Representation of Unstructured Data Across Common Data Models (DI2)

Final Report – Identification of Priority Concepts and Natural Language Processing (NLP) Capabilities Survey

Prepared by: Keith Marsolo, PhD,[1,2] Ruth Reeves, PhD;[3] Li Zhou, MD, PhD;[4] Lesley Curtis, PhD;[1,2] Sarah Palmer, MPH;[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 - Identification of Priority Concepts and Natural Language Processing (NLP) Capabilities Survey

Table of Contents

History of Modifications

Version	Date	Modification	Author
0.1	5/31/2022	Original Draft	Keith Marsolo & Project WG
1.0	01/31/2023	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 paper describes the findings of objectives, 1 and 2a, identify priority concepts that could be extracted via NLP and to survey the NLP tools used across the Sentinel ecosystem, including specific software packages, the context in which they are deployed (e.g., for specific research projects, general use), as well as their extent of use, in terms of notes analyzed and concepts extracted.

Methods

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.

Identification of NLP solutions

The first step in the process was to identify some of the more common NLP solutions utilized within the informatics community. Two systematic reviews were consulted to generate an initial list[1,2], 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.

The NLP solutions were divided into 4 categories: frameworks (generalized platforms that can be used to generate user or site-specific implementations), tools (stand-alone software packages), toolkits (suites of tools where NLP might be one component in a broader set of capabilities), and commercial services (software-as-a-service-type solutions [typically cloud-based] that are offered by a commercial vendor). Tools and toolkits could be open-source or commercial products.

---

[1] Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019 Apr 27;7(2):e12239. doi: 10.2196/12239. PMID: 31066697; PMCID: PMC6528438.

[2] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov 21. PMID: 29162496; PMCID: PMC5771858.

For each solution in the list, we attempted to gather as much information as possible from project websites or other publications. Topics for consideration included:

- Data elements / data domains / metadata extracted
- Terminologies used (by element/domain if applicable)
- General approach
- License / cost
- Source webpage
- Reference publication

Catalog of standard concepts

We sought to generate a set of concepts that could be extracted using the identified NLP solutions. The goal was to generate a “good enough” set of concepts, stopping when we reached saturation. 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). We also limited ourselves to the basic functionality provided by each solution, not every customization that might have been introduced as part of a research project. We also looked at the prior Sentinel projects that leveraged NLP in order to identify additional concepts that were extracted as part of that work.

Concept prioritization

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.

Survey development

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.

Identification of respondents

The survey was distributed in a manner that was compliant with the Paperwork Reduction Act (less than 9 respondents). All organizations that serve as subcontractors to Harvard Pilgrim Healthcare Institute (HPHCI) (e.g., Sentinel Data Partners, Sentinel Innovation Center lead organizations) could be sent the survey and their collective submission would count as a single response. The remaining respondents were selected from the list of participants of the October 28, 2021, webinar that was held to introduce the project to the institutions and organizations that expressed interest in serving as broader Innovation Center partners.

The survey was distributed to Sentinel Data Partners via PopMedNet as a Word document. The remaining partners were contacted via e-mail with a link to an online Qualtrics survey. Sentinel Data Partners also had the opportunity to respond to the Qualtrics survey.

Results

Identification of NLP solutions

The NLP solutions that were identified via the systemic review and through feedback from workgroup members are shown in Table 1 below. We obtained general information for almost every solution in the list, though the specific elements or concepts that could be extracted were most readily available for the solutions in the Tools column. Therefore, when generating the initial list of concepts, we focused our efforts there. All of the information collected on the NLP solutions can be found in **Appendix A – General information on NLP solutions**.

Table 1: NLP solutions identified in the systematic reviews and through feedback from workgroup members. Solutions denoted with a '' indicate a commercial product.*

Frameworks	Tools	Toolkits	Commercial Services
GATE	cTAKES	MALLET	Microsoft*
UIMA	MetaMap	OpenNLP	Amazon*
Protégé	MedLEE/Lumanent Insights*	NLTK	Nuance*
	KnowledgeMap Concept Indexer (KMCI)	SPLAT*	Wolters Kluwer*
	HITEx	RapidMiner*	Linguamatics*
	MedEx		
	MedTagger		
	ARC		
	Medtex		
	CLAMP*		
	MedXN		
	PredMED*		
	SAS Text Miner*		
	MediClass		
	MTERMS		
	BioMedICUS		
	Leo		
	Ether		

Catalog of standard concepts

Table 2 provides a list of concepts that can be extracted via each solution (capabilities based on available documentation). A few general notes about these results:

- Some of these concepts overlap. For instance, the solutions that can extract “all UMLS codes” (all codes from the Unified Medical Language System) should be able to extract diagnoses, procedures, medications, signs/symptoms, etc., as all of terminologies that underlie those domains are part of the UMLS.
- Certain solutions are designed to work with specific document types (e.g., cancer pathology reports), and would not be suitable for general concept extraction.

