OHDSI NLP schema proposal
April 20, 2016

Noémie Elhadad (noemie@gmail.com)
Columbia University

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

Field	Required	Type	Description
note_id	Yes	integer	A unique identifier for each note.
person_id	Yes	integer	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.
note_date	Yes	date	The date the note was recorded.
note_time	No	time	The time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.
note_text	Yes	CLOB	The content of the note.
provider_id	No	integer	A foreign key to the provider in the provider table who was responsible for taking the note.
note_source_value	No	varchar(50)	The source value associated with the origin of the note, as standardized using the note_concept_id
visit_occurrence_id	No	integer	Foreign key to visit

Note table – CDM v5.0

note_time	No	time	The time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.

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 propos

- High-level LOINC taxonomy of [type of service]
- At CUMC, modified mapping from LOINC
- Proposed: compare to at least one more institution

Addendum
Communication
. Consult_Referral
Consult
. Counseling
. . Individual_Counseling
Daily_or_End_of_Shift_Signout
Diagnostic_Study
Education
. Discharge_Instructions
Evaluation_and_Management
. Annual_Evaluation
. Conference
. . Case_Conference
. Crisis_Intervention_(Pyschosocial_Crisis_Intervention)
. Disease_Staging
. Event
. History_and_Physical
. . Admission
. . Comprehensive_History_and_Physical
. . Targeted_History_and_Physical
. Initial_Evaluation
. . Admission
. . Admission_History_and_Physical
. Managment_of_a_Specific_Problem
. . Evaluation_and_Management_of_Anticoagulation
. Medication_Management
. . Medication_List
. Pastoral_Care
. Plan
. . Treatment_Plan
. Progress
. Risk_Assessment_&_Screening
. . Fall_Risk_Assessment
. Subsequent_evaluation
. Summary
. . Discharge_Note
. . Discharge_Plan
. . Discharge_Summary
. . Transfer
. Surgical_Operation
. . Post-Operative
. . Pre-Operative
. Telephone_Encounter
. Tie-in
. Transplant_Donor_Evaluation
. Well_Child_Visit
Procedure
. Diagnostic_Procedure
. Interventional_Procedure
. Operative Procedure
Referral
. Consult_Referral
Triage

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 table –summary of proposed edits

- Note Table proposed edits
  - Note_source_value:
    - extend the string to 250 chars
    - remove reference to standardized terminology
    - maybe change name to note_title_source_value or title_source_value, so that it is clear that it should be the title of the note
  - Proposed 5 elements instead of note_type_concept_id and their potential values/LOINC codes

Storing NLP extracted terms

- Each NLP-extracted observation should be stored in their respective tables.
  - How to handle negation / uncertainty?
  - How to handle NLP system that doesn't do semantic mapping or BOW?
- Hybrid solution:
  - Single Note_NLP table that contains all the NLP extracted concepts, with a flexible structure wrt modifiers that can work for all types of concepts
  - Several NLP_<xxx> tables that provide explicit structure of modifiers for each concept type (e.g., measurement vs condition vs medication)
  - Still possible in ETL to include NLP-extracted information into other tables, but left up to each institution to make sure for all queries to be cognizant of the fact that they could contain NLP outputs

WG Proposed: Note_NLP Table

- New proposed table that stores output of NLP pipeline
- 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
    - Eg. ”low sodium” → “hyponatremia”

Note_NLP Table

Note_NLP_id	Unique identifier for each concept extracted from NLP
note_id	Foreign key identifier to the note the concept was extracted from (Note table).
section_concept_id	Foreign key to predefined concept identifier in the Standardized Vocabularies (LOINC) reflecting the section the extracted concept belongs to.
snippet	Small window of text surrounding term mention
lexical_variant	Raw text extracted from NLP
Note_NLP_concept_id	Foreign key to concept id (Concept Table). Domain concept is provided as part of the Concept table.
NLP_system	String describing system and version used for NLP (data provenance)
NLP_date	Date describing date at which note was processed

- (see later for modifiers)

How to handle modifiers?

- Are there common modifiers to all types of concepts that should be explicit columns?
  - Negated, Temporal expression, Experiencer
  - Or, a logical combination of them?
- How are type-specific modifiers/values stored?
  - Single column that aggregates all of them makes querying a bit difficult, but advantageous for storing in a unified representation
  - Could also look into fact relationship table to be able to add as many modifiers/values as needed in a flexible fashion

How to handle modifiers?

- Proposed approach #1: Note_NLP_modifiers table

Note_NLP_modifiers_id	Foreign key to term mention in Note_NLP
Modifier_concept_id	Foreign key to standard terminology (e.g., “negation_status”, “certainty”)
Value_as_concept_id	Foreign key to standard terminology (e.g., “high”)
Value_as_Number	Float Number (e.g., 30)
Unit_concept_id	Foreign key to unit concepts (e.g., “mg/ml”)

- Pro: structured format for all modifiers
- Con: very large table. Is tradeoff of size of table warranted given frequency of querying these particular modifiers?

How to handle modifiers?

- Proposed approach #2: considering the types of queries on OHDSI-style studies and phenotyping studies, only keep structured a limited set of modifiers, and encode the rest in a string
- Modifiers encoded:
  - Term_exists
  - Term_value (implemented as 3 fields value_concept, value_number, value_unit)
  - Term_temporal

Modifiers – proposed approach #2

Note_NLP_id	Unique identifier for each concept extracted from NLP
note_id	Foreign key identifier to the note the concept was extracted from (Note table).
section_concept_id	Foreign key to predefined concept identifier in the Standardized Vocabularies (LOINC) reflecting the section the extracted concept belongs to.
snippet	Small window of text surrounding term mention
lexical_variant	Raw text extracted from NLP
Note_NLP_concept_id	Foreign key to concept id (Concept Table). Domain concept is provided as part of the Concept table.
NLP_system	String describing system and version used for NLP (data provenance)
NLP_date	Date describing date at which note was processed
Term_exists	Optional boolean; summary modifier that signifies presence or absence of a term for given patient (e.g., not negated, not conditional, not generic, not uncertain → termmention_ispresent=YES)
Value_as_concept_id / Value_as_number / unit_concept_id	Fields describing potential value of term: Optional foreign key to standard terminology (e.g., “high”) / Optional float / Optional foreign key to unit concepts (e.g., “mg/ml”)
Term_temporal	Optional time expression extracted associated to term, “past”, “present”, or “future”

Modifiers – proposed approach #2

- Pro:
  - Single table for all things NLP
  - Common modifiers can be queried easily
  - Keep size of table somewhat reasonable
  - Provide a level of abstraction over "local"/institution specific NLP tables and schemas
- Con:
  - Rarer queries require smarter query functionality to parse the modifier string

To help discussion on modifiers

- Phenotyping use cases
  - Mention of positive concept (not negated, implying specific to the patient, and without any uncertainty, conditional, general)
  - Mention of negated concept
  - Mention of concept with some associated value
  - No mention of concept
- Example:
  - "son has rash"
    term_exist: NO as an optional field
    NLP_modifiers:
    "negated:no,subject=family,certainty=undef,conditional=false,general=false"