Presented by John Snow LABS

Oncology Research And Treatment With Healthcare NLP

Muhammet ŞANTAŞ

Senior Data Scientist

John Snow Labs

Agenda

1. Introduction
2. Case Study Samples
3. Use Case I - Analysis Based On Cancer Types
4. Use Case II - Biomarker Table Generation
5. Coding

John Snow LABS

Healthcare AI Expert - 10 Year Growth

120+ million downloads on PyPI. "Most Widely Used NLP Library in the Enterprise."

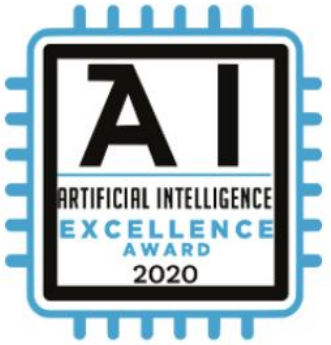
O'Reilly Media

54% share of GenAI in Healthcare projects at large companies use John Snow Labs

Gradient Flow

#1 accuracy on 25 benchmarks in peer-reviewed papers

Papers with Code



Peer-Reviewed, State-of-the-Art Accuracy

John Snow Labs Peer-Reviewed Papers

Deeper Clinical Document Understanding Using Relation Extraction

Accurate Clinical and Biomedical Named Entity Recognition at Scale

Biomedical Named Entity Recognition in Eight Languages with Zero Code Changes

Mining Adverse Drug Reactions from Unstructured Mediums at Scale

Can Zero-Shot Commercial APIs Deliver Regulatory-Grade Clinical Text DeIdentification?

Beyond Negation Detection: Comprehensive Assertion Detection Models for Clinical NLP

Oncology Case Studies from the NLP Summit

Mount Sinai

A Real-time NLP-Based Clinical Decision Support Platform for Psychiatry and Oncology

Roche

Applying Healthcare-Specific LLMs to Build Oncology Patient Timelines and Recommend Clinical Guidelines

