Representation of Unstructured Data Across Common Data Models (DI2)

Executive Summary

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

Executive Summary

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	1/31/2023	Final version created based on content of underlying reports	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).

Objectives 1 and 2a are summarized in the report "**Identification of Priority Concepts and Natural Language Processing (NLP) Capabilities Survey**." The findings of Objective 2b are outlined in "**Availability of Priority Data Elements from Chart Annotation**," and the recommendations of Objective 3 can be found in "**Best Practices for Representing Unstructured Data in the Sentinel Common Data Model (SCDM)**" and "**Natural Language Processing of Unstructured Text in Electronic Health Records at Scale - A Systematic Review."** An Executive Summary of the reports is provided below.

Identification of Priority Concepts

To generate a list of priority concepts to extract from unstructured text, we started with a list of concepts that could be extracted using existing NLP solutions and then asked FDA to add any that might be missing. Two systematic reviews were consulted to generate an initial list and then workgroup members were asked to provide additional suggestions. The intent was not to identify all possible solutions, but rather identify some of the more popular packages in use today that could serve as a baseline for current capabilities. We sought to generate a set of concepts that could be extracted using the identified NLP solutions. We focused on broad categories, not specific items, unless they were called out in the reference documentation (e.g., medications as a concept, not aspirin). The goal was to generate a "good enough" set of concepts, stopping when we reached saturation. FDA was provided with the list of concepts to identify any that were missing, and to assign a priority ranking to each one (high, medium, or low). Highest priority was given to those concepts that are not easily obtained from administrative claims data that are informative for drug safety studies. Concepts that can readily be obtained from administrative claims were assigned a low priority, even if they were important for drug safety studies.

FDA was not asked to rank order the different NLP concepts, but rather to assign an overall priority. This was done purposely to allow for the identification of important concepts and to assess the ability of the community to extract them using existing NLP solutions. Example concepts that were rated high priority include those related to cancer pathology, signs/symptoms, severity, scale and time period for conditions, medical history, genomic information, and attributes related to medications. Additional concepts that were identified include timing and duration of a medication, indication, physical findings, oxygen support and death date and cause. We found that there is a high degree of overlap between the high/medium priority concepts identified by FDA and the capabilities of existing NLP solutions. In addition,

some of the newly added concepts, may already be part of existing solutions, since they can be considered "relationships" that were not readily available via publicly available project websites. Other new concepts, like oxygen use, are considered high priority for a number of other research initiatives, particularly those related to COVID-19, so the ability to extract those data will be part of existing solutions soon, if they are not already.

NLP capabilities survey

Once the priority concepts were identified, a survey was developed to assess the NLP capabilities of partners within the Sentinel ecosystem, in terms of the tool(s) used, the notes processed, context of use (e.g., study-specific research use, to support clinical operations), concepts extracted, etc. The survey allowed us to understand how well the current state of NLP use aligns with the FDA's priorities. Two additional concepts were suggested by the workgroup during the survey development that were not part of the prioritization process – social determinants of health and the ability to detect/assign relationships between concepts. The survey was distributed in a manner that was compliant with the Paperwork Reduction Act. The survey was distributed to 14 Sentinel Data Partners & 8 partners affiliated with the Innovation Center. A total of 17 responses were received by the survey deadline (13 from Sentinel Data Partners). Of the respondents, 12 report using NLP in some capacity, with half using it for project-specific research and half for research and "operational" purposes.

The uptake of NLP solutions across partners varied greatly, with roughly 1/3 of respondents reporting no NLP use, 1/3 reporting use only for project-specific research purposes, and 1/3 with the ability to support more routine use (e.g., extracted information available for use in multiple projects/purposes). There is very little commonality in terms of solutions that are employed. Almost every Partner reported a different approach. The scope of concepts extracted via NLP also varied widely. Diagnoses represent the highest percentage concept, with 9 of 12 reporting the ability to extract them. A handful of other concepts can be extracted by >50% of respondents (e.g., cancer site and histology, smoking status, signs, and symptoms), but most concepts are only extracted by a small number of partners.

Availability of priority data elements from chart annotation

We annotated the charts of selected patients to summarize the availability of priority concepts identified in Objective 1. We focused our annotation efforts on two use cases – hospitalized patients with COVID-19, and cancer. Annotation occurred at two Data Partners. For each use case, a general population definition was drafted, as well as a strategy for selecting patients and notes. Within the COVID-19 cohort, the primary annotation task examined the discharge summary associated with the COVID-19 hospitalization, annotating the existence of priority concepts. We chose the discharge summary because if Sentinel were going to complete an analysis of inpatient encounters that relied on NLP concepts that had already been extracted, discharge summaries would be more likely to have been processed than other specialty note types, even if those specialty notes would be more likely to include documentation related oxygen use. Stratifying by billing codes for supplemental oxygen would ensure there is a mix of patients who did and did not receive oxygen compared with a purely random sample of hospitalized patients. As a secondary activity, we also had each site run a query to quantify the notes that include keywords related to oxygen use (number of notes by type) for the patients in the COVID-19 cohort. The primary annotation task for the cancer use case was to annotate selected concepts of interest found within the physician / clinic note that is associated with the visit where the patient was prescribed darzalex. The concepts of interest included those priority

concepts likely to be present in the note, as well as those that were associated with the label for darzalex. This would allow us to determine whether the physician note contained sufficient information to determine the specific indication.

