OHDSI NLP Working Group
Recommendations for clinical textual data and NLP output storage and representation schema

Latest date: March 17th, 2017
Edited by: Noemie Elhadad, Rimma Belenkaya, Hua Xu

The NLP working group proposes edits to the existing NOTE table and the creation of the NOTE_NLP table.

1. NOTE Table

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_datetime	No	datetime	The date and time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the type, origin or provenance of the Note.
note_class_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the HL7 LOINC Document Type Vocabulary classification of the note.
note_title	No	string(250)	The title of the Note as it appears in the source.
note_text	No	RBDMS dependent text	The content of the Note.
encoding_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the note character encoding type.
language_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the language of the note.
provider_id	No	integer	A foreign key to the Provider in the PROVIDER table who took the Note.
visit_occurrence_id	No	Integer	Foreign key to the Visit in the VISIT_OCCURRENCE table when the Note was taken.

1.1 Proposed Change: new note_type_concept_id

Note_type_concept_id is a foreign key to the CONCEPT table to describe a standardized combination of five LOINC axes (role, domain, setting, type of service, and document kind).
See Section 3. for description of mapping of clinical documents to Clinical Document Ontology (CDO) and standard terminology.

1.2 Proposed Change: note_source_value

This field represents the title of a note. We propose

- changing its name to **note_title**
- extending it to 250 characters
- removing the link to standardization (this now taken care of by note_type_concept_id)

1.3 Proposed Change: new language_concept_id field

Suggest to add a new field of "language_concept_id" to indicate the language of the clinical text.

1.4 Proposed Change: note_text

We propose to modify the type of this field to Varchar(MAX_SIZE), and make it NOT required (for PHI reasons)

2. NOTE_NLP Table (new)

This table will encode all output of NLP on clinical notes. Each row represents a single extracted term from a note.

Field	Required	Type	Description
note_nlp_id	yes	Big Integer	Unique identifier for each term extracted from a note
note_id	yes	integer	Foreign key identifier to the Note table note the term was extracted from
section_concept_id	no	integer	Foreign key identifier to the Concept table.
snippet	no	varchar(250)	Small window of text surrounding term mention
offset	no	integer	Character offset of extracted term in input note
lexical_variant	yes	varchar(250)	Raw text extracted from NLP tool
note_nlp_concept_id	no	integer	Foreign key to Concept table. Represents the normalized concept for extracted term. Domain of the term is represented as part of the Concept table.

note_nlp_source_concept_id	no	integer	A foreign key to a Concept that refers to the code in the source vocabulary used by the NLP system
nlp_system	no	varchar(250)	Name and version of NLP system that extracted the term. Useful for data provenance.
nlp_date	yes	DATE	Date of processing of note. Useful for data provenance.
nlp_datetime	no	TIME	Time of processing of note. Useful for data provenance.
term_exists	no	BOOLEAN	Term_exists is defined as a flag that indicates if the patient actually has or had the condition. Any of the following modifiers would make Term_exists false: Negation = true Subject = [anything other than the patient] Conditional = true Rule_out = true Uncertain = very low certainty or any lower certainties A complete lack of modifiers would make Term_exists true. For the modifiers that are there, they would have to have these values: Negation = false Subject = patient Conditional = false Rule_out = false Uncertain = true or high or moderate or even low (could argue about low)
term_temporal	no	varchar(50)	Term_temporal is to indicate if a condition is “present” or just in the “past”. The following would be past: History = true Concept_date = anything before the time of the report
term_modifiers	no	varchar(2000)	Describes compactly all the modifiers extracted by nlp system. For example, “son has rash” →

			"negated=no,subject=family,certainty=undef,conditional=false,general=false" Value will be saved as one of the modifiers

3. Mapping of clinical documents to Clinical Document Ontology (CDO) and standard terminology

HL7/LOINC CDO is a standard for consistent naming of documents to support a range of use cases: retrieval, organization, display, and exchange. It guides the creation of LOINC codes for clinical notes. CDO annotates each document with 5 dimensions:

· Kind of Document

Characterizes the general structure of the document at a macro level (e.g. Anesthesia Consent)

· Type of Service

Characterizes the kind of service or activity (e.g. evaluations, consultations, and summaries). The notion of time sequence, e.g., at the beginning (admission) at the end (discharge) is subsumed in this axis. Example: Discharge Teaching.

· Setting

Setting is an extension of CMS's definitions (e.g. Inpatient, Outpatient)

· Subject Matter Domain (SMD)

Characterizes the subject matter domain of a note (e.g. Anesthesiology)

· Role

Characterizes the training or professional level of the author of the document, but does not break down to specialty or subspecialty (e.g. Physician)

Each combination of these 5 dimensions should roll up to a unique LOINC code. For example, Dentistry Hygienist Outpatient Progress note (LOINC code 34127-1) has the following dimensions:

According to CDO requirements, only 2 of the 5 dimensions are required to properly annotate a document: Kind of Document and any one of the other 4 dimensions.

