Converting Radiology Free-Text to Structured OMOP Data

Zardar Khan 28 JAN 25

Problem Statement

- Radiology reports as untapped research potential
  - Rich clinical information in unstructured format (free-text)
  - Difficult to query or analyze at scale
- OMOP advantages:
  - Source system independence
  - PHI removal / data privacy
  - Standardized vocabulary
  - Research community access
  - Common data model for multi-center studies

Technical Overview

Technical Overview

Implementation Scope

- Full end-to-end pipeline using microcalcification as a case study

Technical Overview

Pipeline Components: Source System

- FHIR DiagnosticReport
  - Report details (ID, text, radiology codes i.e. MAMMLB, temporal information)
- FHIR Patient (ID, demographics)

Technical Overview

Pipeline Components: NLP Model Architecture

- RadBERT - a domain-adapted BERT model
  - Base model trained on 4M radiology reports from Veterans Affairs hospitals
  - Captures unique linguistic patterns and medical context specific to radiology reports
- Fine-tuned for microcalcification classification
  - Three-class classification task: positive / negative / not stated
  - 94% weighted accuracy on a training set of 503 labelled reports

Technical Overview

Pipeline Components: OMOP CDM Integration

- Sequential table population: `Person -> Note -> Note_NLP -> Observation`
- Transaction-safe operations
- Maintains data integrity and relationships

OMOP Implementation Deep Dive

OMOP Implementation Deep Dive

Table Flow and Relationships

- Person creation
  - Uniqueness managed via person_source_value (stores FHIR ID)
  - Returns existing person if FHIR ID found
- Note creation
  - Links to Person via person_id
  - note_text stores FHIR DiagnosticReport ID only (enables source system backtracking)
  - note_type _concept_id = EHR
  - note_class_concept_id = DIAGNOSTIC_STUDY

OMOP Implementation Deep Dive

Table Flow and Relationships

- Note_NLP entries:
  - lexical_variant: JSON structure with question/answer format

```
{
  "question": {"concept_id": 4132707, "concept_name": "Microcalcification"},
  "answer": {"concept_id": 9189, "concept_name": "Negative"}
}
```

  - nlp_system: Stores model metadata (lab name, model name, version)

OMOP Implementation Deep Dive

Table Flow and Relationships

- Model Metadata

```
nlp_system = ModelMetadata.RADBERT_MICROCALCS_V1
# {"manufacturer": "AMartel Lab",
#  "deviceName": "RadBERT-microcalcs",
#  "version": "1.0.0"}
```

OMOP Implementation Deep Dive

Design Decisions

- If we stop at this step, entries invisible to OMOP tools i.e. ATLAS (cohort discovery)
  - Solutions explored: Extend Note_NLP (add person_id, event fields), Custom linkage table
- Note_NLP to Clinical Tables:
  - Domain Resolution via Concept
    - Question Concept determines target domain (Observation, Measurement, etc.)
    - Automated routing based on concept's domain_id
- Observation value represented as a Concept
- observation_source_value

```
table=note_nlp;
id={note_nlp_id};
str_val={finding['answer']['concept_name]}
```

Implementation Challenges

Design Decisions

- FHIR identifier preservation strategy:
  - Person: person_source_value = fhir_id
  - Note: note_source_value = diagnostic_report_id

Implementation Challenges

Implementation Challenges

Observation Period Management

- A Person's Observation Period is akin to their timeline
  - If any Observations fall outside then ATLAS can't find it during cohort discovery.

Implementation Challenges

Observation Period Management: Implementation Strategy

- Investigation: find_notes_outside_periods()
  - Scans notes against existing periods
  - Identifies temporal mismatches
  - Reports position (before/after)
- Update: update_observation_period()
  - Transaction-safe period extensions
  - Handles both start / end adjustments

Implementation Challenges

Procedure-Note Relationships

- Establishing 1:1 relationship between notes and procedure_occurrence
  - Date matching insufficient (multiple same-day procedures)
  - Use of visit_occurrence too broad (rejected)
  - Must maintain OMOP CDM integrity

Implementation Challenges

Procedure-Note Relationships: Solution Analysis

- Procedure_source_value approach (considered)
  - Simpler but non-standard
  - Limited extensibility
  - Data integrity concerns

Implementation Challenges

Procedure-Note Relationships: Solution Analysis

- FactRelationship Table contains records about the relationships between facts stored as records in any table of the CDM.
  - Example: facts derived from one another (measurements derived from an associated specimen)
- FactRelationship approach (selected)
  - Advantages:
    - OMOP-compliant solution
    - Explicit 1:1 relationships
    - Clear data lineage

Summary and Next Steps

Summary and Next Steps

Current Status

- Pipeline Implementation:
  - FHIR -> OMOP ingestion
  - NLP integration (RadBERT microcalcification classifier)
- Technical Focus:
  - Testing FactRelationship for procedure-note linkage, possible usage with Note_NLP to Clinical Table

Summary and Next Steps

Next Steps

- Scale to Full Dataset
  - Process 2M radiology reports
- Expand NLP Coverage
  - Previous history of cancer classifier
  - Menopausal status model
  - BI-RADS score extraction
  - Background parenchymal enhancement (BPE)

A special acknowledgment to the Biomatrix Team and Richard Mraz’s Information Services Team