- Most solutions have the ability to assign relationships between terms (e.g., timing, body location), but the specific relationships are not always available in the high-level descriptions.
- What some tools consider to be an extractable concept may be a relationship or metadata in another (e.g., the type of diagnosis [primary/secondary, family history]).

Current Sentinel projects were also evaluated to identify additional concepts. These projects are listed below:

- Validation of Anaphylaxis Using Machine Learning
- Validation of Acute Pancreatitis Using Machine Learning and Multi-Site Adaptation for Anaphylaxis
- Sentinel Scalable NLP (COVID-19 focused)
- Improving probabilistic phenotyping of incident outcomes through enhanced ascertainment with natural language processing
- Developing NLP-assisted chart abstraction tool
- Scalable automated NLP-assisted feature extraction
- Augmenting EHR Death Ascertainment in Sentinel

Reviewing these projects yielded the following additional concepts for inclusion: anaphylaxis, acute pancreatitis, COVID-19 (positive or negative) and suicidality.

To aid in the prioritization process, the concepts were reorganized using a categorization described by Microsoft for their Azure text analytics for health solution[3]. This layout is shown in Table 3.

[3] https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/text-analytics-for-health/concepts/health-entity-categories

Table 2: Concepts that can be extracted from each NLP solution.

Concept/Domain	CLAMP	CTAKES	META-MAP	MEDLEE	KMCI	HITEX	MEDEX	MEDXN	PRED-MED	MEDI-CLASS	MTERMS	BIO-MEDICUS	ETHER
All UMLS concepts (includes many of the concepts listed below)	x	x	x		x	x				x	x	x	
Disease	x	x									x		
Medication	x	x					x	x	x		x		
Procedure	x	x											
Disease attributes (body location, severity, uncertainty)	x												
Lab attributes (value)	x												
Medication attributes (dose, form, route, frequency, duration, necessity)	x	x					x	x	x		x		
Bleeding symptoms	x												
Symptoms of colorectal cancer	x												
Smoking status	x	x				x					x		
COVID-19 signs and symptoms	x										x		
Cancer Pathology (Site, Procedure, Histology)	x										x		
Anatomical site		x											
Signs / Symptoms		x									x		
Adverse drug reactions (e.g., allergy, side effects)		x									x		
Medical entities (overlaps with many concepts in the UMLS)				x									
Discharge medications						x					x		
Family history						x					x		x
Primary diagnosis						x							x
Vaccine													x
Secondary diagnosis													x
Medical history													x
Temporal information											x		x

Table 3: Reorganized NLP concepts to be prioritized by FDA.

Domain	Subdomain	Other metadata (applies to corresponding domain/ subdomain)	Description / example
Domains related to patients / patient characteristics			
Allergens			
Anatomy			Body system/sites, anatomic locations/regions
Cancer Pathology	Site		
	Procedure		
	Histology		
Condition	Diagnoses		
	Signs/Symptoms		
		Time period	Acute, chronic
		Nature	Sharp, burning
		Severity	Mild, uncontrolled
		Extensivity	Local, diffuse
		Scale	Stage I, II, etc.
	Condition-specific signs/symptoms		Some tools have developed specific groupings of signs/symptoms
		COVID-19	
		Bleeding	
		Colorectal cancer	
	Type		Some tools can determine the type of mention for a condition
		Family history	
		Primary diagnosis	
		Secondary diagnosis	
		Medical history	
Demographics	Age		
	Gender		
Family Relation			Mentions of family relatives of subject (e.g., father, sister)
Genomics (gene or protein)	Variant		
	Mutation		
	Expression level		
Smoking status			Current smoker, former smoker, etc.

Domains related to care delivery			
Admission-Discharge-Transfer (ADT)-type event			Registration, discharge, transfer
Care setting			Hospital, unit, ER
Diagnostic procedures / tests			
Healthcare profession			Specialty, service

Medication	Name		
	Class		
		Dose	
		Form	
		Route	
Treatment / procedures			
Metadata that apply to multiple domains			
General attributes		Course	Change over time
		Date	
		Direction	
		Frequency	
		Time	Beginning or length
		Measurement unit	
		Measurement value	
		Relational operator	greater than, less than
		Negation	
		Uncertainty	
Concepts from existing Sentinel projects			
Anaphylaxis			
Acute pancreatitis			
COVID-19 (+/-)			
Suicidality			

Concept prioritization

The prioritization of the NLP concepts by FDA is shown in Table 4, along with any comments provided. 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. Several new concepts were added through this process and are listed at the bottom of the table. These additional concepts include timing and duration of a medication, indication, physical findings, oxygen support and death date and cause. *Please note that concept of death date and cause was incorrectly reported as a Low priority in the presentation of this project during the Sentinel public meeting (April 29, 2022). It should have been listed as "High."*

Table 4: Priority rankings assigned to NLP concepts by FDA.

