Optum

Natural language processing methods

Anne-Marie Guerra Currie, PhD, Senior Director, Data Science
Bertrand Lefebvre, PhD, Principal Data Scientist
Kazuki Shintani, MS, Principal Data Scientist
Tonnam Balankura, PhD, Senior Data Scientist
Xiaoren Chen, PhD, Senior Data Scientist
Rachel Kenney, PhD, Senior Data Analyst

Introduction

Optum de-identified clinical assets provide researchers with an extensive range of real-world data (RWD) concepts extracted from electronic health care records (EHR) from health care systems across the United States.

Since the clinical assets were first released more than 16 years ago, ongoing data development efforts include proactively enriching our data by extracting unstructured information from clinical patient notes and making it usable in the form of de-identified structured data for researchers. This is important as specific concepts significant for understanding the progression of diseases are often not available in structured formats. All clinical data remain statistically de-identified during the data processing, allowing researchers to glean insights from these patient data in our Optum EHR data set and Optum® Market Clarity™ data set, which includes RWD that links EHRs with medical claims.

Further, manual review of unstructured data from EHRs is often impractical given the burdensome task and time commitment required. Optum can curate critical and detailed EHR data by utilizing our refined proprietary natural language processing (NLP) system, which extracts information from free-text EHRs sourced from the Optum EHR data asset in an automated, transparent and validated manner for researchers to derive key insights in a user-friendly format. This paper details this NLP methodology, including precision and recall statistics.

Thorough patient journey

Within the Optum EHR data source — our pan-therapeutic, multi-specialty repository of clinical data — there are currently 45.7 million patients and counting who have clinical notes which could be used for NLP projects. Manually reviewing hundreds of millions of documents — and manually extracting clinical data for research — is not a scalable approach. Our NLP system offers an automated solution for providing insights from a large collection of medical notes that continues to grow each day.

The combination of both the enriched data surfaced from the EHR medical notes with the structured data within Optum EHR data allows for a more complete view of the longitudinal, time-based patient journey, including diagnosis, treatment, disease progression, outcomes and cost. The data span from 2007 to the present, allowing for longitudinal research and exploration of new drug therapies in patient outcomes. This in turn can help reduce costs for future innovative and often high-cost treatments, such as in oncology immunotherapy drugs, by facilitating large-scale research studies using these data sources for RWD generation and analyses.

Evidence-based insights extracted from the patient journey can also help understand drivers of diagnosis, treatment choices, treatment evolution and potential areas to inform market strategy and education.

Optum NLP process

The NLP system uses best practices in data science and automation.[1] Our sophisticated system goes beyond simple term-matching and rules-based approaches by incorporating machine learning (ML) and deep learning (DL), to ensure the correct identification of the desired context along with process and output validation.

Our NLP system provides a scalable solution for mining clinical insights from the notes of more than 2.6 billion medical notes and counting. NLP is used for data enrichment for our foundational de-identified clinical data assets as well for conducting custom NLP projects with specific data concepts identified for extraction and validation to support defined internal and external client use cases. Quality control evaluations are integral to our process to be sure we deliver high-quality NLP output to our stakeholders and clients.

The steps of the NLP process include:

1. Development of an annotated data set to use as a gold standard for the disease of interest
2. Development of a supervised machine learning model
3. Running production models
4. Model performance evaluation
5. Post-NLP data processing
6. Manual validation to ensure quality of the data deliverable, as needed.

Our process is presented in Figure 1.

