Representation of Unstructured Data Across Common Data Models (DI2)

Final Report – Best Practices for Representing Unstructured Data in the Sentinel Common Data Model (SCDM)

Prepared by: Keith Marsolo, PhD,[1,2] Ruth Reeves, PhD;[3] Li Zhou, MD, PhD;[4] Lesley Curtis, PhD;[1,2] Tyler Erikson, MS;[2] Judy Maro, PhD;[5] Kathleen Shattuck, MPH;[5] Jill Whitaker, MSN, RN-BC;[3] Tina French, RN, CPHQ;[3] Liz Hanchow, RN, MSN;[3] Suzanne Blackley, MA;[4] John Laurentiev, BS;[4] Sarah Dutcher, PhD, MS;[6] Efe Eworuke, PhD;[6] Aida Kuzucan, PharmD, PhD;[6] Joseph Plasek, PhD;[4]

Author affiliations: [1]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC; [2]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC; [3]Vanderbilt University Medical Center Department of Biomedical Informatics, Nashville, TN; [4]Harvard Medical School and Brigham and Women's Hospital, Boston MA; [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; [6]US Food and Drug Administration, Silver Spring, MD

Version 1.0
January 31, 2023

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

Representation of Unstructured Data Across Common Data Models

Final Report - Best Practices for Representing Unstructured Data in the SCDM

Table of Contents

History of Modifications

Version	Date	Modification	Author
0.1	10/31/2022	Original Draft	Keith Marsolo & Project WG
0.2	12/12/2022	Revision to reflect feedback from WG members and other stakeholders	Keith Marsolo & Project WG
1.0	1/31/2023	Final version based on feedback from FDA and WG members	Keith Marsolo & Project WG

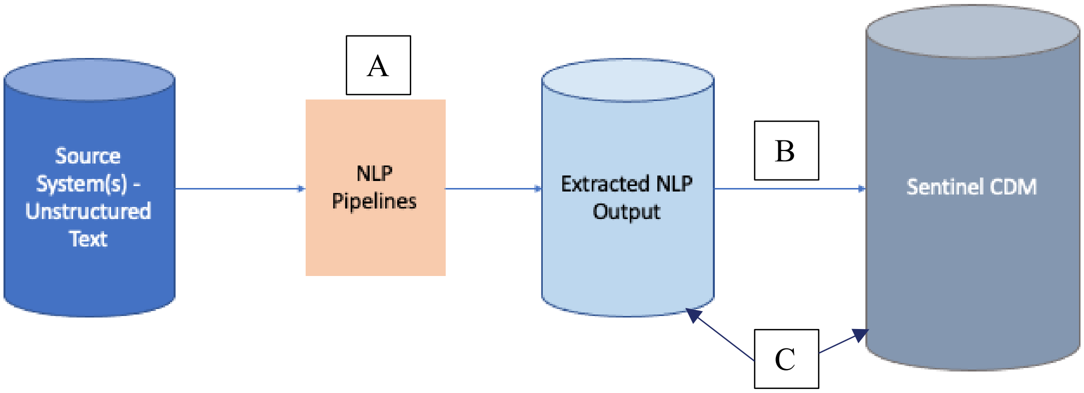
Introduction

The overarching goal of the "Representation of unstructured data across Common Data Models" project is to provide guidance to the Sentinel Network on how best to incorporate information derived from unstructured data into a Common Data Model (CDM) framework. There are three main project objectives, which are to: 1) identify the priority data elements or concepts that are important for pharmacoepidemiological safety studies that FDA could potentially ask data partners to extract from unstructured data; 2a) survey the natural language processing (NLP) solutions that are in use across the Sentinel ecosystem; 2b) assess the overall availability of priority concepts (e.g., medication exposure, smoking status) within unstructured data at two different Data Partners; and 3) develop recommendations on how to best represent natural language processing (NLP)-derived data elements within the Sentinel CDM (SCDM).