However, not all the permutations of the CDO dimensions will necessarily yield an existing LOINC code.[2] HL7/LOINC workforce is committed to

establish new LOINC codes for each new encountered combination of CDO dimensions. [3]

Automation of mapping of clinical notes to a standard terminology based on the note title is possible when it is driven by ontology (aka CDO). Mapping to individual LOINC codes which may or may not exist for a particular note type cannot be fully automated. To support mapping of clinical notes to CDO in OMOP CDM, we propose the following approach.

1. Add all LOINC concepts representing 5 CDO dimensions to the Concept table. For example:

Field	Record 1	Record 2
concept_id	55443322132	55443322175
concept_name	Administrative note	Against medical advice note
concept_code	LP173418-7	LP173388-2
vocabulary_id	LOINC	LOINC

2. Represent CDO hierarchy in the Concept_Relationship table using the "*Subsumes*" – "*Is a*" relationship pair. For example:

Field	Record 1	Record 2
concept_id_1	55443322132	55443322175
concept_id_2	55443322175	55443322132
relationship_id	*Subsumes*	*Is a*

“LP173387-4 Administrative note” “*Subsumes*” “LP173388-2 Against medical advice note”

“LP173388-2 Against medical advice note” “*Is a*” “LP173387-4 Administrative note”

These hierarchies should also be represented in the Concept_Ancestor table.

3. Add LOINC document codes to the Concept table (e.g. Dentistry Hygienist Outpatient Progress note, LOINC code 34127-1). For example:

Field	Record 1	Record 2
concept_id	193240	193241
concept_name	Dentistry Hygienist Outpatient Progress note	Consult note
concept_code	34127-1	11488-4
vocabulary_id	LOINC	LOINC

4. Represent dimensions of each document concept in Concept_Relationship table by its relationships to the respective concepts from CDO. Use the “Member Of” – “Has Member” (new) relationship pair. Using example from the Dentistry Hygienist Outpatient Progress note (LOINC code 34127-1):

concept_id_1	concept_id_1	relationship_id
193240	55443322132	Member Of

55443322132	193240	Has Member
193240	55443322175	Member Of
55443322175	193240	Has Member
193240	55443322166	Member Of
55443322166	193240	Has Member
193240	55443322107	Member Of
55443322107	193240	Has Member
193240	55443322146	Member Of
55443322146	193240	Has Member

Where concept codes represent the following concepts:

Content	Description
193240	Corresponds to LOINC 34127-1, Dentistry Hygienist Outpatient Progress note
55443322132	Corresponds to LOINC LP173418-7, Kind of Document = Note
55443322175	Corresponds to LOINC LP173213-2, Type of Service = Progress
55443322166	Corresponds to LOINC LP173051-6, Setting = Outpatient

554433 22107	Corresponds to LOINC LP172934-4, Subject Matter Domain = Dentistry
554433 22146	Corresponds to LOINC LP173071-4, Role = Hygienist

Most of the codes will not have all 5 dimensions. Therefore, they may be represented by 2-5 relationship pairs.

5. If LOINC does not have a code corresponding to a permutation of the 5 CDO encountered in the source, this code will be generated as OMOP vocabulary code. Its relationships to the CDO dimensions will be represented exactly as those of existing LOINC concepts (as described above). If/when a proper LOINC code for this permutation is released, the old code should be deprecated. Transition between the old and new codes should be represented by “Concept replaces” – “Concept replaced by” pairs.

6. Mapping from the source data will be performed to the 2-5 CDO dimensions.

Query below finds LOINC code for Dentistry Hygienist Outpatient Progress note (see example above) that has all 5 dimensions:

```
SELECT
FROM Concept_Relationship
WHERE relationship_id = ‘Has Member’ AND
(concept_id_1 = 55443322132
OR concept_id_1 = 55443322175
OR concept_id_1 = 55443322166
OR concept_id_1 = 55443322107
OR concept_id_1 = 55443322146)
GROUP BY concept_ID_2
```

If less than 5 dimensions are available, HAVING COUNT(n) clause should be added to get a unique record at the intersection of these dimensions. n is the number of dimensions available:

```
SELECT
FROM Concept_Relationship
WHERE relationship_id = ‘Has Member’ AND
(concept_id_1 = 55443322132
OR concept_id_1 = 55443322175
OR concept_id_1 = 55443322146)
GROUP BY concept_ID_2
HAVING COUNT(*) = 3
```

To identify appropriate dimension while mapping source documents, use the following concept classes:

Note Provider Role
Note Domain
Note Setting
Note Service Type
Note Kind

The proposed approach will ensure that any combination of the 5 CDO dimensions encountered in the source data has a corresponding concept in the vocabulary. It will also support consistent approach to the OMOP CDM/Vocabulary conventions:

· One required _type_concept_id field will be populated in a corresponding domain table, NOTE.