National Cancer Centre Singapore
SingHealth

Large Language Models to Facilitate Building of Cancer Data Registries

COTA

Leveraging Healthcare NLP Models in Regulatory Grade Oncology Data Curation

myNEO

Applying Natural Language Processing to Cancer Genomics

Fred Hutch Cancer Center

"Be Like Water" – NLP and GenAI Opportunities and Challenges in Healthcare

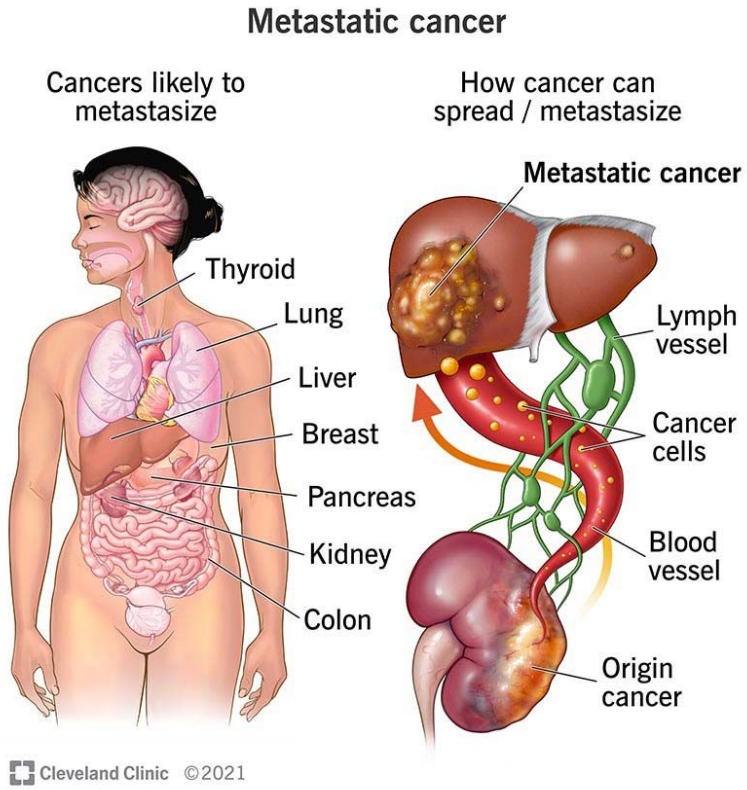
MIBA

AI-Enhanced Oncology Data: Unlocking Insights from EHRs with NLP and LLMs

Use Case I - Analysis Based On Cancer Types

Analyzing Oncological Entities and Disease Progression

To assess disease progression and identify cases where cancer has metastasized beyond its primary site, clinical documents can be filtered for mentions of metastasis. Analyzing oncological entities within these documents, along with their associations to specific body parts, offers critical insights into cancer spread patterns. This information can support more accurate staging and guide personalized treatment planning.



Real-World Oncology Use Case

Solution Steps

John Snow LABS

01

Document Filtering

Filter the documents which contain “metastasis” entity.

02

Entity Extraction (NER)

Create a robust NER pipeline, extract oncological entities and body parts.

03

Assertion Status Detection

Check the assertion status of the detected oncological entities.

04

Relation Extraction

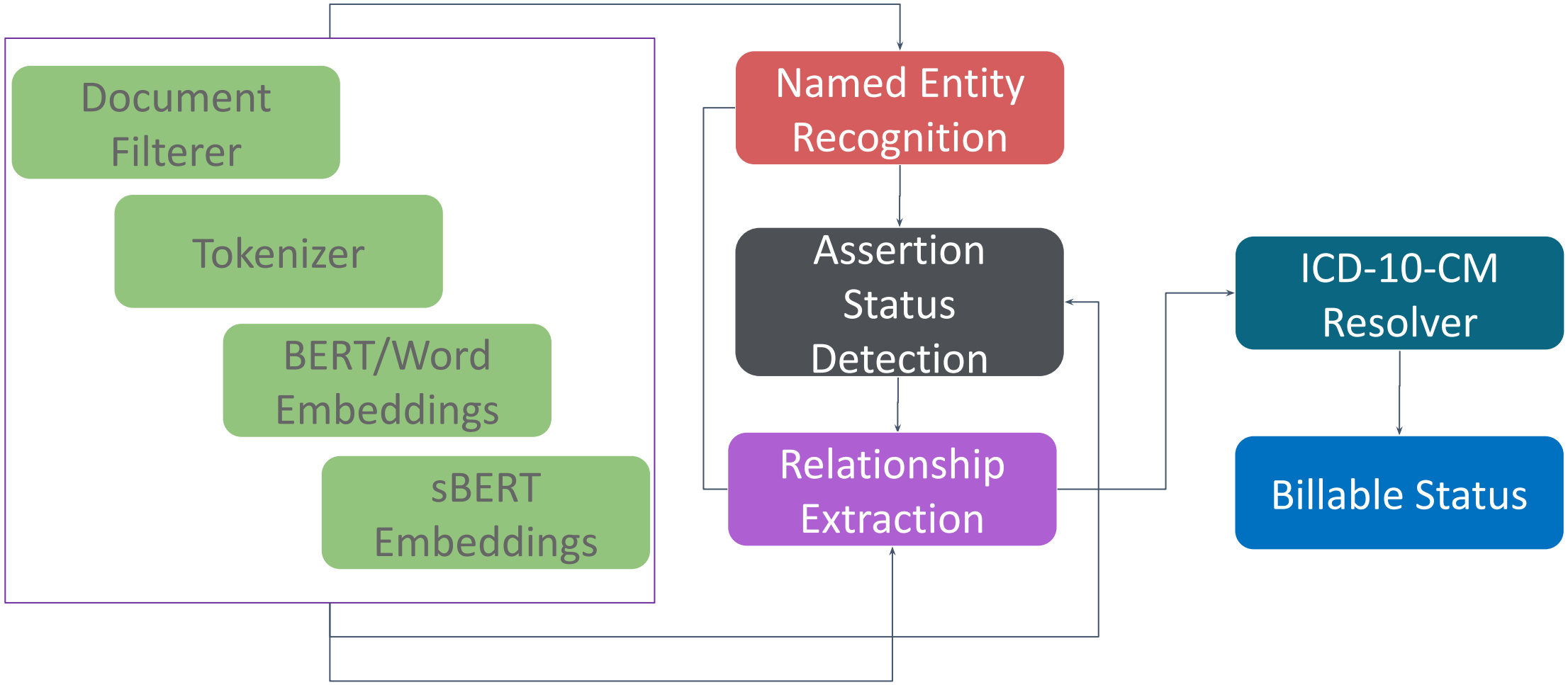
Extract relations between the oncological entities and body parts.