Draft query code was developed to assist in the process of identifying patients, but because the clinical notes were obtained from local systems, each site tailored the process to fit their local environment. Annotations were completed using the extensible Human Oracle Suite of Tools (eHOST) software package. An annotation guide was drafted to assist in this task. Each of the main priority concepts (e.g., medications) were defined as a "primary class," with additional "attributes" to indicate other metadata (e.g., positive mention, historical / resolved status). Some primary classes also had associated "secondary" sub-classes (e.g., medication dose). When present in the same portion of the text, these primary and secondary classes could be linked with a relationship. The annotation guide included examples of the text that would be included for each class as well as guidance on how to proceed in certain instances.

Within each cohort, each team of two annotators was asked to double-annotate a set of notes (5-6 in total) and compute the inter-annotator agreement at the class / attribute level. If the overall percentage was above 80%, that site could proceed and single-annotate the remainder of the notes within that cohort. Descriptive statistics were computed on the single-annotation notes, summarizing the characteristics of the "primary" class concepts and the associated attributes across Data Partner sites, along with the secondary class concepts. For the primary medication class, we also computed statistics on the presence of associated secondary classes (e.g., how often was dose recorded alongside medication, alone and in combination with route, frequency, etc.). The COVID-19 cohort numbers are broken out into two categories: patients WITH a billing code for supplemental oxygen (oxygen cohort) during the hospitalization, and patients WITHOUT a billing code (non-oxygen cohort).

Annotation results

When looking at the annotations for the COVID-19 cohort, most of the primary concepts (e.g., medications, conditions) are present in almost every note across both Data Partner sites, and the percentage of overall annotations that correspond to those concepts are also similar (e.g., conditions ~60%, medications ~15-20%). When there are differences across Data Partner sites, such as with discharge disposition, the percentages within site between the oxygen and non-oxygen groups are more similar, which may be an indicator of the difference in documentation patterns. Some of the biggest differences across Data Partner sites are seen in the overall level of documentation of oxygen support, though there are again similarities when looking just at positive mentions of oxygen support. For instance, 100% of the patients in the oxygen cohort at DP1 have an indication of oxygen use, with 85% having a positive mention. Within oxygen cohort at DP2, 88.6% have a mention of oxygen, with all having a positive mention. In the non-oxygen cohort at DP1, 84.8% of patients have a mention of oxygen use, compared with 45.7% among DP2 patients. This is a fairly large difference, but again looking at positive mentions, there is a smaller spread. Approximately 45% of DP1 patients have a positive mention of oxygen support, and ~31% of DP2 patients. This illustrates both the potential drawbacks of relying on billing codes to indicate oxygen use and the fact that documentation of positive/current items may be more consistent across Data Partner sites than negative mentions.

Looking at the mentions of oxygen use across note type, we see mentions of oxygen use in over 40 different note types. A key takeaway from this analysis is that it is important to understand the underlying documentation practices in order to select the correct notes. Manual review / annotation is typically not feasible for such a large number of notes, particularly over a large

population, and even automated processing can be a challenge depending on the complexity of the extraction task. This is especially true for more novel concepts that are not part of existing pipelines. Therefore, engagement with potential data partners and clinical experts is key in order to make sure that the relevant content is targeted for any downstream analysis.

For the Cancer cohort, there were some differences in the notes that were selected, particularly the underlying specialty, which led to different levels of information density. This is in contrast to the COVID-19 cohort, where both Data Partners annotated hospital discharge summaries. DP2 notes have more than 5 times that number of annotations within their cancer cohort compared with DP1. Both Data Partner sites chose notes that met the inclusion criteria, though the DP2 note selection was less random, specifically targeting the note types with the most information. For conditions, the overall percentage of the total annotations was roughly the same across Data Partner sites (~25%), though they were found in more DP2 notes compared with DP1 (100% to 75%). There were far more mentions of tests and procedures within the DP2 notes. Smoking status was not well documented across either Data Partner site. All notes mentioned medications, as was expected given the cohort definition, and the total percentages were similar across Data Partner sites. For the gene/protein class, there were more notes in the DP2 cohort that contained a mention compared to DP1 (96.4% compared to 25%), though DP1 had more annotations overall and they represented a higher percentage of the total (18.5% DP1; 1.5% DP2). There were a limited number of overall mentions of stem cell transplants or patients being refractory (non-responsive) to a treatment, which are of interest because of their inclusion in the label indication. It is likely that this information is somewhere within the patient's chart, but the particular note, or visit within their overall care journey may be different by institution. The variation across classes by Data Partner site is an informative finding, illustrating the differences that can occur when selecting note types, particularly as they vary by specialty.

