Integration of Scalable Natural Language Processing to the Atlas Cohort Building Workflow

Pavan Parimi[1], Selvin Soby[1], Pavel Goriacko[2], Chandra has Nelapatla[1], Boudewijn Aasman[1], Manuel Wahle[1], Reetam Nath[1], Parsa Mirhaji[2]
[1] Montefiore Medicine, [2] Albert Einstein College of Medicine at Montefiore

Background

Einstein Atlas facilitates the design and execution of computable phenotyping and cohort-based analysis on standardized, patient-level, observational data. There are several methodologies utilized by the community to build cohorts from data derived from clinical text on OHDSI's OMOP datasets.[1,2, 3] This project describes a mechanism for researchers to build advanced cohorts using discrete data from the OMOP CDM as well as concepts derived from clinical text using an NLP engine built on cTAKES and Elastic Search.

Clinical Text Analysis and Knowledge Extraction System (cTAKES) is an open-source natural language processing system for information extraction from EHR's clinical free-text. The system is based on Unstructured Information Management Architecture (UMIA) framework and the OpenNLP toolkit. The components are specifically trained for the clinical domain, based on Java and can be used to identify and extract entities specific entities, relationships between those entities, part-of-speech tagging, and dependency parsing.[2]

Methods

We built our NLP engine using a stepwise approach. To prepare the data for processing, the first step involved obtaining the source database containing the clinical text. The data is then preprocessed by understanding the various sections within the clinical structure, such as impression, plan, and labs. Next, the cTAKES software is configured, including the integration of the UMLS and OMOP dictionaries, relation-extraction, negation and context extraction, and any other required components. The output from cTAKES and other analytic engines, along with the related metadata, is serialized as JSON-LD (JSON for Linked Data) and integrated into an Elastic Search cluster for storage.

Additional analytic engines are employed to perform tasks like named entity recognition, part-of-speech tagging, and uniform text deidentification process that uses a large language model to tag protected information in clinical text. The text body, metadata, and annotations generated by the analytic pipelines are then placed into the Elastic Search database. Finally, necessary indexing is implemented to facilitate efficient retrieval of the processed data.

A specific user experience and interaction model was developed to expose cohorts generated or shared via Atlas to the NLP engine for just-in-time querying. A versatile cohort-based query engine was developed to enable submitting complex pattern search queries or terminology-based queries to JSON-LD and free text components. This engine allows query notes within an Atlas generated cohort and to integrate the results to research baskets and/or Atlas cohorts that could be further characterized, shared, or analytically used inside the Atlas framework.

All NLP queries are stored and tracked for IRB auditing and/or sharing with other users. A specialized visualizer, browser was developed to assist with reviewing and navigating clinical documents as well as extracted metadata and query results for validation.