NLP general process

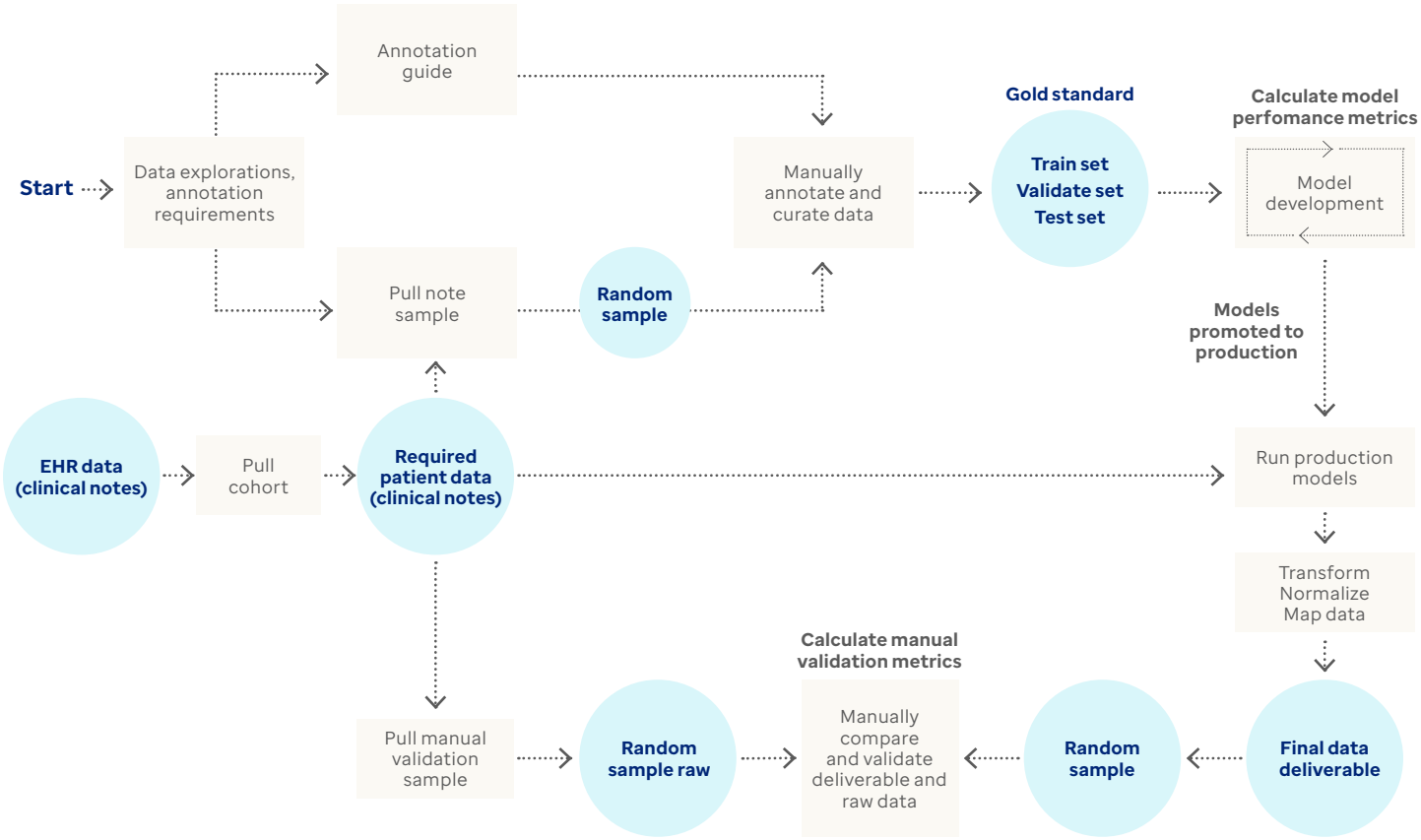


Figure 1. Supervised NLP model development and quality process

Our targeted NLP system is designed to identify the positive occurrences of desired concepts, such as disease diagnosis, procedures, stage and biomarkers, related symptoms and medication usage, as well as enable the exclusion of semantic contexts that are not desired contexts. An example of a concept would be a disease diagnosis such as prostate cancer, and an example of a context related to the concept would be positive or negative (for prostate cancer).