This report describes activities related to the third project objective, the development of recommendations for representing NLP-derived data elements in the SCDM. This includes a summary of approaches for representing NLP concepts (derived data elements) within CDMs, as well as suggested best practices for the FDA's Sentinel system. The image below illustrates the general process of transforming unstructured text to records within a CDM.[1] Unstructured text is processed through one or more NLP pipelines, generating a set of extracted NLP outputs (e.g., presence or absence of specific concepts). Algorithms can be executed on these outputs in order to derive or compute records that then are stored in the SCDM (e.g., presence of concepts X, Y, and Z indicate severe disease, while presence of concepts X and Z only indicates mild disease). We focused on 3 aspects of the process, labeled A, B, and C in the image:

A. If Sentinel or a Data Partner were choosing an NLP pipeline to implement locally, what information is available related to performance, and how does that compare to other published studies on NLP performance? This information is presented in the form of a literature review and can be found as a separate document.
B. Considerations when creating derived records in the SCDM from NLP outputs.
C. Approaches for representing and integrating NLP outputs and NLP-derived records within the SCDM

Additional detail is provided on Topics B and C in the text below.



[1] Image adapted from Hua Xu webinar - Representing and Utilizing Clinical Textual Data for Real World Studies: An OHDSI Approach (https://www.sentinelinitiative.org/news-events/meetings-workshops-trainings/representing-and-utilizing-clinical-textual-data-real)

Considerations when creating a derived record from NLP outputs

It is possible to generate a range of outputs from NLP pipelines, which can be combined in multiple ways to create derived records within the Sentinel Common Data Model (SCDM). There are a number of factors to consider on how best to handle this derivation process. It is important to note that there is not necessarily a right or wrong answer. It is best to optimize for the typical Sentinel use cases, with the recognition that this optimization may make support of non-standard use cases or alternatives more costly or cumbersome in the future.

Factors to consider:

- *What is the general objective of NLP within Sentinel?* – To date, the use of NLP within Sentinel has been primarily to extract health outcomes of interest (HOI) that cannot be captured well in administrative claims or common structured EHR data. The algorithms used to derive these outcomes are sophisticated, like a computable phenotype, and are analogous to a trained reviewer going through a patient's chart and deciding about the presence or absence of that outcome due to information within the clinical narrative. We expect these activities to continue going forward.

  There are other uses of the NLP that do not exactly fit this model. For instance, Data Partners could use NLP to derive information that is commonly found in structured data but can also be found in the clinical narrative (e.g., medication exposure). There is an added complexity in this use case, in that unstructured text may represent information that was pulled into a note at a certain point of time (e.g., medications administered in an inpatient setting), and may not exactly match the corresponding information that is stored in structured fields, due to dose modifications or other changes. Discrepancies often exist among different data sources and even within the same data source (e.g., different data fields or time frame). Reconciliation of these differences can be important, as the narrative may reveal adherence issues, medications discussed but not prescribed, or other insights that are not readily apparent from structured data.

  NLP can also be used to extract information that is commonly found in semi-structured formats (e.g., echocardiogram results, radiology findings), but that may also be recorded in structured fashion within certain Data Partners. As EHRs evolve, data elements or domains that fall into the latter category may begin to look more like the former at a network level. The documentation of smoking status within the EHR is one such example, with structured fields added over the course of several years to capture smoking status in ways that were compliant with Meaningful Use regulations. The addition of fields to capture patient-level social determinants of health is another example. NLP can be useful during this transition to structured documentation. For instance, until such variables are reliably captured as part of clinical workflow, NLP can help reduce missing values by extracting corresponding information from the clinical narrative. The working assumption is that Sentinel would like to support all three of these use cases: algorithms to derive health outcomes of interest; NLP to extract information that can also be found in existing data domains; NLP to extract semi-structured information that is not routinely found in structured data, but that the first one would remain a priority. Certain design choices proposed within this document may not be necessary if some of the other use cases are ultimately considered out of scope.

- *Should all HOI algorithms be required to use the same NLP pipeline? Are Data Partners free to choose their own to implement locally?* Prior Sentinel projects that utilized unstructured text have leveraged a variety of natural language processing tools, as shown in the table below.

Table 1: Prior Sentinel projects that utilized unstructured text. Also shown are the concepts of interest and the NLP techniques that were used to extract information.