Domain	Subdomain	Other metadata	Priority	FDA Comments
Domains related to patients / patient characteristics				
Allergens			**Low**	Rarely used in FDA pharmacoepi studies
Anatomy			**Medium**	Often captured in claims (diagnosis) codes
Cancer Pathology				
	Site		**High**	There have been a number of ARIA insufficiency determinations due to lack of detailed data on cancer (e.g., staging)
	Procedure		**High**	
	Histology		**High**	
Condition				
	Diagnoses		**Medium**	Often captured in claims via codes
	Signs / Symptoms		**High**	Less likely to be captured in claims but would be useful for different aspects of studies (e.g., population of interest, outcome of interest)
		Time period	**High**	
		Nature	**Medium**	
		Severity	**High**	Severity of disease is an important covariate often missing in our claims-based data studies
		Extensivity	**Medium**	
		Scale	**High**	
	Condition-specific signs / symptoms	COVID-19	**High/Med**	Dependent on whether the condition is of interest in a particular study, but this is likely valuable since claims do not typically capture signs and symptoms well
		Bleeding	**High/Med**	
		Colorectal cancer	**High/Med**	
	Type	Family history	**Medium**	May be useful for some studies but not all
		Primary Dx	**Medium**	Primary/secondary diagnosis positions can be arbitrary; this info is typically captured in claims
		Secondary Dx	**Medium**	
		Medical history	**High**	Longitudinality often missing in EHR, so medical history info is important to capture
Demographics				
	Age		**Low**	Important variables, but not important to capture in EHR if already captured in claims
	Gender		**Low**	
Family Relation			**Low**	
Genomics (gene or protein)				
	Variant		**High**	High priority for drugs where a genetic test is relevant, but many drugs/conditions do not have an applicable generic test
	Mutation		**High**	
	Expression level		**High**	
Smoking status			**High**	Key covariate in many FDA studies; under captured in claims

Domains related to care delivery		
Admission-Discharge-Transfer (ADT)-type event	**Medium**	Many transitions (admission, discharge) are captured in claims; smaller care transitions/settings (e.g., ICU) would be important to capture
Care setting	**Medium**	
Diagnostic procedures / tests	**High**	Imaging and laboratory? Claims data have fact of, not results
Healthcare profession	**Medium**	

Medication				Depends on setting - inpatient high priority, OTC medications (not covered in claims), free samples of new medications (not covered in claims)
	Name		**High/Med**	
	Class		**Low**	Can be determined from drug name
		Dose	**High/Med**	
		Form	**High/Med**	
		Route	**High/Med**	
Treatment / procedures			**High/Med**	
Metadata that apply to multiple domains				
General Attributes				Priority largely depends on the domain being described by the metadata – *Examples where important*
		Course		Medications administration; when titration schedule matters
		Date		Medication administrations; exact timing for acute adverse events
		Direction		*Unclear utility; likely based on a judgement*
		Frequency		Medication administrations
		Time		Medication administrations; oxygen support; medical history
		Measure-ment unit		Lab result units - necessary for interpretation
		Measure-ment value		Lab result values
		Relational operator		Lab result values, if exact value is not available
		Negation		Rule out diagnoses
		Uncertainty		*Unclear utility; likely based on a judgement*
Concepts from existing Sentinel projects				
Anaphylaxis			**Medium**	Outcomes of interest for drug safety studies, but ongoing work to identify using claims algorithms
Acute pancreatitis			**Medium**	
COVID-19 (+/-)			**High**	COVID-19 diagnosis can largely be captured in claims, but other aspects of COVID-19 are of interest: PASC, antibody testing, vaccine status, etc.
Suicidality			**High**	
Additional items to consider				
Timing and duration of medications			**High**	May be captured above, but this is particularly important for inpatient medications which are difficult to identify in claims
Findings collected during physical exam (e.g., height, weight)			**High**	Key covariates in many FDA studies; under captured in claims
Indication for a drug			**High**	This may be captured by diagnoses and/or procedures, but EHRs should note specifically what a drug is being used for
Oxygen support			**High**	Largely relevant for COVID-19 studies
Death (date) and cause of death			**High**	Capture of death information varies widely by Sentinel DP
Hospice care			**Medium**	Indicates imminent death; often impacts health service utilization

Survey results

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 survey questions can be found in **Appendix B – NLP Survey**.