To further illustrate, consider the following example. To identify patients with prostate cancer, the Optum NLP system identifies different semantic contexts and appropriately extracts the desired contexts into a structured format. Clients can then easily search these concepts within our enriched data assets. Some examples of the contexts that occur within the notes are shown in Table 1:

Sample text	Context
"Patient has stage II prostate cancer"	Patient **positive** for prostate cancer
"Patient does not have prostate cancer"	Patient **negative** for prostate cancer
"If prostate cancer is found, patient may require additional imaging"	**Hypothetical** prostate cancer situation
"Might be prostate cancer"	**Hedged** prostate cancer statement
"Prostate cancer is a common cancer among males"	Prostate cancer **not relevant** to patient

Table 1. Sample of contexts for cancer statements

Information extraction process: Concepts, relations and frames

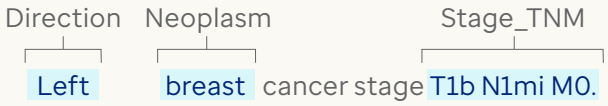
As the NLP system processes the clinical notes data, our trained models extract relevant concepts in the text and the relationships between them using 4 approaches:

1. **Concept extraction:** the extraction of a concept or entity represented by lexical units or phrases in the free text
2. **Relation extraction:** the extraction of the relationships between concepts
3. **Frame extraction:** the extraction of the logical semantic group of lexical units and the collection of any relevant relations
4. **Auxiliary tag extraction:** the extraction of the context of the concept, including the descriptive information regarding the nature of the tagged concepts, and the temporal aspects of tagged concepts.

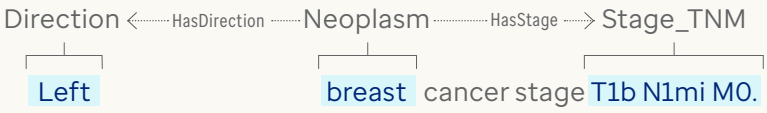
Example A shows the concept, relations and frame extraction. Individual concepts are tagged or labeled and linked to one another via relations. Relation extraction links one tag to another tag. In Example A, "neoplasm" tag links to "direction" and "stage_tnm" tags. Frame extraction groups relations originating from the same parent concept into a structure that is more easily consumable as table-like data. The frame is a logical set of semantic units, and the frame for the cancer stage context is shown extracted into table format in Example A.

Example A. Concept, relation, frame tagging and extraction. Sentence: "Left breast cancer stage T1b N1mi M0."

Concept tagging:



Relation linking:



Concept, relation, frame extraction:

Frame	Neoplasm	Direction	Stage_TNM
Neoplasm	Breast	Left	T1b N1mi MO

Annotation guide

The thoughtful crafting and designing of an annotation scheme and appropriate sampling of notes to annotate are critical to ensure high-performing NLP models. In close consultation with clinical experts (including disease or therapeutic area physician specialists, other clinicians, pharmacists, molecular biologists and medical informaticists), annotation design and sampling methods are developed by NLP data scientists specializing in clinical NLP.

During the annotation design stage, the team collaboratively outlines the concepts and contexts to annotate and extract and creates an annotation guide to meet NLP system needs. The design focuses on both the clinical context and the generalizability of the concept space to ensure scalability and extensibility of the NLP approach for our overall data enrichment.

The annotation guides are regularly updated and improved over time. Changes are tracked and reviewed in version control to ensure documentation, consistency and reliability of our process. An iterative and careful review is conducted on the annotation design by a team of diverse clinical and data science subject matter experts for clinical content and data science design structure.

Sampling strategy

During the random sampling stage, efforts are made to match population distributions of age, gender and disease diagnosis in the sample drawn, depending on project specific details. The sampling cohort is selected to ensure it contains enough information related to special sets, or rarer concepts, related to the custom project specifications.