Project Title	Concept of interest	NLP technique(s)
Validation of Acute Pancreatitis Using Machine Learning and Multi-Site Adaptation for Anaphylaxis	Anaphylaxis, acute pancreatitis	MetaMap, ETHER
Advancing Scalable Natural Language Processing Approaches for Unstructured Electronic Health Record Data	COVID-19; COVID-19 Severity	MetaMap, PheNorm
Improving Probabilistic Phenotyping of Incident Outcomes through Enhanced Ascertainment with Natural Language Processing	Suicide	Word2vec, lexical association
Enhancing Causal Inference in the Sentinel System: An Evaluation of Targeted Learning and Propensity Scores for Confounding Control in Drug Safety	NLP used in the conduct of the study, but not the main focus of development	Bag of words

Adopting a common NLP pipeline (or set of pipelines) for use in developing and deploying NLP algorithms across Sentinel would lead to efficiencies across the network, particularly when working with commercial Data Partners. Utilizing multiple pipelines may require multiple software licenses, security reviews, etc., increasing the implementation time and cost. There may be methodological or performance reasons why a project team would choose to use a particular pipeline or set of pipelines when developing a new algorithm or determining feasibility, but in the long-term, Sentinel could consider standardizing to a subset of tools when it comes to implementing or validating algorithms at scale.

At the same time, when surveying Sentinel Data Partners and those affiliated with the Innovation Center about their current use of NLP (objective 2a of the project), we found a wide variety of tools in use, including SAS, locally developed Python scripts or other in-house tools, Health Discovery (from Averbis), n-gram models, cTAKES and CLAMP. This demonstrates that NLP experience exists within the Sentinel ecosystem, but there is little in the way of commonality and consensus about standard approaches. It may take time to move the network to a standard set of tools, so Sentinel could decide that for the time being, Data Partners may choose to use their own local pipelines to process notes. If that is the case, then it may be worthwhile for Data Partners to validate their pipeline on a set of representative notes and report their local performance on a common set of measures. Even if all Data Partners use the same pipeline, it may still be worthwhile to conduct this validation to understand the variation across the network.

One potential framework for reporting results has been defined in Velupillai et al,[2] which includes data source characteristics (e.g., type, content, size, sampling characteristics), the approach to NLP (e.g., task, approach, parameters, gold/silver standard used) and evaluation criteria (e.g., method, metric, results). Encouraging the sharing of data

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6986921/

product (Enabling the Sentinel Data Partners to share data products (e.g., models, NLP tools, lexicons, ontologies, intermediate results, etc.) could enhance the generalizability of models trained across multiple institutions and allay concerns about reproducibility.[3]

- *What is the starting point for each NLP project?* Most NLP projects begin with the raw unstructured text and derive the concepts of interest, but as pipelines become more sophisticated, it is possible to ask Data Partners to pre-process certain notes so that "standard" concepts are extracted and readily available for new analyses (e.g., medications, procedures, signs/symptoms). The PheNorm approach, leveraged by the "Scalable NLP" Innovation Center project is essentially designed to support this type of model, where key terms/phrases are identified from "knowledge sources" like Wikipedia entries or scientific manuscripts and then extracted from unstructured text. Pre-extracting terms would dramatically shorten the overall development process. While there may always be a need for a project to process notes specifically for their analysis, either for an emerging disease (e.g., COVID-19) or to look at adverse events for a medication that is new-to-market, pre-processing the notes could decrease the cost and time needed for algorithm development. If the notes are pre-processed, one option may be to limit the use of modules within NLP pipelines to those activities with higher order logic, such as only positive mentions of concepts that are current/active. Analyses that require more complex use of NLP (e.g., negations, hypothetical, history or determining anatomical location) could start from the raw unstructured text to deploy custom module selections.

- *Will development and implementation of all NLP-based HOI algorithms be driven by Sentinel, or will external contributions be incorporated*? The use of NLP in Sentinel has been driven by FDA priorities. While we expect this to continue to be the case, there is a tremendous amount of NLP expertise that exists outside of Sentinel that could also be leveraged. If Sentinel defined an expected level of validation or rigor around external algorithms, many researchers may choose to make them "Sentinel ready." While this community-driven approach is somewhat of a departure from existing practices, it might help lower the costs and timelines around the development and deployment of new algorithms.

Approaches for representing NLP outputs and NLP-derived records within the SCDM

As Sentinel works to incorporate NLP into network analyses, standardized approaches are needed to store both the outputs of NLP pipelines and NLP-derived records within the SCDM. The NLP outputs may not be directly used in Sentinel analyses, but the expectation is that NLP-derived records would be incorporated into the core SCDM. In the section below, we describe some of the considerations for the structure of the various tables needed to represent information extracted by NLP in relation to other Sentinel data model needs. We then follow with a discussion on different terminologies that can be used to represent these data.