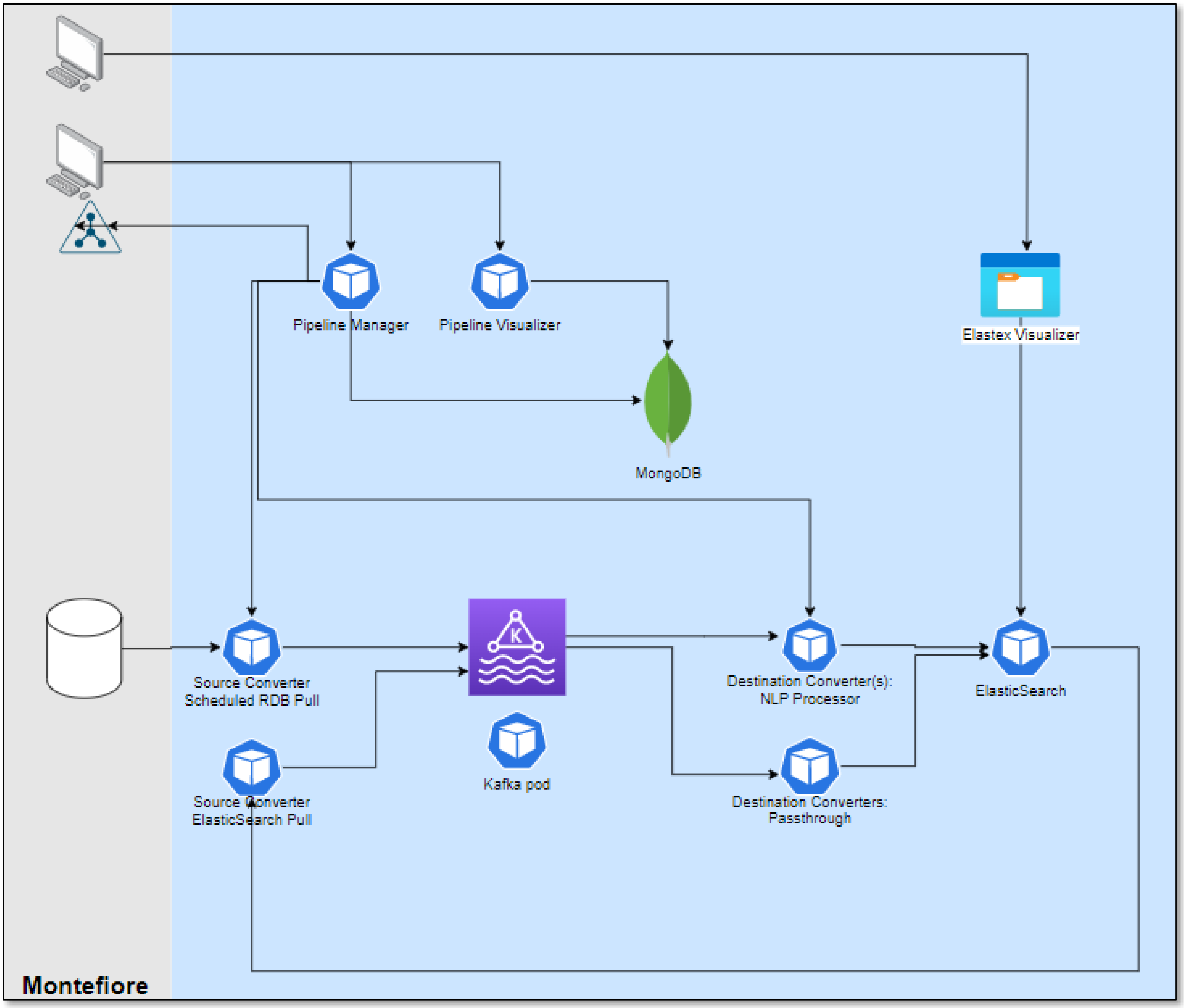