Negative sampling is used to ensure the model is trained to not only understand and extract pertinent concepts, but to recognize the absence of these concepts. Sampling strategies are developed based on project specifications and may differ for custom projects. Some examples of strategies that may be used include:

1. **Filtering notes by ICD-10-CM codes** to ensure there will be enough concepts of interest
2. **Sampling selected note snippets** based on keywords
3. **Including negative controls** – for example, including chronic kidney disease (CKD) stage 4 concepts to ensure the model could distinguish this from lung cancer stage 4 concepts
4. **Including educational notes** to test the model's ability to identify information not directly linked to the patient

Gold standard development

Once the annotation design and the random sampling steps are completed, a random sample of data is drawn, and additional refinements are made to the specifications during the annotation process. Each note in the sample is annotated separately by two annotators and any conflicts are resolved in a third review by a curator. This process occurs with each document in our sample.

To quality check the annotation reliability, Krippendorff's Alpha (KAlpha)[2] values are calculated. Across NLP initiatives encompassing gastroenterology and neurology clinical areas, the range of KAlpha scores were 0.87–1.0, indicating strong inter-annotator agreement. While 2 of the calculated KAlpha scores were 0.87, the majority of KAlpha scores for concept and context annotation were above 0.93, demonstrating excellent reliability.[3]

Model development and performance results

Once the annotated data go through the review process (curation), which resolves disagreement between the 2 annotators, these curated data become the gold standard. Before beginning the NLP modeling phase, this ground truth, or gold standard, data set is then split into 3 subsets: training, validation and test data.

Using statistical models, such as conditional random fields and support vector machines, NLP models learn the contextual patterns from the annotated training data and then apply labeling predictions (sequence or class) to incoming data.[4,5] Classifiers are used to extract contextual information, which provides descriptive information about the tagged concepts, such as if the disease statement was affirmative or negated for the disease.

Our models are then evaluated against a holdout annotated test set, which is the third partitioned set of the gold standard annotated data that neither the models nor the NLP engineers have seen before. The results of the model performance on the holdout annotated test set are evaluated to determine if the model meets specific success criteria and will be generalizable to new data.

Performance of our NLP models is measured by calculating:

- **Precision,** also known as positive predictive value (PPV)
- **Recall,** also known as sensitivity
- **F1,** which is the harmonic mean of precision and recall

Definitions of these metrics are in Appendix 1.

Next, the best performing model that had optimal precision and recall scores is selected, guided by our internal baseline target goals of 0.80 precision (PPV) and 0.70 recall (sensitivity) for all the key concepts required for the custom project. High precision relates to a low false positive rate while high recall relates to a low false negative rate.

Precision for all key concepts in our recent NLP initiatives — including oncology, gastroenterology and neurology — exceeded 0.90, demonstrating excellent model performance. F1 scores allow us to find a balance between precision and recall. Shown below in Table 2 is an example of metrics for our oncology product.

Table 2. Oncology NLP model performance metrics

	Model	Precision (PPV)	Recall (Sensitivity)	F1 Score	Support
Oncology Neoplasm					
Neoplasm	Concept Recognition	0.97	0.75	0.85	5781
Oncology Histology					
Histology	Concept Recognition	0.94	0.85	0.89	4246
Relation to Neoplasm		0.97	0.65	0.78	2527
Oncology Stage					
Stage Value	Concept Recognition	0.96	0.93	0.95	1154
Stage TNM	Concept Recognition	0.96	0.90	0.93	952
Oncology Metastasis					
Metastasis	Concept Recognition	0.89	0.72	0.80	1584

Note: Performance metrics are calculated from the application of the NLP models on the test set of clinical notes. The support is the number of occurrences of the concept in the entire note sample for all patients. Each note can have multiple occurrences or no occurrences of a concept.

Once models are finalized, they are run at scale in a distributed manner on our collection of patient notes. Key concepts are extracted, normalized and, when possible, linked to controlled vocabularies and ontologies to reduce the variability of the output and facilitate analysis.

Challenges of NLP data

Working with real-world clinical data presents several challenges. For one, the data reflect clinical care processes that are inherent throughout in the U.S. health care system and that includes what's entered (or not entered) into the EHR record, and how. The NLP team can only extract information that is documented in the clinical notes. Notes can contain misspellings and typos. They can be biased, incomplete or ambiguous. They may even contain incorrect or contradictory information.