05

Entity Resolution

Map the entities to their corresponding ICD-10-CM codes and check their billable status.

Real-World Oncology Use Case



(Resolvers can be mapped to each other using the pretrained mapper module or resolved from scratch using the same sBert Embeddings)

Oncology NLP Models

John Snow LABS

Medical NER Models

ner_biomarker	ner_oncology_diagnosis_langtest
ner_biomarker_langtest	ner_oncology_emb_clinical_large
ner_cancer_genetics	ner_oncology_emb_clinical_medium
ner_cancer_types	ner_oncology_langtest
ner_oncology	ner_oncology_limited_80p_for_benchmarks
ner_oncology_anatomy_general	ner_oncology_posology
ner_oncology_anatomy_general_healthcare	ner_oncology_posology_langtest
ner_oncology_anatomy_general_langtest	ner_oncology_response_to_treatment
ner_oncology_anatomy_granular	ner_oncology_response_to_treatment_langtest
ner_oncology_anatomy_granular_langtest	ner_oncology_test
ner_oncology_biomarker	ner_oncology_test_langtest
ner_oncology_biomarker_docwise	ner_oncology_therapy
ner_oncology_biomarker_healthcare	ner_oncology_therapy_langtest
ner_oncology_biomarker_langtest	ner_oncology_tnm
ner_oncology_demographics	ner_oncology_tnm_langtest
ner_oncology_demographics_langtest	ner_oncology_unspecific_posology
ner_oncology_diagnosis	ner_oncology_unspecific_posology_healthcare
ner_oncology_diagnosis_langtest	ner_oncology_unspecific_posology_langtest

Relation Extraction Models

re_oncology	redl_oncology_biobert
re_oncology_biomarker_result	redl_oncology_biomarker_result_biobert
re_oncology_granular	redl_oncology_granular_biobert
re_oncology_location	redl_oncology_location_biobert
re_oncology_size	redl_oncology_size_biobert
re_oncology_temporal	redl_oncology_temporal_biobert
re_oncology_test_result	redl_oncology_test_result_biobert

Assertion Status Detection Models

assertion_oncology
assertion_oncology_demographic_binary
assertion_oncology_family_history
assertion_oncology_problem
assertion_oncology_response_to_treatment
assertion_oncology_smoking_status
assertion_oncology_test_binary
assertion_oncology_treatment_binary

Sequence Classification

bert_sequence_classifier_biomarker
bert_sequence_classifier_response_to_treatment