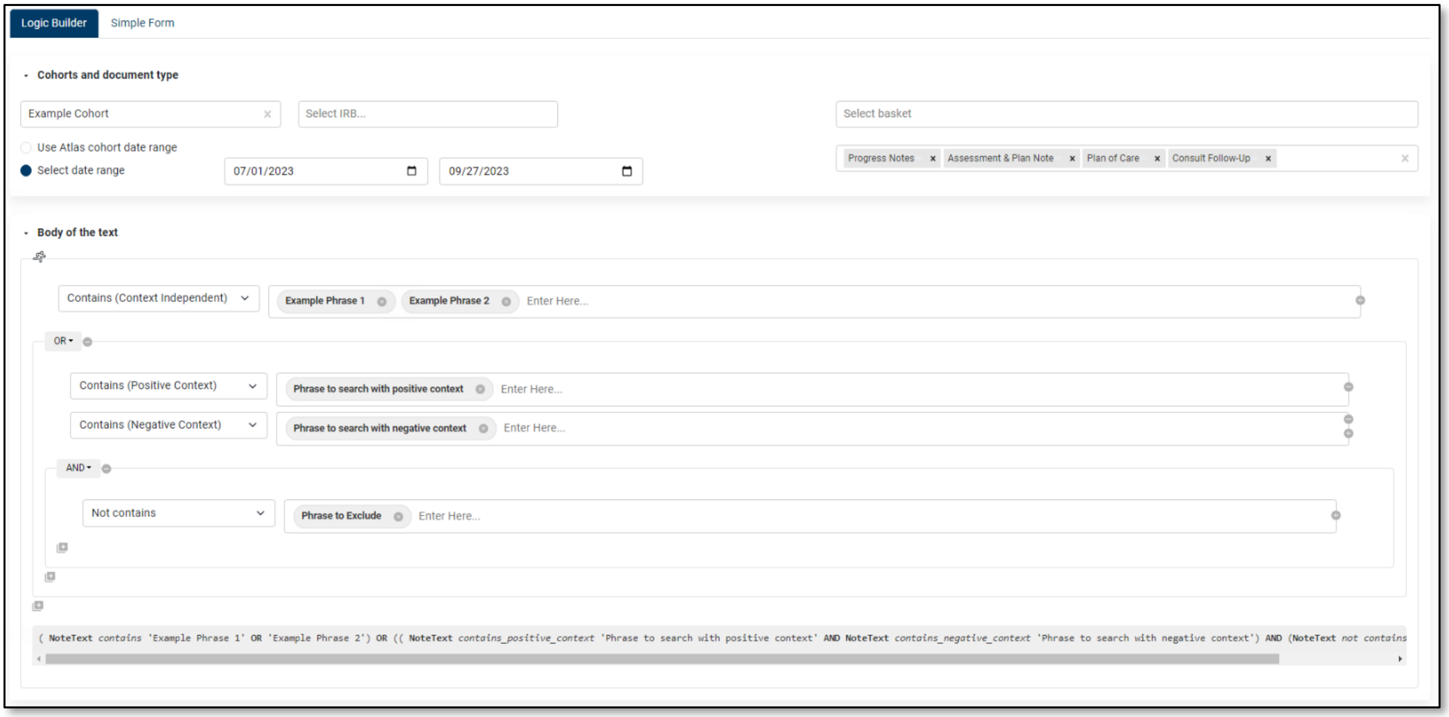