When looking at the presence of medication metadata within a medication annotation in the DP2 cancer cohort, we see that most annotations do not contain any other metadata attributes (~44%). Timing is the most mentioned attribute, but it only appears in ~22% of annotations. One important note about this result is that the annotators skipped any templated text that might have been pulled in from structured portions of the EHR (e.g., medication list). These results were captured in the "Template Start" concept and there is at least one in almost every DP2 cancer note, so it is possible that pulling in the text from the medication list would increase those numbers. However, since those data are already available in a structured field, it would be more straightforward to just use the data in that format than to parse it out via NLP. As a result, one of the main takeaways from this particular sub-analysis is that medication mentions in the free-text portion of the note do not include much additional metadata.

Finally, many of the concept classes used in the annotation exercise are broad (e.g., conditions, treatments and procedures), but these groupings were chosen because existing NLP pipelines are generally robust at extracting data of these types. For more novel concepts where there has been less NLP development (e.g., oxygen support, refractory for stem cell transplant), having a more defined concept definition can make the annotation task more straightforward, while also generating a training corpus for the development of a pipeline that can automatically extract these terms.

Best Practices for Representing Unstructured Data in the SCDM

The general process of transforming unstructured text to records within a CDM is that the 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 to derive or compute records that then are stored in the CDM (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, which are described below:

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?
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

Literature review on NLP performance

EHRs contain a wealth of information that is stored in unstructured text. There have been tremendous advances in NLP tools, to the point where they have become commodity services offered by cloud providers like Microsoft and Amazon. If we wished to select an NLP solution to deploy across dozens of sites to support research or public health surveillance, how would we decide which tool to use? Or if we wanted to give sites an opportunity to select a tool that is "good enough," how would we demonstrate that? In our study, we examined the state of the NLP literature and detailed the various characteristics reported in those articles to provide a guide for those who wish to deploy NLP solutions across multiple sites.

Papers evaluating NLP tools for processing unstructured EHR data (both institutional and open-challenge datasets) were retrieved from five literature databases – PubMed, Association for Computing Machinery Digital Library, Ovid Medline, Scopus, and Web of Science. The years of publication ranged from 2005 to 2021. We focused on Unified Medical Language System (UMLS)-based NLP tools for EHR information retrieval, specifically CTAKES, MetaMap, CLAMP, MTERMS, KMCI and HITEX. We included studies focusing on clinically related resources, having quantitative evaluations, written in English and using English corpus for evaluation. We reviewed and summarized the study objectives, NLP methods used and their validation, software implementations, the performance on the dataset used, and any reported use in practice.

The search identified 479 papers, of which 149 remained after selection on title and abstract, and 40 after full-text review. The most common objective of the papers was to propose and evaluate a new pipeline, with the next most common objective to evaluate existing NLP tools on different domains (i.e., a specific research area like oncology or identification of adverse events). Eleven studies evaluated a single tool, while the rest evaluated more than one. A majority of studies clearly reported whether they addressed a specific domain, and a majority of studies also used gold standard annotations for validating the NLP tools and all three performance measures customary in evaluating NLP results: precision, recall and F-measure.

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.

One of the key questions will be to define the general objective of the use of NLP within Sentinel. The working assumption is that Sentinel would like to support the following 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 may not be necessary if some of the other use cases are ultimately considered out of scope.

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. However, as identified in the NLP capabilities survey, while NLP experience exists within the Sentinel ecosystem, there is little in the way of commonality and consensus about standard approaches. Therefore, it may take time to move the Network to a standard set of tools. For the time being, it is likely more practical for Data Partners to use their own local pipelines to process notes. 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 eventually use the same pipeline, it likely worth conducting this validation to understand the variation across the network.

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). Pre-extracting terms would dramatically shorten the overall development process. If there is concern about the "quality" of pre-processed concepts, one option may be to limit the use of modules within NLP pipelines to those activities that are more robust, 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.

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 direct 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 some fashion.

The OHDSI CDM provides a strong foundation, with tables to store both note text and the outputs of NLP pipelines. The structure of these tables could be adapted by Sentinel, though some of the specific OHDSI requirements, such as the use of 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 use, whether to include them as separate fields or store them 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.

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. Sentinel should provide implementation guidance to ensure the underlying NLP outputs are stored in preferred terminologies.

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. If creating dedicated provenance fields is not feasible, separate tables can be created specifically for NLP concepts as a stopgap, but this may become unsustainable as the number of concepts increases. Finally, a more general table to store "conditions" or other health 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.