35 NER Models

7 Relation Extraction Models

7 Relation Extraction DL Models

8 Assertion Status Detection Models

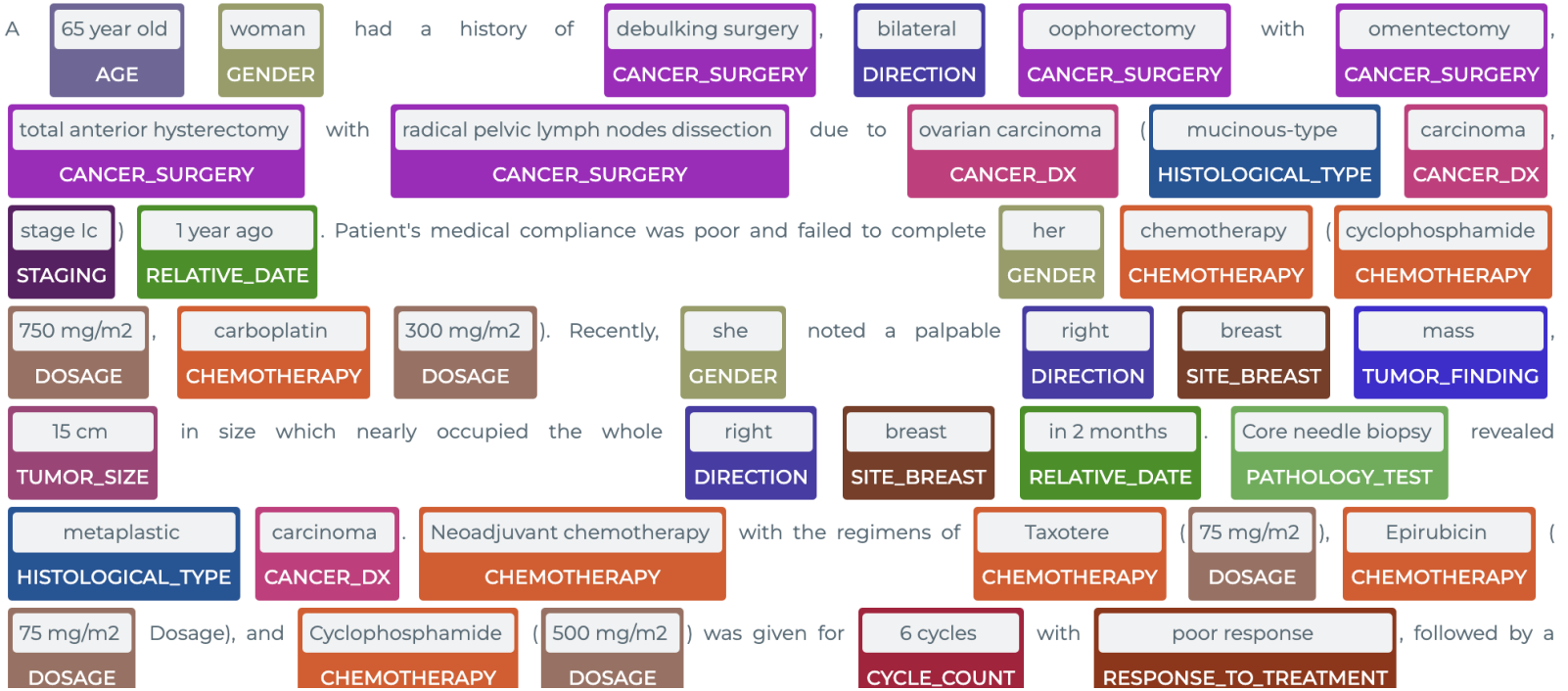
2 Sequence Classification Models

Entity Extraction

John Snow LABS

91 Entity Detection Labels

Adenopathy	Leukemia_Type
Age	Leukemia
Anatomical_Site	Line_Of_Therapy
Biomarker	Lymph_Node
Biomarker_Measurement	Lymph_Node_Modifier
Biomarker_Quant	Lymphoma_Type
Biomarker_Result	Melanoma
Body_Site	Metastasis
CNS_Tumor_Type	Oncogene
CancerDx	Other_Tumors
CancerModifier	Pathology_Result
Cancer_Score	Pathology_Test
Cancer_Surgery	Performance_Status
Cancer_Therapy	Predictive_Biomarkers
Carcinoma_Type	Prognostic_Biomarkers
Chemotherapy	RNA
Cycle_Count	Radiation_Dose
Cycle_Day	Radiological_Test
Cycle_Number	Radiological_Test_Result
DNA	Radiotherapy
Date	Response_To_Treatment
Death_Entity	Route
Direction	Sarcoma_Type
Dosage	Smoking_Status
Drug	Staging
Duration	Size_Trend
Ethnicity	Targeted_Therapy
Frequency	Test
Gender	Test_Result
Grade	Tumor
Histological_Type	Tumor_Size
Hormonal_Therapy	cell_line
Imaging_Test	cell_type
Immunotherapy	…