· Vocabulary-related attributes are stored in a vocabulary data model in a uniform way

· Usage of a standard vocabulary, LOINC, is ensured where possible

· Introduction of new OMOP concepts when a standard does not provide adequate coverage of the source data

A similar mapping approach can be applied to labs.

4. Use cases

4.1. Example 1 - Left ventricular ejection fraction

Left ventricular ejection fraction is an important indicator of heart health. It is measured during echocardiogram procedures but also during a range of various procedures. The value is frequently reported in clinical reports and has to be extracted using natural language processing.

Note_NLP_id	123456
note_id	123446425
section_concept_id	<foreign key to "Echocardiogram Report">
snippet	ejection fraction was estimated at 60%
lexical_variant	ejection fraction
Note_NLP_concept_id	<foreign key to "Left Ventricular Ejection Fraction" concept>
NLP_system	EchoExtractor_EF(v.2016)
NLP_date	3/30/16
Term_exists	TRUE
Value_as_concept_id	null
Value_as_number	60.0
Unit_concept_id	<foreign key to "percent">
Term_temporal	present
Term_modifiers	null

4.2 eMERGE Phenoytpes

Existence of specific report or specific note section
1. Presence of a Pathology Report [Appendicitis].
2. Must contain at least two Past Medical History sections and Medication lists (could substitute two non-acute clinic visits or requirement for annual physical) [Hypothyroidism].
3. At least 1 abdominal CT or colonoscopy [Diverticulosis].
4. Patients have to have had a colonoscopy [colonPolyp].
5. Must have at least a problem list and/or note containing non-empty (can say “none”) medication list and past medical history before or immediately after the time of the ECG [QRS].

Term/Concept mentioning in notes or specific sections

6. Positive result of inflammation and non-inflammation concept (CUI) in post-surgical biopsy report [Appendicitis].
7. Reported History of Appendicitis [Appendicitis].
8. Individual's patient chart includes one or mentions of an ADHD or hyperkinesia [ADHD].
9. SSTI cases must have the following or similar keywords in the text results of a bacterial culture lab test, such as skin, wound, boil, abscess, but also recognizing that anatomic sites (e.g. foot/hand/leg/buttock, etc.) [caMRSA].
10. At least on diagnosis code for C. diff and at least one affirmative mention of C. diff infection (unqualified by negation, uncertainty, or historical reference) in progress notes [CDiff].
11. Retrieve DSM-IV Symptom criteria (Social Interaction/Communication/Behavior, Interests and Activities) terms from notes to confirm Autism [Autism].
12. Patient has colonoscopy without positive mention of diverticulosis as control [Diverticulosis].
13. Positive mention of HF in the problem list through either NLP or structured problem list [HeartFailure].
14. Cases are those that have polyps in any of their colonoscopy or associated pathology reports [colonPolyp].
15. Notes contain no evidence of heart disease concepts (NLP for notes, Problem Lists at or near ECG time, ignoring Family Medical History and Allergy sections (using section tagger), ICD9 and CPT codes at or near ECG time describing heart disease) before ECG time or within one month following [QRS].

Related terms mentioning in the same line or adjacent lines

16. Potential cases were identified if they contained at least one term from List 1 (terms identifying an ace-inhibitor, see below) AND List 2 (terms identifying cough, see below) one the same line (e.g., sentence) within the "Allergy section", "Medication section" or within the entire "Patient summary section" of the EMR [ACEIcough].
17. At least one non-negated "Disorder related terms" mention and "Anatomical site related terms" mention either in the SAME or adjacent sentences in a 'section of interest' [VTE].

Numeric values with/without temporal constraints

18. Exclude all patients with an Ejection Fraction (EF or LVEF) <35% within 1 year before or after meeting the CASE 1 definition [ResHTN].
19. Have evidence from a carotid imaging study of >50% carotid artery stenosis (at least unilaterally) [CAAD].
20. Classify the type of HF using the numeric EF results (use the lowest EF recorded in the time window) [HeartFailure].

21. In defining “Normal” ECG, QRSd between 65-120ms, ECG designed as “NORMAL”, Heart Rate between 50-100, ECG Impression must not contain evidence of heart disease concepts [QRS].