Pipeline Manager
Pipeline Visualizer
Elastex Visualizer
MongoDB
Source Converter Scheduled RDB Pull
Kafka pod
Destination Converter(s): NLP Processor
ElasticSearch
Source Converter ElasticSearch Pull
Destination Converters: Passthrough
Montefiore

Figure 1: Architecture of NLP Engine



Cohorts and document type

Example Cohort ×

Select IRB...

Select basket

○ Use Atlas cohort date range

● Select date range 07/01/2023 09/27/2023

Progress Notes × Assessment & Plan Note × Plan of Care × Consult Follow-Up ×

Body of the text

Contains (Context Independent): Example Phrase 1, Example Phrase 2, Enter Here...

OR

Contains (Positive Context): Phrase to search with positive context, Enter Here...

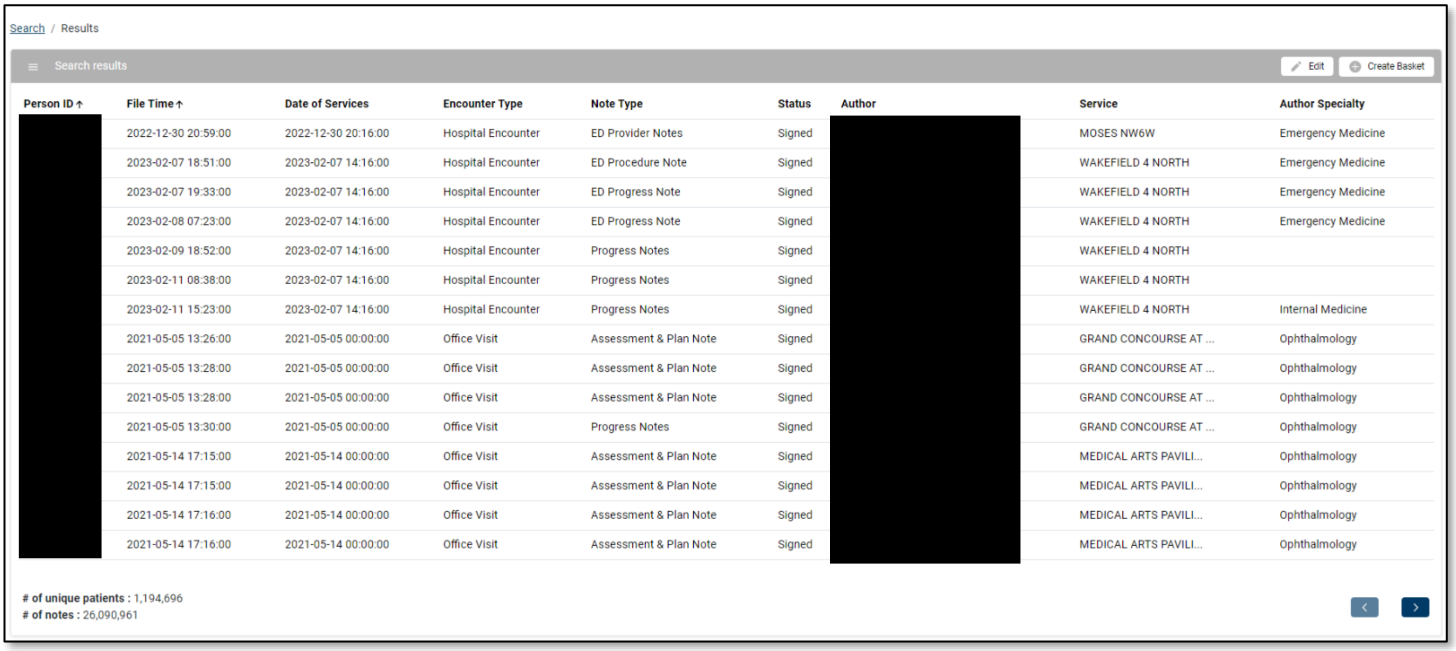
Contains (Negative Context): Phrase to search with negative context, Enter Here...