Respondents reported a wide variety of tools used to extract information from text. These include SAS, locally developed Python scripts or other in-house tools, Health Discovery (from Averbis), n-gram models, cTAKES and CLAMP. In terms of notes processed, some respondents report being able to extract information from any clinical notes, typically from the point when their EHR / source system(s) went live, while others are limited to certain specialty types (e.g., pathology or radiology reports, laboratory tests).

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), as indicated in Figure 1. But most concepts are only extracted by a small number of partners.

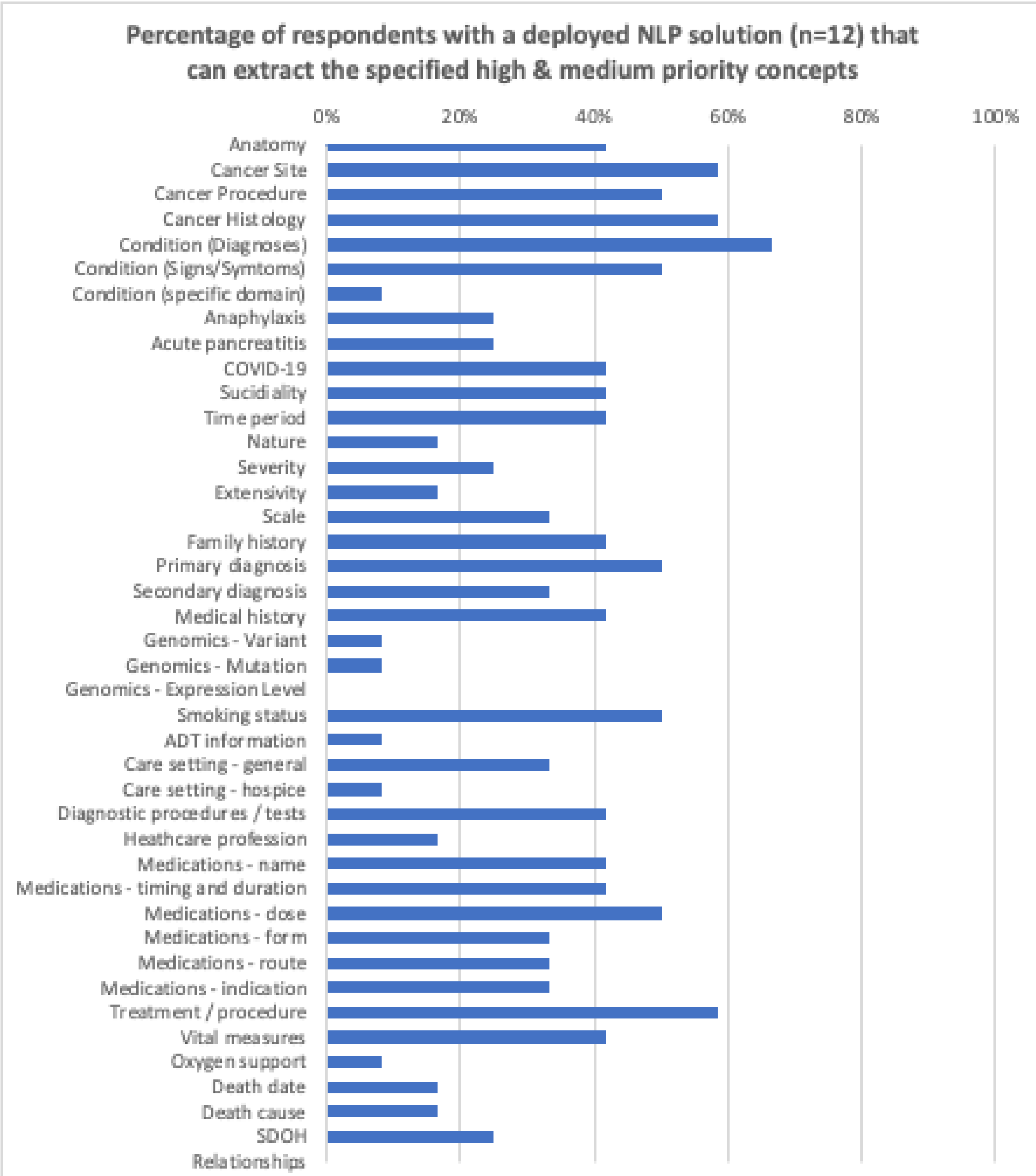

Figure 1: NLP concepts that are extracted by survey respondents.

Discussion

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. To that end, 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, such as the indication behind medication prescription/order or the timing and duration of a medication exposure, 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.