Sometimes, notes may also contain irrelevant or non-useful information, such as instructions or generic statements about symptoms or medications not pertaining to the patient, and it can be difficult to distinguish patient-centric data. Use of provider-specific acronyms and abbreviations in clinical notes can cause ambiguity that is difficult for the NLP models to differentiate. For example, consider the challenge for an NLP model to distinguish "Her," referencing the patient, from "Her," referencing "Her 2" receptor in a note. Data may even be formatted into a table in the notes that is challenging to solve as plain text input.

Given the large volume of notes that Optum can access, it is not within our scope to verify the clinical accuracy of the information presented in each individual note. To help address this, Optum performs quality checks to ensure the integrity of the data received from our data sources. We verify that the notes have gone through the extract, transfer and load (ETL) process without any loss of information. However, Optum does not attempt to reconcile the content of the notes themselves.

For instance, a clinician may document in the same note that a patient sometimes has nausea but sometimes does not, which may lead to both a present context and an absent context for nausea for the same note. The Optum NLP models cannot compensate for, or overcome, incomplete, discordant or erroneous documentation.

Limited data volume is also a perennial challenge for machine learning tasks. Many concepts rarely occur in the notes, so our models do not have enough information to make accurate inferences about them.

As we accumulate more data over time, our models evolve accordingly, but we cannot extract concepts that are not present in sufficient volumes in the notes. Additionally, models cannot always accommodate all data patterns and modeling efforts consist of trade-offs, meaning it is unlikely that all models will have 100% precision (PPV) and 100% recall (sensitivity). These constraints are the same for any organization seeking to extract data from notes.

While EHR systems are extremely valuable, there are certain limitations associated with the use of EHR data, including:

- Presence of a written prescription does not indicate that the patient filled the medication, although a written prescription for the same medication over time is an appropriate proxy. Also, patient-reported medications within the EHR may assist in determining what the patient is consuming at each visit
- Patients may seek care outside of their usual health care provider network and these data may not be present in the EHR data available to Optum
- EHR system users will often use diagnosis codes when they are “ruling out” an event, so special attention needs to be paid to additional supporting information when determining diagnoses, such as multiple diagnoses on multiple encounters, supporting lab results or treatments
- EHR systems will often not indicate when an event, diagnosis or medication has been discontinued
- The study patient sample will be drawn from patients treated at specific locations and may not be generalizable to all patients with the disease of interest

Patient cohort selection

Project-specific cohorts are created by including patients who meet specific eligibility criteria, such as having a disease of interest-related ICD-9-CM or ICD-10-CM code. Eligibility criteria are determined through discussions between the NLP team and the client and are based on custom elements related to the project. Potential selection biases are considered, and efforts are made to minimize these biases in the final cohort.

Post-NLP data processing

Once the finalized NLP models have been run, the Optum Data Development team performs data transformations to prepare the data for clients. The goal of these transformations is to ensure that the final deliverable conforms to client needs and project specifications. The Optum Data Development team designs final data deliverables that are concise and easy to query, while faithfully and accurately representing the data extracted from the note.

Manual validation quality check

Along with providing model metrics to clients, we can provide additional transparency on the quality of the NLP output by conducting a manual validation study on the client delivered data, if required. Manual validation allows us to provide more insight into the quality of the data in the final tables delivered to our clients.

Using a sample of the delivered data, we may conduct a manual review of the NLP-extracted structured concepts and compare this output with the original source notes. The following section walks through the process of sample selection and manual comparison of the final NLP table output with the original note source data.

Sampling strategy for manual validation

To perform manual validation, a representative sample is drawn from the final deliverable for the concepts we want to validate. The sample is selected so that there are similar proportions of demographic groups to those in the final deliverable. The demographic variables we consider are gender, race, ethnicity and geographic region. We use Cochran's formula[6] to determine the minimum sample size needed to ensure a robust sample size with a 95% confidence interval and ±5% precision.

