OHDSI NLP schema proposal
April 13, 2016

Noémie Elhadad, Karthik Natarajan
Columbia University

noemie@gmail.com
Outline

• Proposed schema for storing output of NLP pipeline into the OHDSI CDM

• Edits to Note table
• New table: Note_NLP
# Note table – CDM v5.0

<table>
<thead>
<tr>
<th>Field</th>
<th>Required</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>note_id</td>
<td>Yes</td>
<td>integer</td>
<td>A unique identifier for each note.</td>
</tr>
<tr>
<td>person_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key identifier to the person about whom the note was recorded. The demographic details of that person are stored in the person table.</td>
</tr>
<tr>
<td>note_date</td>
<td>Yes</td>
<td>date</td>
<td>The date the note was recorded.</td>
</tr>
<tr>
<td>note_time</td>
<td>No</td>
<td>time</td>
<td>The time the note was recorded.</td>
</tr>
<tr>
<td>note_type_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.</td>
</tr>
<tr>
<td>note_text</td>
<td>Yes</td>
<td>CLOB</td>
<td>The content of the note.</td>
</tr>
<tr>
<td>provider_id</td>
<td>No</td>
<td>integer</td>
<td>A foreign key to the provider in the provider table who was responsible for taking the note.</td>
</tr>
<tr>
<td>note_source_value</td>
<td>No</td>
<td>varchar(50)</td>
<td>The source value associated with the origin of the note, as standardized using the note_concept_id</td>
</tr>
<tr>
<td>visit_occurrence_id</td>
<td>No</td>
<td>integer</td>
<td>Foreign key to visit</td>
</tr>
</tbody>
</table>
# Note table – CDM v5.0

<table>
<thead>
<tr>
<th>note_type_concept_id</th>
<th>Yes</th>
<th>integer</th>
<th>The time the note was recorded. A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.</th>
</tr>
</thead>
</table>

- Pathology Report
- Discharge Summary
- Nursing Report
- Outpatient Note
- ED Note
- Inpatient Note
- Radiology
- Ancillary Report
- Note
- Admission Note
Note Table proposed edits

• Replace Note_type_concept_id with 5 elements
  – Note_role_concept_id (Role)
  – Note_domain_concept_id (Subject Matter Domain)
  – Note_setting_concept_id (Setting)
  – Note_service_concept_id (Type of Service)
  – Note_kind_concept_id (Document Kind)
Note – Role proposed

• High-level LOINC taxonomy of roles
• Filtered based on note type frequency at CUMC

Physician
Nurse
Assistant
Student
Therapist_Technician
Case Manager
Patient
Note – Domain proposed

• High-level LOINC taxonomy of subject matter domains
• Filtered based on note type frequency at CUMC

• 53 original domains or slightly filtered out?
  – Filter out Ethics, Forensic, Pastoral Care, Pharmacy?
Note – Setting proposed

• High-level LOINC taxonomy of **settings**

• At CUMC
  – Home
  – Inpatient
  – Outpatient
    • Rehab, ICU, ED
  – Telephone

• Propose to stick to original LOINC codes
Note – Type of Service proposed

• High-level LOINC taxonomy of type of service

• At CUMC, modified mapping from LOIN

• Proposed: compare to at least one more institution
Note – Document Kind proposed

- High-level LOINC taxonomy of kind of document
- Filtered based on CUMC note types

- Note
- Report
- Letter
- Instruction
- Advanced Directive
- Administrative Note
Note_NLP Table

• New proposed table that stores output of NLP pipeline
• Restrictions for now
  – Only store disease/disorders semantic group
  – Only store positive mentions of a disease/disorders term
• Keep data provenance at the concept level
• Similar to Condition_occurrence table in CDM
  – E.g. Condition_era contains more inferred information
  – Inferences about NLP outputs belong to a different table
# Note_NLP Table

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>note_NLP_id</td>
<td>Unique identifier for each concept extracted from NLP</td>
</tr>
<tr>
<td>note_id</td>
<td>Foreign key identifier to the note the concept was extracted from (Note table).</td>
</tr>
<tr>
<td>section_concept_id</td>
<td>Foreign key to predefined concept identifier in the Standardized Vocabularies (LOINC) reflecting the section the extracted concept belongs to.</td>
</tr>
<tr>
<td>lexical_variant</td>
<td>Raw text extracted from NLP</td>
</tr>
<tr>
<td>note_NLP_concept_id</td>
<td>Foreign key to concept id (Concept Table). Restriction to SNOMED-CT as standard vocabulary</td>
</tr>
<tr>
<td>semantic_type</td>
<td>Associated semantic type – UMLS semtypes—need to add as a standard terminology in OMOP?</td>
</tr>
<tr>
<td>NLP_system</td>
<td>String describing system and version used for NLP</td>
</tr>
<tr>
<td>NLP_modifiers</td>
<td>String describing modifiers “uncert</td>
</tr>
<tr>
<td>Temporal_expression</td>
<td>Date if NLP identified an associated temporal expression with disorder</td>
</tr>
</tbody>
</table>