The uptake of NLP solutions across partners varies 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 Sentinel Network may have an opportunity to ask Data Partners to adopt a specific NLP solution for processing their notes, though there may be higher cost than if Data Partners were just allowed to use whatever solution(s) they have locally. The potential benefit of a standard solution is that there will be more consistency in the quality of the results, since the performance of these tools will often vary based on the concepts being extracted or the type of notes that are processed. As noted above, FDA was not asked to create a ranked list of concepts. One of the key decisions will be whether Sentinel asks Data Partners to only extract targeted concepts in service of a specific type of analysis, or whether they should extract a broad set of concepts for potential future use. This calculus may depend on the NLP pipeline(s) and notes available to the partners, along with timeline and budget. If there is a fixed cost per partner to do a single run through a set of notes, a decision might be made to extract as much as possible to keep the overall costs down. Conversely, if there are partners with access specialty notes (e.g., cancer pathology), a more tailored approach may be suitable to focus on the information that is unique to those document types.

Going forward, it will also be important for the Sentinel Network / FDA to decide if they want to ask Data Partners to extract concepts from unstructured text that are likely to be present in structured fields within the EHR – such as findings from physical exam, social history items (e.g., alcohol use or lifestyle behaviors), social determinants of health, or medications administered within an inpatient setting. When comparing structured data with the (nominally) same concepts derived from unstructured data, it is likely that there will be “gaps” in both data streams – an inpatient note may contain information on medications administered during a hospital stay at another facility, for instance, while flowsheets may include vital signs that were not pulled into a progress note. One option may be to simply limit the concepts extracted from unstructured text to those are unlikely to be present in structured fields (e.g., medical history that was recorded at another healthcare facility), as it will be challenging to limit any extraction to those concepts that are only present in the unstructured text. Another option would be for Data Partners to pull in data from all possible streams, and then execute quality checks that attempt to ascertain completeness (or assess information gain) based on provenance, but this will result in a non-trivial amount of work, for both the Operations Center and Data Partners. Therefore, it will be important to clearly define any use case for incorporating information extracted from unstructured text.

Appendix A – General information on NLP solutions

Table 5: General information on NLP solutions evaluated as part of this objective.

Solution	License / Cost	Source Webpage	References	Description
CLAMP	CLAMP is free for academic users for their individual research projects.	https://clamp.uth.edu/	https://clamp.uth.edu/publications.php	CLAMP (Clinical Language Annotation, Modeling, and Processing Toolkit) is a comprehensive clinical Natural Language Processing software that enables recognition and automatic encoding of clinical information in narrative patient reports. In addition to running clinical concept extraction as well as annotation pipelines, the individual components of the system can also be used as independent modules.
cTAKES	Open source/free	https://ctakes.apache.org/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/	cTAKES is a natural language processing system for extraction of information from electronic medical record clinical free text. Originally developed at the Mayo Clinic, it has expanded to being used by various institutions internationally. The clinical Text Analysis and Knowledge Extraction System (cTAKES) supported by Apache is an open-source NLP tool. The cTAKES annotations are the foundation for methods and modules for higher-level semantic processing of clinical free text.
MetaMap	Free	https://metamap.nlm.nih.gov/	https://academic.oup.com/jamia/article/17/3/229/738417	MetaMap is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. Besides being applied for both IR and data-mining applications, MetaMap is one of the foundations of NLM's Medical Text Indexer (MTI) which is being used for both semiautomatic and fully automatic indexing of biomedical literature at NLM.
MedLEE / Lumanent Insights*	Commercial, unavailable	https://healthfidelity.com/lumanent/	https://cdc.confex.com/recording/cdc/phin2008/ppt/free/4db77adf5df9fff0d3ca	MedLEE (Medical Language Extraction and Encoding) extracts, structures, and encodes clinical information in narrative patient reports. With support from CDC, development began by Columbia University Medical Center in 1991 and has been used there since 1995. In 2012, MedLEE granted an exclusive license