Manual validation precision metrics

Manual validation is a 2-part quality check process. First, 2 validators independently verify that the extracted concepts are found in the original note. For example, if there is a discrepancy, each validator would go into the original note, check whether the note truly contains the concept mentioned, and evaluate if the concept was extracted correctly.

The second part of the process involves a third validator, who resolves any differences that result from the first step. The third validator's results are used to calculate a precision score for each concept as described in the model development and performance section. Results of the manual validation for recent custom projects demonstrate excellent precision scores, ranging from 0.87–0.99. While one precision metric score was slightly lower at 0.87, the remaining scores were 0.91 and above. Manual validation requirements are specified within the contract terms.

Summary: Value of Optum NLP data

A recent review article published in 2022 found 123 articles using NLP methods to extract oncology-related concepts for research publications.[7] While a wide range of data sources were used for these studies, none reached the volume and range of the Optum EHR data source, which captures 45.7 million patients and 2.6 billion notes, of which 5 million patients and 411 million notes have oncology-related information.

The advantage of using supervised machine learning algorithms is the ability to accurately identify the appropriate contexts in an automated fashion over highly variable text. Our supervised machine learning models are trained to identify broader patterns that are not explicitly and manually noted by a human as a rule, but instead the machine learns from a sample of labeled data that will then enable the system to generalize to relevant contexts.

Our models are evaluated against a holdout annotated test set which the algorithm has not seen before. The results of this test help ensure we are not overfitting to the training data and that the model will remain reliably accurate with new data.

Despite its limitations, the NLP process and the data it generates has many strengths. The ability to use EHR clinical data from clinical notes and documents has many implications for understanding medication efficacy, disease management and risk stratification for patients with various diseases.

Automated NLP algorithms catalyze availability of structured data from an unstructured source which would take an unmeasurable number of hours to review and analyze manually. These models provide reliable data regarding symptomology and medication use in a reasonable time horizon and in an avenue that has not been previously explored. The Optum Market Clarity data set[8,9] — our integrated claims and EHR asset — includes data with diversity in age, gender, race, ethnicity and region that increases generalizability to the general population compared to studies including only a small sample or patients from one area or racial or ethnic group.

The advantages of the Optum NLP approach include methodological rigor, combination of scalable techniques, comprehensive extraction, and extraction that is consistent and reliable. Overall, the combination of rules, traditional machine learning and deep learning techniques leads to effective and highly accurate results. The extraction results for specific concepts for various diseases are consistently above 80% precision, and most often exceed 90% precision. These high-quality results allow our clients to be confident that our data are robust enough for their research purposes.

Appendix 1: Definition metrics definitions

Precision = true positives / (true positives + false positives)

Precision, also called positive predictive value, is a measure of correctly identified positive cases from all the predicted positive cases.

It aims to measure how many of all tags that were labeled were correct.

Recall = true positives / (true positives + false negatives)

Recall, also called sensitivity, is the measure of the correctly identified positive cases from all the actual positive cases.

It aims to measure how many of all the tags that truly occurred were accurately captured by the NLP model.

F1 = 2 x precision x recall / (precision + recall)

The F1 score is the harmonic mean of precision and recall.

The score takes both false positives and false negatives into account.

Sources

1. Optum. Ensuring responsible use of AI in health care. Accessed 2023
2. Krippendorff, K. (2013) *Content Analysis. An Introduction to Its Methodology.* 3rd ed. Thousand Oaks, California. Sage Publications.
3. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* (Zagreb) 2012;22:276-282
4. Reading Turchioe M, Volodarskiy A, Pathak J, et al. Systematic review of current natural language processing methods and applications in cardiology. *Heart*. May 25, 2022;108:909-916.
5. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med.* March 2023;155:106649.
6. Cochran WG. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons, 1977.
7. Wang L, Fu S, Wen A, et al. Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. *JCO Clin Cancer Inform*. July 6, 2022;e2200006.
8. Optum. Market Clarity: Linked EHR and claims data.
9. Optum. Market Clarity Data.

Optum

optum.com