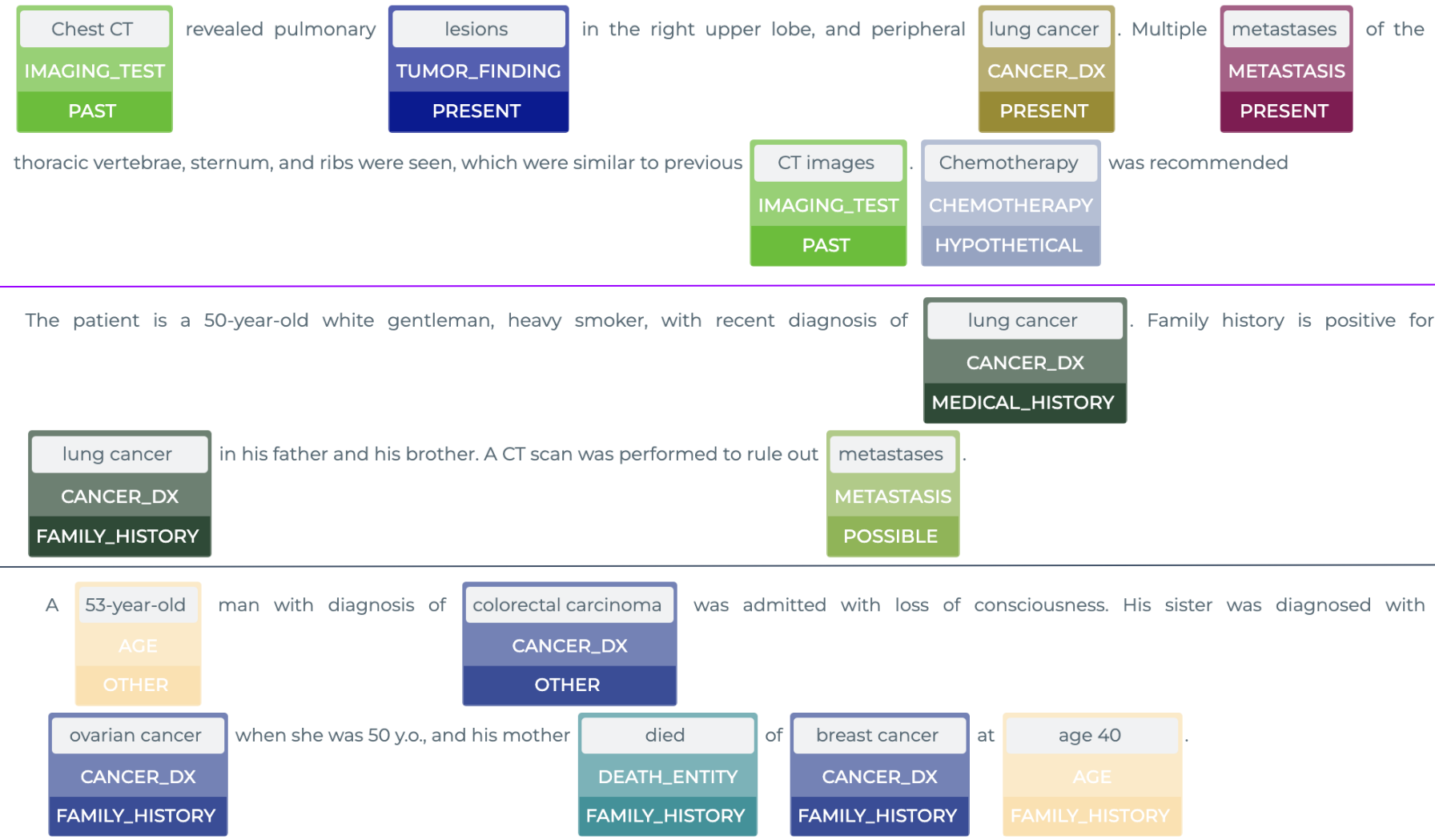

```
labels = ["Adenopathy", "Age","Biomarker","Biomarker_Result","Body_Part","Cancer_Dx","Cancer_Surgery",
        "Cycle_Count","Cycle_Day","Date","Death_Entit","Directio","Dosage","Duration","Frequency",
        "Gender","Grade","Histological_Type","Imaging_Test","Invasion","Metastasis","Oncogene","Pathology_Test",
        "Race_Ethnicity","Radiation_Dose","Relative_Date","Response_To_Treatment","Route","Smoking_Status",
        "Staging","Therapy","Tumor_Finding","Tumor_Size"]

pretrained_zero_shot_ner = PretrainedZeroShotNER().pretrained("zeroshot_ner_oncology_medium", "en", "clinical/models")\
    .setInputCols("sentence", "token")\
    .setOutputCol("ner")\
    .setPredictionThreshold(0.5)\
    .setLabels(labels)
```