Storing NLP-derived records within the Sentinel CDM

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284141/

The current version of the SCDM (v8.1) does not include any tables or fields to specifically handle NLP-derived records. A draft table specification has been defined to handle "Engineered Features." This table structure is shown below.

FEATURE ENGINEERING Table Content (SCDM draft)

Concept	Description / Comments
Patient Identifier	Unique identifier for each patient
Replicated fields related to ENCOUNTER and PROVIDER details	Fields copied from ENCOUNTER (Encounter ID, Admit Date, Encounter Type) and PROVIDER (Provider ID) to simplify the analytical query process. This is a common feature of many SCDM core tables to avoid additional table joins as part of the query process.
Fields for Feature value and Term Type	Code for the Engineered Feature and the corresponding terminology type.

This table structure is sufficient for Sentinel's limited NLP efforts to date but will likely need to be extended to handle more NLP-derived features. For instance, the proposed table can handle positive mentions of an engineered feature (i.e., anaphylaxis), but nothing about the status (active, inactive, resolved). Nor can it handle negative mentions (e.g., denied, rule out, not present, discussed), "experiencer" (e.g., family history mention relates to relative rather than patient) or other potential modifiers (e.g., disease severity or temporality). If Sentinel chooses to move beyond the use of NLP to identify "ever" events (e.g., anaphylaxis), this additional metadata becomes critical. Supporting items like status and temporality adds a degree of relativism to each record (e.g., active as of a certain time point), so additional fields could be considered to capture information like the timestamp of a positive or negative mention. Decisions on which records to incorporate in an analysis could be made by investigators in the initial design stage.

Use of NLP in other CDMs

Of all the CDMs routinely used in clinical research today, the Observation Health Data Sciences and Informatics (OHDSI) CDM is the most robust in terms of representing NLP outputs and will be used as the motivating example (it is possible to create similar representations within the PCORNet CDM, but there are no specific tables as with OHDSI). Within OHDSI, there are tables to store both note text (NOTE) and the outputs of NLP pipelines (NOTE_NLP). These tables are summarized below (details obtained from v5.4 of the OHDSI CDM specification- https://ohdsi.github.io/CommonDataModel/cdm54.html). Unless specified as Required, all fields are optional. Within the OHDSI CDM, concepts labeled as "Standardized metadata" are assigned values based on the OHDSI standard vocabularies.

NOTE Table Content (OHDSI v5.4)

Concept	Description / Comments
Note Record Identifier	Unique identifier for each note (Required).
Foreign keys to link to other tables within the CDM	Foreign keys exist for PERSON, PROVIDER, VISIT_OCCURRENCE and VISIT_DETAIL tables (Required).
Date and time Note was generated in the source system	Date is a required field, but time is optional. Time is defaulted to midnight if missing.

Standardized metadata about the note – Provenance and Document Type	Provenance would typically be "EHR." Document type is assigned from HL7 LOINC Document Type Vocabulary (Required). *EHRs do not natively assign an HL7 document type to each note/note type, so this mapping is done manually.*
Standardized metadata about the note – Encoding and Language	Encoding would typically be something like 'UTF-8' or 'ASCII'. (Required)
Title of the Note	Note title, if available
Note content	Raw text of the note itself (Required).
RAW values of the Note Type prior to mapping	Allows for later verification of the Note Type mapping, if required.
Links to additional records if in the CDM	If note record is related to another record in the CDM, fields exist to store the primary key and ID of the corresponding

NOTE_NLP Table Content (OHDSI v5.4)

