Transforming Plain Text Radiology Reports into Structured Data Using Common Data Elements and FHIR Standards

Michael Hood, MD, Massachusetts General Hospital; Roshan Fahimi, MD; Heather Chase; Tarik Alkasab, MD, PhD

Introduction/Background

This project aims to enable transforming plain text radiology reports into structured data using Common Data Elements (CDEs) and Fast Healthcare Interoperability Resources (FHIR) standards. This effort enhances downstream clinical applications and improves the compatibility and exchange of radiology data across healthcare systems.

Methods/Intervention

A large language model (LLM) was employed to generate preliminary CDE definitions from anonymized chest CT reports. These reports were segmented into overlapping chunks, with semantic vectors generated. For given chest CT findings, relevant chunks were retrieved using cosine-similarity search and reranked. A GPT model was then prompted to create structured data models using the report chunks as context. The model was prompted to generate models with attributes such as identification, characteristics, and associated findings. Models then underwent iterative refinement and expert radiologist review.
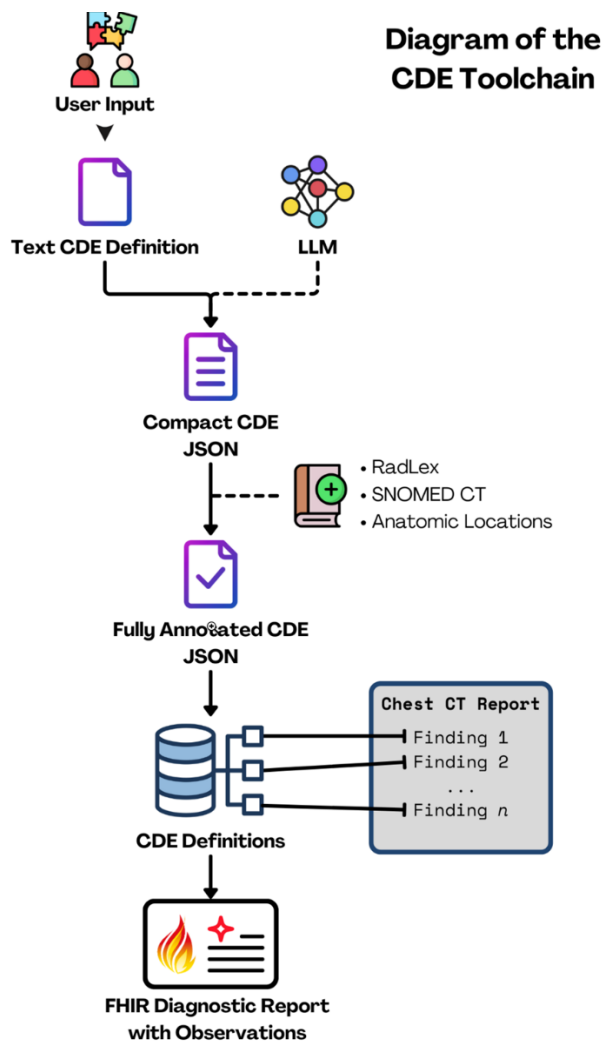
Results/Outcome

This initial pilot generated refined CDE definitions for over 200 chest CT findings, demonstrating LLM capability in rapidly producing preliminary CDEs. Our toolkit successfully transformed these into fully annotated JSON files for generating CDE-labeled FHIR Observation objects. This new process for appending ontological tags is robust — analysis of 82 chest CT reports revealed a total of 1190 findings, 83.2% of which could be successfully encoded as CDE-labeled FHIR Observations.

Conclusion

The developed methodologies and tools significantly expedite the generation and application of CDE definitions, which will enable structured and standardized representation of radiology findings using a standard FHIR structure.

Statement of Impact

This project demonstrates the feasibility of transforming unstructured radiology reports into structured data by integrating tailored LLMs, a streamlined toolchain, and standardized semantics. This improves compatibility and data exchange across healthcare systems, empowering downstream clinical workflows and ultimately improving patient care.

Diagram of the CDE Toolchain

- User Input
- Text CDE Definition
- LLM
- Compact CDE JSON
  - RadLex
  - SNOMED CT
  - Anatomic Locations
- Fully Annotated CDE JSON
- CDE Definitions

Chest CT Report
- Finding 1
- Finding 2
- ...
- Finding n

- FHIR Diagnostic Report with Observations

Summary of LLM-aided generation of Common Data Element definitions for radiology findings based on radiology reports to enable extraction of structured data from radiology reports.

Keywords