AND

Not contains: Phrase to Exclude, Enter Here...

```
( NoteText contains 'Example Phrase 1' OR 'Example Phrase 2') OR (( NoteText contains_positive_context 'Phrase to search with positive context' AND NoteText contains_negative_context 'Phrase to search with negative context') AND (NoteText not contains
```

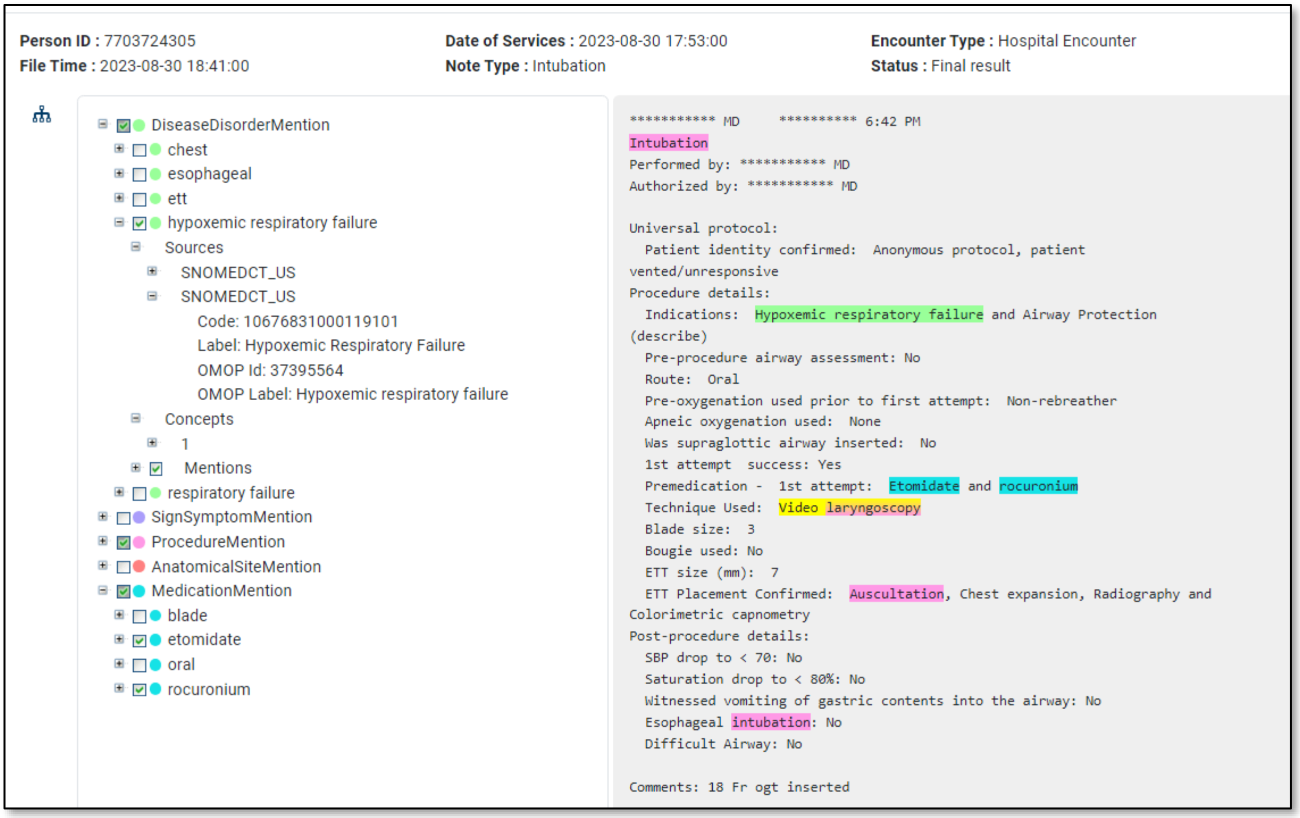
Figure 2: Text-based Query Builder



Person ID ↑	File Time ↑	Date of Services	Encounter Type	Note Type	Status	Author	Service	Author Specialty
[redacted]	2022-12-30 20:59:00	2022-12-30 20:16:00	Hospital Encounter	ED Provider Notes	Signed	[redacted]	MOSES NW6W	Emergency Medicine
[redacted]	2023-02-07 18:51:00	2023-02-07 14:16:00	Hospital Encounter	ED Procedure Note	Signed	[redacted]	WAKEFIELD 4 NORTH	Emergency Medicine
[redacted]	2023-02-07 19:33:00	2023-02-07 14:16:00	Hospital Encounter	ED Progress Note	Signed	[redacted]	WAKEFIELD 4 NORTH	Emergency Medicine
[redacted]	2023-02-08 07:23:00	2023-02-07 14:16:00	Hospital Encounter	ED Progress Note	Signed	[redacted]	WAKEFIELD 4 NORTH	Emergency Medicine
[redacted]	2023-02-09 18:52:00	2023-02-07 14:16:00	Hospital Encounter	Progress Notes	Signed	[redacted]	WAKEFIELD 4 NORTH	
[redacted]	2023-02-11 08:38:00	2023-02-07 14:16:00	Hospital Encounter	Progress Notes	Signed	[redacted]	WAKEFIELD 4 NORTH	
[redacted]	2023-02-11 15:23:00	2023-02-07 14:16:00	Hospital Encounter	Progress Notes	Signed	[redacted]	WAKEFIELD 4 NORTH	Internal Medicine
[redacted]	2021-05-05 13:26:00	2021-05-05 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	GRAND CONCOURSE AT ...	Ophthalmology
[redacted]	2021-05-05 13:28:00	2021-05-05 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	GRAND CONCOURSE AT ...	Ophthalmology
[redacted]	2021-05-05 13:28:00	2021-05-05 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	GRAND CONCOURSE AT ...	Ophthalmology
[redacted]	2021-05-05 13:30:00	2021-05-05 00:00:00	Office Visit	Progress Notes	Signed	[redacted]	GRAND CONCOURSE AT ...	Ophthalmology
[redacted]	2021-05-14 17:15:00	2021-05-14 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	MEDICAL ARTS PAVILI...	Ophthalmology
[redacted]	2021-05-14 17:15:00	2021-05-14 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	MEDICAL ARTS PAVILI...	Ophthalmology
[redacted]	2021-05-14 17:16:00	2021-05-14 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	MEDICAL ARTS PAVILI...	Ophthalmology
[redacted]	2021-05-14 17:16:00	2021-05-14 00:00:00	Office Visit	Assessment & Plan Note	Signed	[redacted]	MEDICAL ARTS PAVILI...	Ophthalmology