Field	**Description**
NOTE_NLP Record Identifier	Unique identifiers for the record of NLP output (Required)
Foreign key to link to the Note table	ID of the Note associated with the NOTE_NLP record. (Required).
Standardized metadata about NLP record – Section	Section of the Note from which the NOTE_NLP record was extracted. *The existing OHDSI concepts for this value set may not adequately represent all possible note/section types.*
Metadata associated with the extracted text – text snippet, offset, lexical variant	Text snippet represents the small window of text surrounding the extracted term. Offset represents the character offset of the extracted term and Lexical Variant is the raw text extracted from the NLP tool (Lexical variant is Required).
Fields for standardized representation of the NLP concept – Code and Term Type	Term type and Code assigned based on the extracted NLP output (e.g., ICD-9 255.0). In the OHDSI CDM, all outputs are assigned a standardized code.
Metadata about the NLP pipeline and processing steps – NLP system, date and time	NLP system represents the name and version of the NLP pipeline that was used to extract the term. Date corresponds to the date when the note was processed (Required). The Time the note was processed can be recorded as well.
Metadata about the NLP concept – Exists, Temporal, Other Modifiers	Exists is a flag to indicate patient has or had the condition in question; Temporal is used to indicate if a condition is present or just in the past; Other Modifiers are used to store additional information about the extracted concept (negation, subject, uncertainty, etc.). In OHDSI, these terms are concatenated together into single string.

Within the EHR, a note typically consists of information recorded by a provider at a single point in time (e.g., nursing notes), or constitutes the summary of a procedure (e.g., pathology reports), though they can also contain information from prior notes (using copy-paste or pull-forward functionality). Defining a NOTE table within the SCDM (or as a supplemental table if it not considered part of the core CDM) may not be necessary if there is some flexibility in how partners implement their NLP pipelines locally and if there is no expectation that the raw source data be available for external analyses. Even so, a common structure would simplify cross-network activities, even if there is some local variation on implementation.

A version of the NOTE_NLP table would be useful for Sentinel NLP activities, however decisions are needed in two major areas in order to finalize any table structure. The first is whether the text snippets / lexical variants (i.e., raw text from the note associated with the extract concepts) will be stored in the table along with the corresponding code / term type. This can be useful when verifying or validating the output, but there is a small chance that the extracted text can contain personal health information (PHI). Given that Sentinel activities are considered Public Health Surveillance, there are fewer regulatory hurdles to accessing PHI than in a typical research project, but there still may be concerns about patient privacy or institutional risk. It is not strictly necessary to include these data in any Sentinel-facing table but asking Data Partners to at least keep these data available locally behind institutional firewalls may be a suitable compromise. Another alternative would be to pass the text snippets through a de-identification tool to automatically redact PHI, though this is another NLP pipeline that will need to be created, validated and maintained. The second decision is around the metadata to include with the extracted concept (e.g., exists, temporal, other modifiers as noted above). In the annotation tasks for this project, we defined general attributes of "Assertion," with values for positive, negative, uncertain and hypothetical, and "TimePerspective," with values of current, history and predicted. We also defined relationships between primary and secondary classes (i.e., Medication and Dose, Cancer and Stage). Those secondary classes (i.e., Dose, Cancer Stage) could be considered modifiers of the original concept and included as part of that record. They could also be represented as completely separate records with another field in the table that links them. The former approach makes it easier to retrieve everything associated with a given concept, but the latter allows for more nuanced representations, such as when information on cancer staging is present but not a corresponding description of the cancer). Both primary and secondary class concepts may have temporality, severity, negation or other modifiers, so figuring out how to best store this chain of metadata appropriately is necessary to have the full context and extract the proper meaning. There are pros and cons to both representations in terms of record/table size and ease of querying, and the ultimate format should be driven by the main analytical use cases. These use cases should also drive the decision-making on what metadata to routinely include, the different value sets, whether to represent them as separate fields or as other formats (i.e., CSV, JSON, XML), etc.

Representing conditions

One potential option to consider in extending the SCDM (as of v8.1) is a stand-alone table to represent Conditions (e.g., disease, medical condition, symptom, or co-morbidity). Conditions could be patient self-reported or recorded by a provider in a healthcare setting and may include signs or symptoms. They could be sourced from structured fields (e.g., problem list, review of symptoms) or unstructured text notes. They are distinct concepts from diagnosis codes assigned as part of clinical care. The OHDSI CDM groups traditional diagnoses with conditions into a common table (CONDITION_OCCURRENCE), while the PCORnet CDM maintains a separate table given the different analytical meaning of these data streams. Many of the health conditions used in Sentinel studies, including indications, co-morbid conditions and HOIs can be viewed as Conditions, particularly those extracted via NLP, so having a way to represent this information could lead to ease in downstream access for certain analytic activities. However, it should be noted that analyses of regulatory importance in Sentinel typically require highly specific information regarding presence as well as onset timing for health conditions of interest, especially when they are used as outcomes. Therefore, assessment of validity for the various modes of capturing conditions in a Conditions table merit special attention to meet Sentinel specific use cases. As addition of a Conditions table to SCDM likely requires significant resource commitment by data partners, and prioritization of conditions that are captured may also need to be considered to focus on a limited set of conditions which are not captured well in other tables using structured data.