			f5cafe28f496/paper17028_5.ppt	to Health Fidelity (healthfidelity.com). Health Fidelity's NLP engine Lumanent Insights was created based on the NLP engine MedLEE developed at Columbia University by Carol Friedman, PhD.
KnowledgeMap Concept Indexer (KMCI)	TBD	https://www.vumc.org/cpm/cpm-blog/kmci-knowledgemap-concept-indexer	N/A	The KnowledgeMap Concept Indexer (KMCI), housed at Vanderbilt University Medical Center, is the underlying natural language processing engine used in the KnowledgeMap and Learning Portfolio website, and has been used for many clinical and genomic research studies. It identifies biomedical concepts, mapped to Unified Medical Language System concepts, from natural language documents and clinical notes. KMCI has performed favorably in comparison to MetaMap and has been validated in a variety of clinical and education contexts (see publications). Later additions to KMCI include the ability to detect negated terms (e.g., "no chest pain) via a Perl implementation of NegEx.
HITEx	Open source/free	https://www.i2b2.org/software/projects/hitex/hitex_manual.html	N/A	HITEx (Health Information Text Extraction) is an open-source NLP software application developed by a group of researchers at the Brigham and Women's Hospital and Harvard Medical School. HITEx is built on top of Gate framework and uses Gate as a platform. HITEx consists of the collection of Gate plug-ins that were developed to solve problems in medical domain, such as princial diagnoses extraction, discharge medications extraction, smoking status extraction and others.
MedEx	Open source/free	Not available	https://www.sciencedirect.com/science/article/pii/S1532046417302563	NLP tool used to recognize drug names, dose, route, and frequency from free-text clinical records
MedTagger	Open source/free	https://github.com/OHNLP/MedTagger	http://ohnlp.org/index.php/MedTagger_Project_Page	MedTagger contains a suite of programs that the Mayo Clinic NLP program has developed in 2013. It includes three major components: MedTagger for indexing based on dictionaries, MedTaggerIE for information extraction based on patterns, and MedTaggerML for machine learning-based named entity recognition.
ARC	Open source/free	Not available	https://dl.acm.org/doi/10.1145/1882992.1883065	Open-source natural language processing (NLP) frameworks have made it easier for NLP developers and researchers to develop more reusable and modular components and to capitalize on the work of others. With the Automated Retrieval Console (ARC) we attempt to build upon this foundation by streamlining

				the many processes surrounding the development, evaluation, and deployment of natural language processing technologies. Toward this end, ARC offers graphical user interfaces to facilitate corpus import, reference set creation, annotation, and inter-annotator agreement calculation. To speed task-specific information extraction development, ARC combines NLP-generated features from UIMA pipelines with machine learning classifiers and calculates performance statistics against a reference set.
Medtex	Open source/free	https://aehrc.com/medical-text-processing/		Medtex works by learning what statements to look for, and uses SNOMED CT, the internationally defined set of clinical terms, to unify and reason with the language across information sources. It incorporates domain knowledge to bridge the gap between natural language and the use of clinical terminology semantics for automatic medical text inference and reasoning.
MedXN	Open source/free	https://github.com/OHNLP/MedXN	https://pubmed.ncbi.nlm.nih.gov/24637954/	Medication Extraction and Normalization (MedXN, pronounced [med-eks-en]) is an Apache UIMA-based medication information extraction system that focuses on assigning the most specific RxNorm RxCUI to medication description. MedXN finds medication and its complete attributes and normalize them to the most specific RxNorm RxCUI using flexible matching, abbreviation expansion, inference, etc. MedXN uses externalized resources (ie, medication dictionary, attribute definitions, and regular expression attribute patterns) to allow a simple customization process for the needs of end users.
PredMED*	Commercial, unavailable		Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan; 77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov	NLP application developed by IBM to extract full prescriptions from narrative clinical notes

			21. PMID: 29162496; PMCID: PMC5771858.	
SAS Text Miner*	Commercial, unavailable		Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan; 77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov 21. PMID: 29162496; PMCID: PMC5771858.	A plug-in for SAS Enterprise Miner environment provides tools that enable you to extract information from a collection of text documents and uncover the themes and concepts that are concealed in them.
MediClass	Open source/free		https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205600/	The MediClass system was built from open-source components. It uses three distinct informatics technologies: (1) HL7's CDA for representing the clinical encounter including both structured (coded) and unstructured (free-text) data elements;10 (2) natural language processing (NLP) techniques for parsing and assigning structured semantic representations to text segments within the CDA; and (3) knowledge-based systems for processing semantic representations addressing specific subdomains of medicine and clinical care and for defining logical classifications over the semantic contents of a clinical encounter.
MTERMS	Available / free for academic users	https://mterms.bwh.harvard.edu/mterms/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243163/	Medical Text Extraction, Reasoning, and Mapping System (MTERMS) is a natural language processing (NLP) system for biomedical text. Originally designed to extract medication information from clinical notes to facilitate real-time medication reconciliation, MTERMS has been extended to identify diverse clinical information and support a variety of clinical informatics applications and projects, some of which have been integrated with the Epic EHR system in real-time patient care via research