# of unique patients : 1,194,696
# of notes : 26,090,961

Figure 3: Search Results



Person ID : 7703724305
File Time : 2023-08-30 18:41:00

Date of Services : 2023-08-30 17:53:00
Note Type : Intubation

Encounter Type : Hospital Encounter
Status : Final result

- ⊟ ☑ ● DiseaseDisorderMention
  - ⊞ □ ● chest
  - ⊞ □ ● esophageal
  - ⊞ □ ● ett
  - ⊟ ☑ ● hypoxemic respiratory failure
    - ⊟ Sources
      - ⊞ SNOMEDCT_US
      - ⊟ SNOMEDCT_US
        - Code: 10676831000119101
        - Label: Hypoxemic Respiratory Failure
        - OMOP Id: 37395564
        - OMOP Label: Hypoxemic respiratory failure
    - ⊟ Concepts
      - ⊞ 1
    - ⊞ ☑ Mentions
  - ⊞ □ ● respiratory failure
- ⊞ □ ● SignSymptomMention
- ⊞ ☑ ● ProcedureMention
- ⊞ □ ● AnatomicalSiteMention
- ⊟ ☑ ● MedicationMention
  - ⊞ □ ● blade
  - ⊞ ☑ ● etomidate
  - ⊞ □ ● oral
  - ⊞ ☑ ● rocuronium

```
*********** MD     ********** 6:42 PM
Intubation
Performed by: *********** MD
Authorized by: *********** MD

Universal protocol:
  Patient identity confirmed:  Anonymous protocol, patient
vented/unresponsive
Procedure details:
  Indications:  Hypoxemic respiratory failure and Airway Protection
(describe)
  Pre-procedure airway assessment: No
  Route:  Oral
  Pre-oxygenation used prior to first attempt:  Non-rebreather
  Apneic oxygenation used:  None
  Was supraglottic airway inserted:  No
  1st attempt  success: Yes
  Premedication -  1st attempt:  Etomidate and rocuronium
  Technique Used:  Video laryngoscopy
  Blade size:  3
  Bougie used: No
  ETT size (mm):  7
  ETT Placement Confirmed:  Auscultation, Chest expansion, Radiography and
Colorimetric capnometry
Post-procedure details:
  SBP drop to < 70: No
  Saturation drop to < 80%: No
  Witnessed vomiting of gastric contents into the airway: No
  Esophageal intubation: No
  Difficult Airway: No

Comments: 18 Fr ogt inserted
```

Figure 4: Example clinical text with annotations, mappings and deidentification

Results

As of October 2023, over 134 million notes have been ingested and annotated by the NLP platform. Each note's annotations are made available through the scalable infrastructure. Four cohort projects which required the use of clinical text were attempted as a proof of concept using the platform. Each cohort's clinical text component was successfully completed using the NLP engine. Result lists for the queries are retrieved in microseconds, and the just-in-time deidentification of the note text is currently completed at an average rate of around 2 seconds. The deidentification is currently being optimized to distinguish between names of people and concept/medical terminology names. Research users have been able to successfully create complex cohorts using variables obtained from the clinical text.

Conclusion

The NLP engine that we developed and integrated to Atlas has provided our users the ability to build advanced cohorts using discrete data from the OMOP-CDM as well as concepts derived from the clinical text. The cTAKES and Elastic Search backend has been successfully implemented to provide extremely quick annotations and retrieval of text, most of which is completed in microseconds for real-time, highly scalable searches. The addition of LLM enabled deidentification is processed just-in-time for the research user which enables them to successfully create cohorts using highly complex queries pursuant to their needs without any preordination. IRB linking with systems such as IRIS and BRAINY allows for research users to view de-identified notes while maintaining an audit trail to protect PHI. The results of this project demonstrate the potential of integrating NLP engines to Atlas's cohort building capabilities.

References:

1. Yuan, C., Ryan, P. B., Ta, C. N., Guo, Y., Li, Z., Hardin, J. R., Makadia, R., Jin, P., Shang, N., Kang, T., & Weng, C. (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, *26*(4), 294–305. https://doi.org/10.1093/jamia/ocy178

2. Tu, S. W., Peleg, M., Carini, S., Bobak, M. T., Ross, J. M., Rubin, D. L., & Sim, I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics*, *44*(2), 239–250. https://doi.org/10.1016/j.jbi.2010.09.007
3. Weng, C., Wu, X., Luo, Z., Boland, M. R., Theodoratos, D., & Johnson, S. M. (2011). EliXR: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association*, *18*(Supplement 1), i116–i124. https://doi.org/10.1136/amiajnl-2011-000321
4. Savova, Guergana; Masanz, James; Ogren, Philip; Zheng, Jiaping; Sohn, Sunghwan; Kipper-Schuler, Karin and Chute, Christopher. 2010. Mayo Clinic Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA 2010;17:507-513 doi:10.1136/jamia.2009.001560