As one possible example, the format of the PCORnet CONDITION table is outlined below. While some of the fields captured in this table could be applicable for Sentinel, the exact structure of the table should be tailored to best support Sentinel analyses. One notable gap is that the PCORnet CONDITION table can only handle variations of a positive association (e.g., active, inactive, resolved). It does not handle negative indications (e.g., ruled out, not present) or other modifiers (e.g., disease severity). If there is an effort to include a CONDITION-type table in the SCDM, it will be important to decide whether to support such modifiers within the table, or simply create additional codes to denote the different statuses (i.e., a code for positive COVID-19; a code for negative COVID-19; a code for each different severity of COVID-19, etc.). This decision will also be driven by whether a standard terminology is used to represent conditions (e.g., SNOMED, ICD) or if Sentinel will rely more on custom code sets.

CONDITION table content (PCORnet v6.0)

Concept	**Description / Comments**
Condition Record Identifier	Unique identifier for each condition record
Foreign keys to link to other tables within the CDM	Foreign keys for PATIENT and ENCOUNTER tables
Dates related to the Condition – Report Date, Onset Date, Resolved Date	Report date represents the date the condition was recorded/noted. Onset date corresponds to the date the condition started. The Resolved date is the date at which the condition resolved (if applicable).
Status of the Condition	Indicate whether the Condition is Active, Inactive or Resolved. *Can add other options if needed.*
Condition code and term type	Code and Terminology used to represent the Condition. Example terminologies include SNOMED, ICD, Human Phenotype Ontology, Algorithmic (e.g., assigned by computable phenotype).
Provenance	Source of the condition. Examples include patient self-report, healthcare setting (i.e., EHR problem list), PCORnet algorithm (for network-wide activities), registry, derived (local algorithmic work), and flavors of null (Other / Unknown / No information).
RAW values of the Condition and metadata	Raw source values of all fields prior to mapping to the PCORnet CDM.

Representing data provenance

Many CDM tables that store similar data from multiple sources contain provenance fields to distinguish records (e.g., diagnosis extracted from unstructured EHR text, diagnosis entered by clinician into structured EHR fields, diagnosis generated from EHR billing system, diagnosis from health plan claim). The SCDM does not yet incorporate provenance fields because most of the data tends to come from a common source (health plans) that do not include variations of the same domain (e.g., ordered and billed procedures) or in cases where there are variations of similar data, there are analytical reasons to keep them separate (i.e., separate tables for prescriptions, dispensing and inpatient administrations). As Sentinel works to incorporate more EHR data into the SCDM, and as more NLP-derived records are added, including provenance fields would allow data to be stored in the most logical table, while also ensuring that those records can be readily identified so they can be included or excluded from an analysis. Provenance fields can also be helpful in model building for determining the reliability or stability of a given data source in evaluating model performance. Adding provenance fields to the SCDM is non-trivial exercise, however, with several downstream effects. Beyond the work required of Data Partners to add these fields to their extract-transform-load procedures, data checks will need to be defined to verify that these fields are appropriately populated, and perhaps most importantly, all of the Sentinel analytic tools will need to be updated to query or filter records based on these fields (this is true of the other changes discussed in this report as well (e.g., Condition table), but adding provenance fields is potentially of greater magnitude as they are part of all (or most) tables).

If such fields are not defined, Sentinel will still need to make decisions on how EHR-based Data Partners should handle data provenance. At a minimum, it will be necessary to define provenance in some fashion to provide guidance to these Data Partners as what to load or exclude in their SCDM and to verify that they have correctly followed that guidance (e.g., load clinician-entered EHR diagnoses only and exclude records from the billing system). Another option is to begin creating EHR and/or NLP-specific versions of most tables (e.g., NLP_LAB_RESULTS), which simplifies the process of populating the SCDM, but may complicate analyses by having to check multiple tables for the presence of an observation.