				studies/clinical trials (e.g., detection of abnormal cancer screening results, allergy reconciliation).
BioMedICUS	Open source/free	https://github.com/nlpie/biomedicus3	https://nlpie.github.io/biomedicus/	The BioMedical Information Collection and Understanding System (BioMedICUS) is a system for large-scale text analysis and processing of biomedical and clinical reports. The system is being developed by the Natural Language Processing and Information Extraction Program at the University of Minnesota Institute for Health Informatics. This is a collaborative project that aims to serve biomedical and clinical researchers, allowing for customization with different texts.
Leo	Open source/free	https://department-of-veterans-affairs.github.io/Leo/index.html	https://github.com/department-of-veterans-affairs/Leo	The Department of Veterans Affairs' VINCI-developed Natural Language Processing (NLP) infrastructure is a set of services and libraries that facilitate the rapid creation and deployment of Apache UIMA-AS annotators focused on NLP. Leo, named for the Spanish word meaning "I read", was first built to support scalable deployment of NLP pipelines (VINCI now has more than 2 billion clinical text notes available). It extends the open-source Apache Unstructured Information Management Architecture (UIMA).
ETHER	Unavailable	Unavailable	https://www.sciencedirect.com/science/article/pii/S1532046416300776?via%3Dihub	ETHER was developed within CBER for VAERS narratives and has more recently been applied to extracting events from product labels. There is an ongoing project within OSE that uses ETHER to extract some information from FAERS narratives (e.g., diagnoses, medical history, timing) and display it in a visualization platform (INFOViP) to assist with case evaluation.
MALLET	Open source/free	Undetermined	http://mallet.cs.umass.edu/index.php	MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. Connection to University of Massachusetts
OpenNLP	Open source/free	https://opennlp.apache.org/docs/	https://opennlp.apache.org/	OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, language detection and coreference resolution.
NLTK	Open source/free	https://www.nltk.org/		NLTK is a leading platform for building Python programs to work with human language data.
SPLAT*	Commercial, unavailable	https://www.microsoft.com/en-us/research/project/msr-splat/		Developed by Microsoft, Statistical Parsing and Linguistic Analysis Toolkit (SPLAT) is a linguistic analysis toolkit. Its main goal is to allow easy access to the linguistic analysis tools produced by the Natural Language Processing group at Microsoft

				Research. The tools include both traditional linguistic analysis tools such as part-of-speech taggers and parsers, and more recent developments, such as sentiment analysis (identifying whether a particular of text has positive or negative sentiment towards its focus)
RapidMiner*	Commercial, unavailable	https://rapidminer.com/why-rapidminer/		RapidMiner brings artificial intelligence to the enterprise through an open and extensible data science platform. Built for analytics teams, RapidMiner unifies the entire data science lifecycle from data prep to machine learning to predictive model deployment. More than 700,000 analytics professionals use RapidMiner products to drive revenue, reduce costs, and avoid risks.
Microsoft	Commercial, unavailable	https://appsource.microsoft.com/en-us/product/web-apps/sytrue.nlpos?tab=Overview		SyTrue Natural Language Processing Operating System (NLP OS™) is a Microsoft web app. NLP OS™ cascades a single medical record into multi-purpose content in less than a second. Want to know which ICD-10, CPT, LOINC, SNOMED, HCC codes are represented within the same medical record? That's easy. Need to identify allergies, medications, hedging terms or critical findings? Yep… What about personal history, medical necessity, pain or inference? SyTrue's clinical analyzers dive deep into the content contained within medical records shedding valuable insights on the patients' encounter. One medical record – multiple purposes.
Amazon	Commercial, pricing based on amount of text processed on a monthly basis. See https://aws.amazon.com/comprehend/medical/	https://aws.amazon.com/comprehend/medical/	https://pubmed.ncbi.nlm.nih.gov/34042745/	Amazon Comprehend Medical is a HIPAA-eligible natural language processing (NLP) service that uses machine learning to extract health data from medical text–no machine learning experience is required.
Nuance	Commercial, unavailable	https://www.nuance.com/healthcare.html		Nuance AI solutions transform the way we work, connect, and interact with each other to advance the effectiveness of your organization and further your positive impact on the world. Nuance is being acquired by Microsoft.
Wolters Kluwer	Commercial, unavailable	https://www.wolterskluwer.com/e		The cNLP (clinical NLP) solution, part of the Health Language platform, supports extraction of patient data found in

		n/solutions/health-language/resource-center/solution-information		free-text physician notes and patients' electronic health records to improve the quality of data across payer and provider organizations.
Linguamatics	Commercial, unavailable	https://www.linguamatics.com/products/linguamatics-nlp-platform	https://www.linguamatics.com/solutions/real-world-data	Owned by IQVIA, The Linguamatics Natural Language Processing (NLP) platform offers an exceptional combination of flexibility, scalability, and data transformation power to effectively address the challenges of analyzing unstructured data, and support organizational goals to: Boost innovation, Speed R&D and clinical processes, optimize quality and improve efficiency, reduce risk and costs, Improve patient outcomes.

