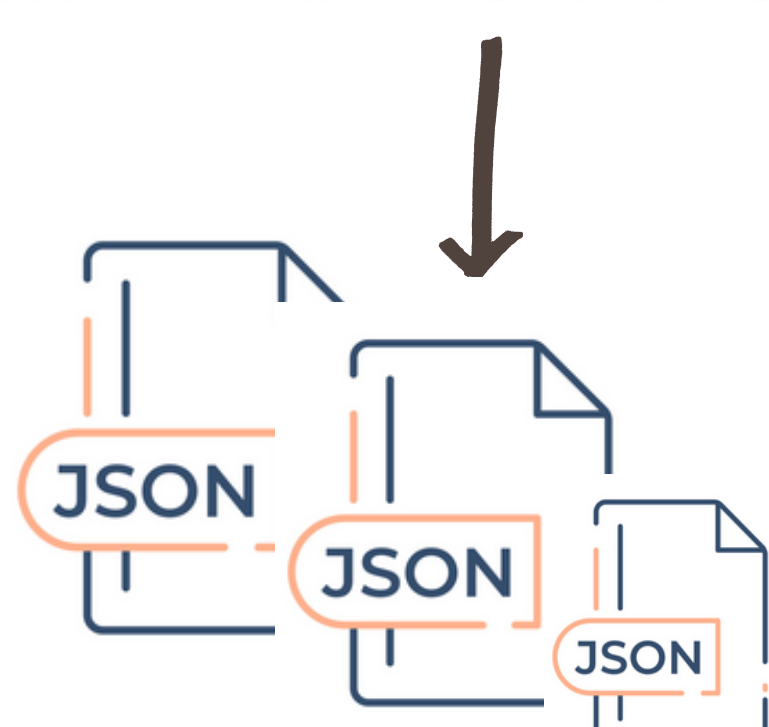
"Extract the following clinical entities from the patient's encounters:

1. Diagnoses
2. Medications
3. Procedures
4. Lab tests
5. Notes (any significant clinical observations)

For each encounter:

- Identify relationships between diagnoses and treatments (e.g., which medication is prescribed for which condition).
- Identify temporal relationships (e.g., which medications were prescribed during specific encounters).
- Extract clinical trends or changes (e.g., HbA1c levels improving over time).

Format the extracted data into a structured JSON that includes these entities and relationships."



We used synthetic clinical notes generated by Synthea as our data source. We used synthetic clinical notes generated by Synthea as our data source (example on the left). We first converted all notes into a standardized format, which organized each patient and their encounters in a clear, chronological sequence.

For clinical information extraction, we employed OpenAI's GPT-4o large language model. We designed specific prompts to guide the model to identify relevant diagnoses, treatments, medications, procedures and immunizations for each patient. The model also recognized how these entities related to each other within the patient's journey. All extracted data were stored as structured JSON files.

To support analysis and interpretation, we created interactive visualization tools for the knowledge graphs. These tools allowed us to view shared and unique events among patients. We also implemented filters for gender, ethnicity, and age to make subgroup exploration and real-world evidence generation easier.

Legend: Knowledge Graph Shapes

- Hexagon: Patient
- Star: Diagnosis
- Triangle: Treatment
- Square: Medication
- Ellipse: Immunization
- Triangle Down: Encounter/Date
- Leave (any shape): Common Event (shared among patients): usual color/shape

Thick Red Arrows: Patient connection to any common event (shared among patients)

Basic Filters (at least one field must match):

Gender: Age: Ethnicity:

17 of 29 patients shown

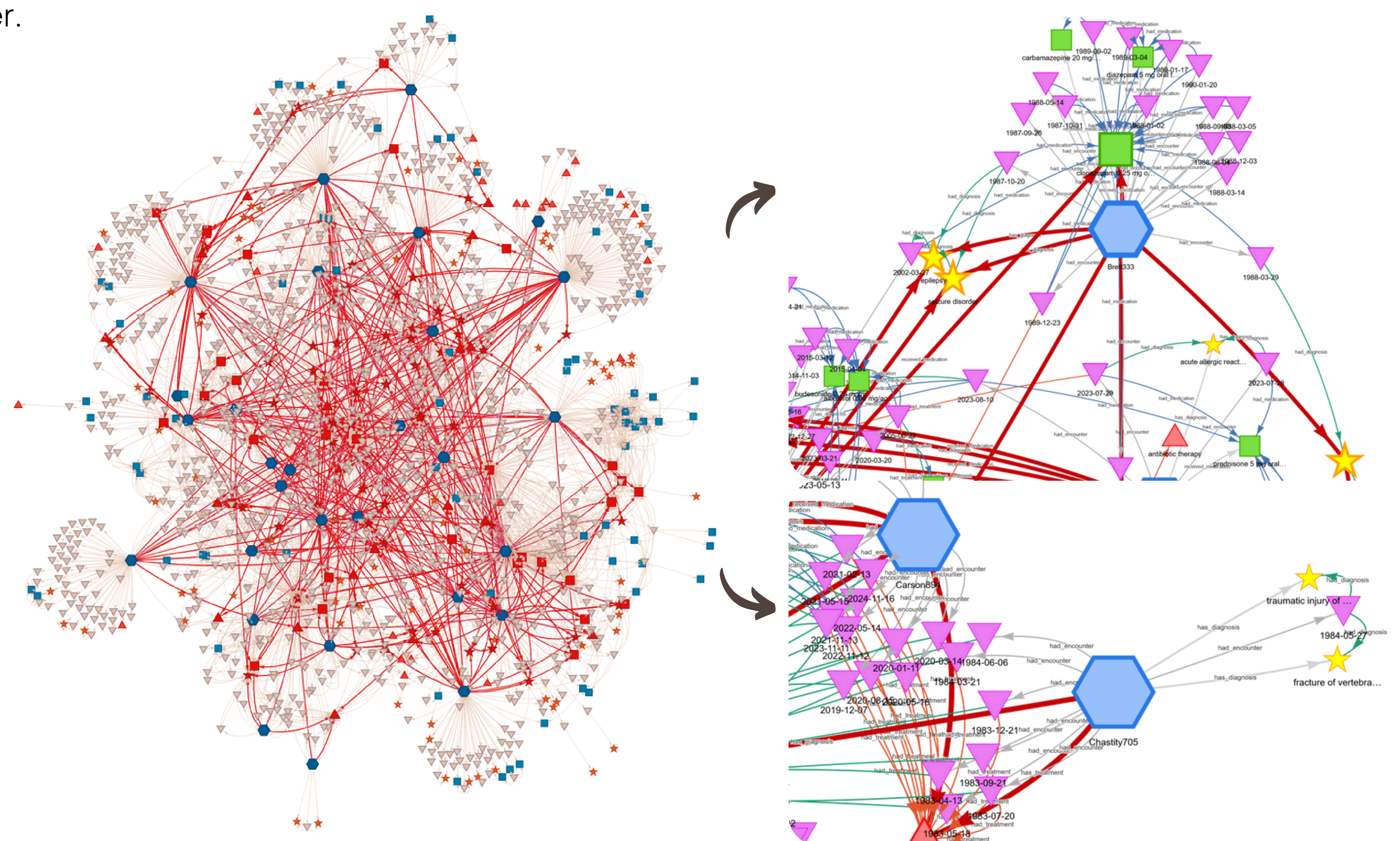


Figure 1: Nodes represent patients, clinical events (diagnoses, treatments, medications, procedures), or encounter timepoints. Edges illustrate the chronological and clinical relationships along each patient's journey.

CONCLUSION

Knowledge graphs constructed from synthetic clinical notes are feasible and informative. Natural language processing and large language models enable structured extraction of patient journeys. These graphs reveal care pathways, risks, and subgroup patterns for analytics. Scaling and real-world application can further advance healthcare insights.