Within the PCORnet CDM, provenance values are defined based on the potential sources to a given table. Many of the traditional data domains (e.g., diagnoses, procedures) contain 5 values – EHR, EHR billing, claim, derived (e.g., NLP), flavors of null (Other/Unknown/No Information), but other domains, such as patient vitals, may contain others – patient-reported, patient device feed, EHR, healthcare device feed, derived, Other/Unknown/No information. OHDSI has defined a common value set for all provenance fields (that can be filtered by type), but because this value set is fairly large (79 as of October 2022), it can be difficult for Data Partners to find and select the most appropriate value. If Sentinel opted to have a common value set for all provenance fields, the recommendation would be to provide specific guidance to partners on which options to consider for each table. This is an area that could be piloted by sites within the Development Network to determine the most appropriate approach for Sentinel.

Terminologies used to represent NLP outputs

Many of the more popular NLP pipelines in use rely on the Unified Medical Language System (UMLS) as their underlying “dictionary.” Extracted text terms are typically mapped to a UMLS concept, coded by a concept unique identifier (CUI), which supports further mapping to codes from standardized vocabularies (e.g., SNOMED-CT, RxNORM, LOINC), depending on the concept. Since a UMLS concept typically maps to many different terminologies, it is often possible to represent a concept at multiple levels of granularity. In general, it is best to select the terminology that most appropriately represents the concepts of interest. For instance, SNOMED-CT may be ideal if concepts are being derived from a review of symptoms or a patient history. However, if concepts were extracted from a structured list of diagnoses that originated in ICD-9 or ICD-10, then the use of SNOMED-CT may result in the generation of overly specific codes.

SNOMED CT vs. MedDRA

Adverse events in clinical trials are typically coded in MedDRA. MedDRA is not typically used within the EHR, so there is no structured data within the EHR that would be natively coded to MedDRA. MedDRA is a terminology that is supported within the UMLS, so any tools that support UMLS should be able to generate text terms that have been mapped to MedDRA. If

Sentinel always asks Data Partners to extract concepts from scratch, or there are efforts to explicitly generate outputs in MedDRA, then this should not be an issue. The use of SNOMED-CT is more prevalent in the NLP community when representing the same types of concepts that might typically be coded to MEDRA. If data have been pre-processed with SNOMED-CT, it may be time or cost-prohibitive to rerun everything through a pipeline. There are mappings between MedDRA and SNOMED-CT, but they are not 1:1, and since the lower-level terms are not always the same, higher levels of the hierarchies may not be equivalent.[4]

Conclusion

The specific design of any SCDM updates should be driven by the common Sentinel NLP use cases and approaches to NLP in general. While there are efficiency gains in having Data Partners adopt the same NLP pipeline (e.g., minimizing number of software licenses, security reviews, etc.), there may not be sufficient consensus as to the most appropriate solution for the Network. In either case, Sentinel should consider having Data Partners validate their pipeline on a set of representative notes and report their local performance on a common set of measures. Doing so will provide a sense of the variation across Data Partners.

The OHDSI CDM provides a strong foundation for the representation of NLP outputs, though the Standardized Vocabularies for Document Type, Section Type, etc. may not fully encompass all of the different notes available across Sentinel Data Partners. The decision about which concept modifiers (e.g., assertion, time perspective) to include as separate fields and which to store in a container-type format (e.g., CSV, JSON, XML) should be made to best facilitate the process for deriving records in the SCDM and avoid extraneous data manipulation. The same applies for the underlying terminologies used to represent the concepts.

The expansion of existing SCDM tables to include data provenance will likely be necessary to allow NLP-derived records to be properly labeled, and most productively deployed in model-building and other automated reasoning systems. As a stopgap, separate tables can be created specifically for NLP concepts, but this may be unsustainable as the number of concepts increases. Finally, a more general table to store “conditions” or other outcomes of interest may be needed, or at least an expansion of the “engineered features” table being considered for the SCDM. Specifically, the need to store different codes and code types, statuses, start/end dates, etc., may prove to be a valuable feature for future analyses. Small scale pilots to test this process from end-to-end will help determine the most appropriate data model design.

[4] https://forums.ohdsi.org/t/relation-between-meddra-and-snomed/6556/19
https://www.meddra.org/how-to-use/basics/hierarchy