OPTUM®

Clinical natural language processing:

Unearthing deeper oncology insights by mining unstructured medical notes

Oncology

Anne-Marie Guerra Currie, PhD, Director, Data Science

Bertrand Lefebvre, PhD, Principal Data Scientist

Kazuki Shintani, MS, Principal Data Scientist

Tonnam Balankura, PhD, Sr. Data Scientist

Optum oncology data initiatives include enriching our data by extracting essential information from the oncology patient's medical records and making it usable for researchers. Specific oncology concepts important in understanding the progression of the disease are often not available in structured formats, particularly the tumor, node and metastasis (TNM) values, stage information and biomarkers.

2M patients who have at least one solid tumor ICD code

We surface this critical and detailed oncology data by leveraging our proprietary natural language processing (NLP) system that performs automated information extraction on the free-text medical records repository within the Optum electronic health record (EHR) data asset to provide key oncology-related insights to our clients in an easy-to-use format.

Our oncology-focused NLP system is designed to identify the positive occurrences of desired oncology concepts, such as cancer type, TNM, stage and biomarkers, as well as enable the exclusion of semantic contexts that are not desired oncology contexts.

For example, if the goal is to identify patients with prostate cancer, the Optum NLP system identifies different semantic contexts and appropriately extracts the desired contexts into a structured format. The concepts are then able to be easily searched by our clients. Some examples of the contexts that occur within the notes are shown in Table 1:

Table 1. Sample of contexts for cancer statements

Sample text	Concept
"Patient has stage II prostate cancer"	Patient **positive** for prostate cancer
"Negative for 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

Within the Optum EHR data source, there are 2 million patients who have at least one solid tumor ICD code. 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.

Information extraction process: Entities, relations and frames

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

1 **Entity 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 entities.

3 **Frame extraction**
The extraction of the logical semantic group of lexical units and the collection of any relevant relations.

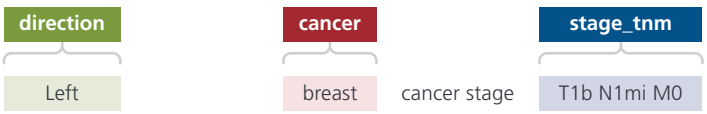
Example A shows the entity, relations and frame extraction. Individual entities are tagged, or labeled and linked to one another via relations. Relation extraction links one tag to another tag.

In Example A, "cancer" 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.

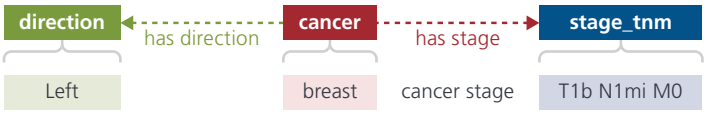
Example A. Entity, relation, frame tagging and extraction

Sentence: "Left breast cancer stage T1b N1mi M0."

Entity tagging:



Relation linking:



Entity, relation, frame extraction:

Frame	Cancer	Direction	Stage_tnm
cancer	breast	left	T1b N1mi M0

Modeling approach

The Optum oncology NLP system leverages best practices in data science and automation. Our sophisticated system goes beyond term-matching and rules-based approaches by incorporating machine learning and deep learning, in order to ensure the correct identification of the desired oncology context.

The advantage of leveraging supervised machine learning models 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 created 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 held-out annotated test set, which the models 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.

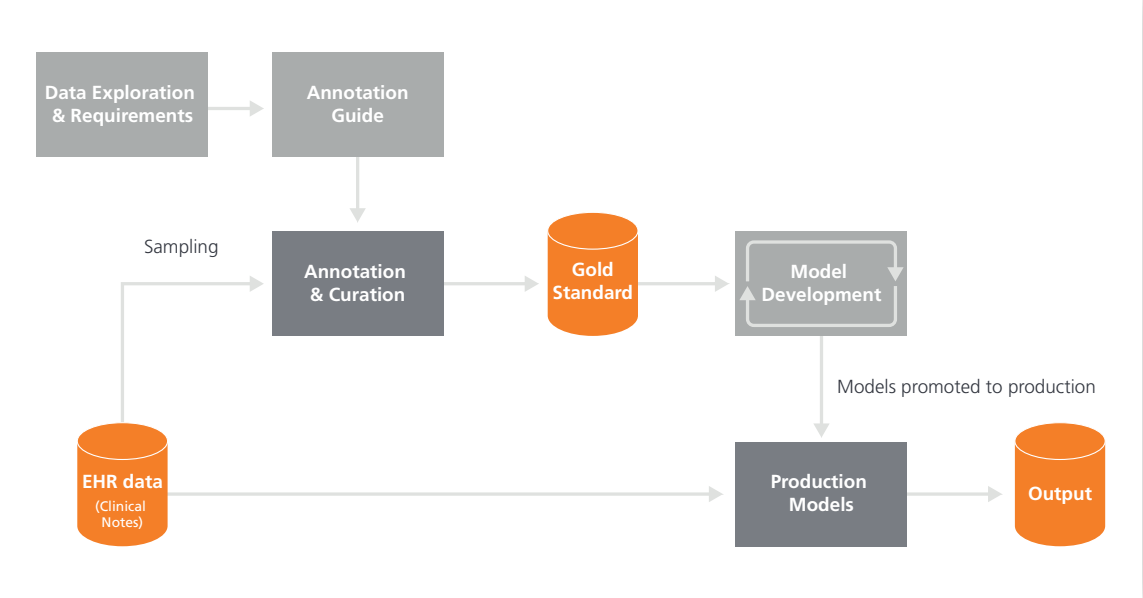
Annotation design and gold standard data development

The thoughtful crafting and designing of an annotation scheme and appropriate sampling of notes to annotate are critical to ensure high-performing NLP models. The annotation design and sampling methodology is systematically developed with NLP data scientists specialized in the field of clinical NLP in close consultation with clinical experts (oncologists, oncology clinicians, pharmacists, molecular biologists, medical informaticists and other physicians). During the annotation design stage, the team carefully and thoughtfully outlines the entities and relations to annotate and extract. This design focuses on both the clinical context as well as the generalizability of the concept space to ensure scalability and extensibility of the NLP approach for our overall data enrichment.

Our annotation guides are iteratively improved over time, and changes are tracked and reviewed in version control to ensure 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, as well as for data science design structure. Once the annotation design and the random sampling methodology are refined, a random sample of data is drawn and additional refinements may be made to the specifications during the annotation process. Each note in the sample is double-annotated by two annotators and any conflicts are resolved in a third review by a curator. This process occurs with each document in our sample. The sample is then subdivided into the subsets of train, validation and test.

Once models are finalized, these models are run at scale in a distributed manner on our collection of notes. Extracted entities are normalized in order to reduce the variability of the output and to facilitate analysis, and whenever possible, linked to controlled vocabularies and ontologies.

Figure 1. Model development and production process



Benefits & Results

The advantages of our rigorous approach and combination of techniques is scalability, comprehensive extraction, and extraction that is methodically 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 oncology concepts for cancer, stage and TNM are consistently above 90% precision and all 58 biomarkers in our product are above 80% precision. These high-quality results allow our clients to be confident that our data is robust enough for their research purposes.

OPTUM®

optum.com

11000 Optum Circle, Eden Prairie, MN 55344