Appendix B – NLP Survey

This survey is intended to assess the use of natural language processing (NLP) solutions by Sentinel-affiliated partners through a projected funded by the Sentinel Innovation Center. We are interested in gathering more information about the NLP solution(s) that may be deployed within each organization to process their EHR data, as well as the scale of information that is extracted from unstructured text. These findings will be used to inform the planning of the Sentinel Network as it looks to incorporate information derived from unstructured text into the Sentinel Common Data Model and distributed analyses. While you can provide contact information as part of your submission, all responses will be reported in aggregate.

1. Does your organization have an NLP solution deployed?
   - Yes
   - No
2. [If yes to Question #1] In what context is that solution used?
   - Operational (i.e., information is extracted to support clinical care, operations or improvement activities)
   - Research
   - Both
3. [If yes to Question #1] What solution(s) are deployed?
   - Free text response
   -
4. [If no to Question #1] Do you plan to deploy a solution in the next 1-2 years?
   - [If yes to Question #4] What solution(s) do you plan to deploy?
     - Free text response
   - [If no to Question #4] Thank you for your input. You do not need to complete the remainder of the survey.

[Complete remaining questions if you answered "yes" to Question #1 or #4]

5. What notes/narrative do you process with your NLP solution? Please indicate timeframe (e.g., all available notes, notes after X date), care setting (e.g., ambulatory, inpatient, surgery), patient population or content from ancillary system(s) (e.g., pathology notes), as appropriate.
   - Free text response
   -

6. The following concepts have been identified as medium / high priority by the FDA. Which, if any, do you routinely extract from narrative text? Please mark all that apply. Note that some of these concepts may be more reliably retrieved from structured fields. Even so, we are interested in whether your organization extracts them from narrative text.

Domain / Concept		**Description**	**Available?**
Anatomy			
	Anatomy details	Body system, anatomic locations/regions, body sites, etc.	
	Free text - other items to note about this concept/domain?		
Cancer Pathology			
	Site		
	Procedure		
	Histology		
	Free text - other items to note about this concept/domain?		
Condition			
	Diagnoses		
	Signs/Symptoms		
Condition-specific signs/symptoms	Free text – do you extract any signs/symptoms that are targeted to a specific domain (e.g., COVID-19-related)		
Conditions from Sentinel projects	Note: these conditions are highlighted due to their use in existing Sentinel projects.		
	Anaphylaxis		
	Acute pancreatitis		
	COVID-19 (+/-)		
	Suicidality		
	Free text – other items to note about this concept/domain?		
Condition Metadata			
	Time period	Acute, chronic	
	Nature	Sharp, burning	
	Severity	Mild, uncontrolled	
	Extensivity	Local, diffuse	

	Scale	Stage I, II, etc.	
	Free text - other items to note about this concept/domain?		
Condition Type			
	Family history		
	Primary diagnosis		
	Secondary diagnosis		
	Medical history	Co-morbidity	
	Free text - other items to note about this concept/domain?		
Genomics (gene or protein)			
	Variant		
	Mutation		
	Expression level		
	Free text - other items to note about this concept/domain?		
Smoking status			
	Status	Current smoker, former smoker, etc.	
	Free text - other items to note about this concept/domain?		
Admission-Discharge-Transfer (ADT)-type event			
	ADT information	Registration, discharge, transfer	
	Free text - other items to note about this concept/domain?		
Care setting			
	General setting	Inpatient, ER, ambulatory	
	Hospice care		

	Free text - other items to note about this concept/domain?		
Diagnostic procedures / tests			
	Diagnostic procedure/test info		
	Free text - other items to note about this concept/domain?		
Healthcare profession			
	Profession information	Specialty, service	
	Free text - other items to note about this concept/domain?		
Medications			
	Name		
	Timing and duration		
	Dose		
	Form		
	Route		
	Indication	Reason ordered	
	Free text - other items to note about this concept/domain?		
Treatment / procedures			
	Treatment/procedure info		
	Free text - other items to note about this concept/domain?		
Physical exam findings			
	Vital measurements	Height, weight, blood pressure, etc.	
	Free text - other items to note about this concept/domain?		

Oxygen support			
	Info on oxygen support	Use of supplemental oxygen	
	Free text - other items to note about this concept/domain?		
Death			
	Date		
	Cause of death		
	Free text - other items to note about this concept/domain?		
Social Determinants of Health (SDOH)			
	SDOH	Food instability, transportation issues, financial strain	
	Free text – other items to note about this concept/domain?		
Relationships			
	If you assign relationships between concepts (e.g., tumor location, temporal relationships between records, etc.), please describe.		
Other			
	Free text – any additional information you would like to provide?		