Assertion Status Detection

13 Assertion Status Labels

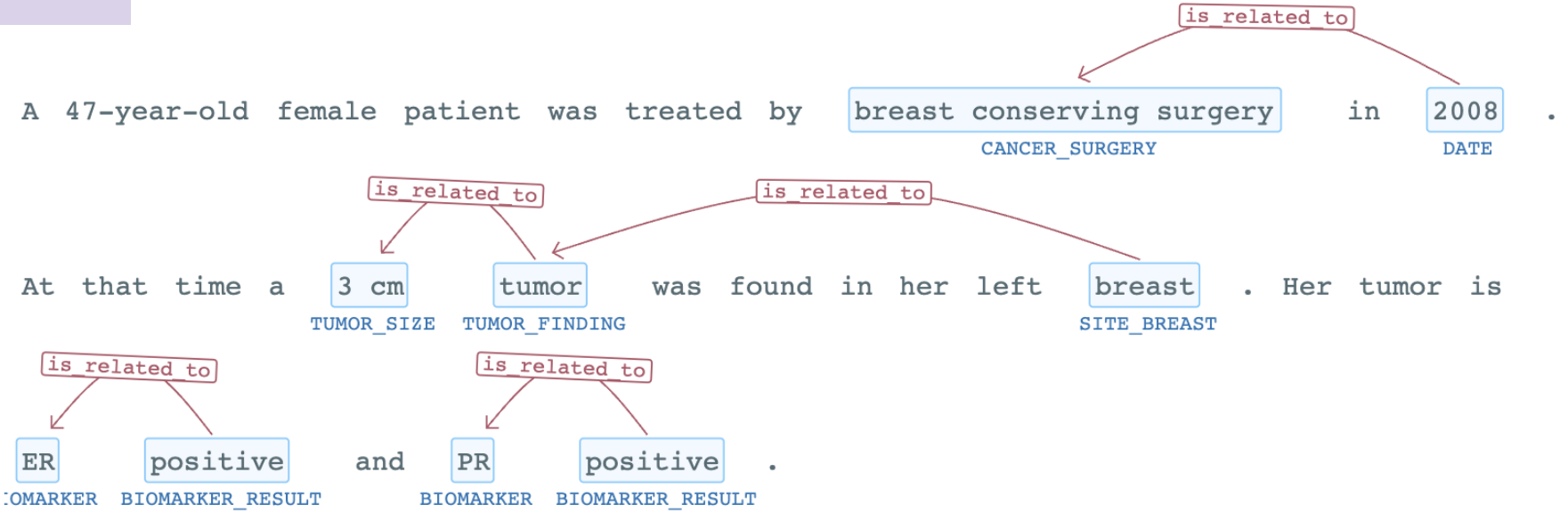
```
Present
Past
Present_Or_Past
Absent
Someone_Else
Family
Family_History
Hypothetical
Hypothetical_Or_Absent
Possible
Patient
Medical_History
Other
```

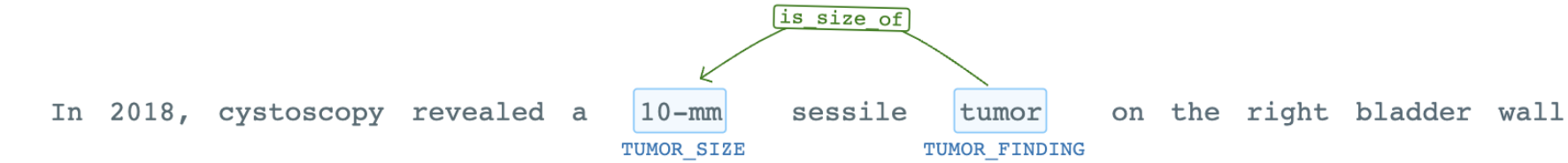


Relation Extraction

5 Relation Extraction Labels

- is_date_of
- is_finding_of
- is_location_of
- is_related_to
- is_size_of





Entity Resolution to Standard Terminologies

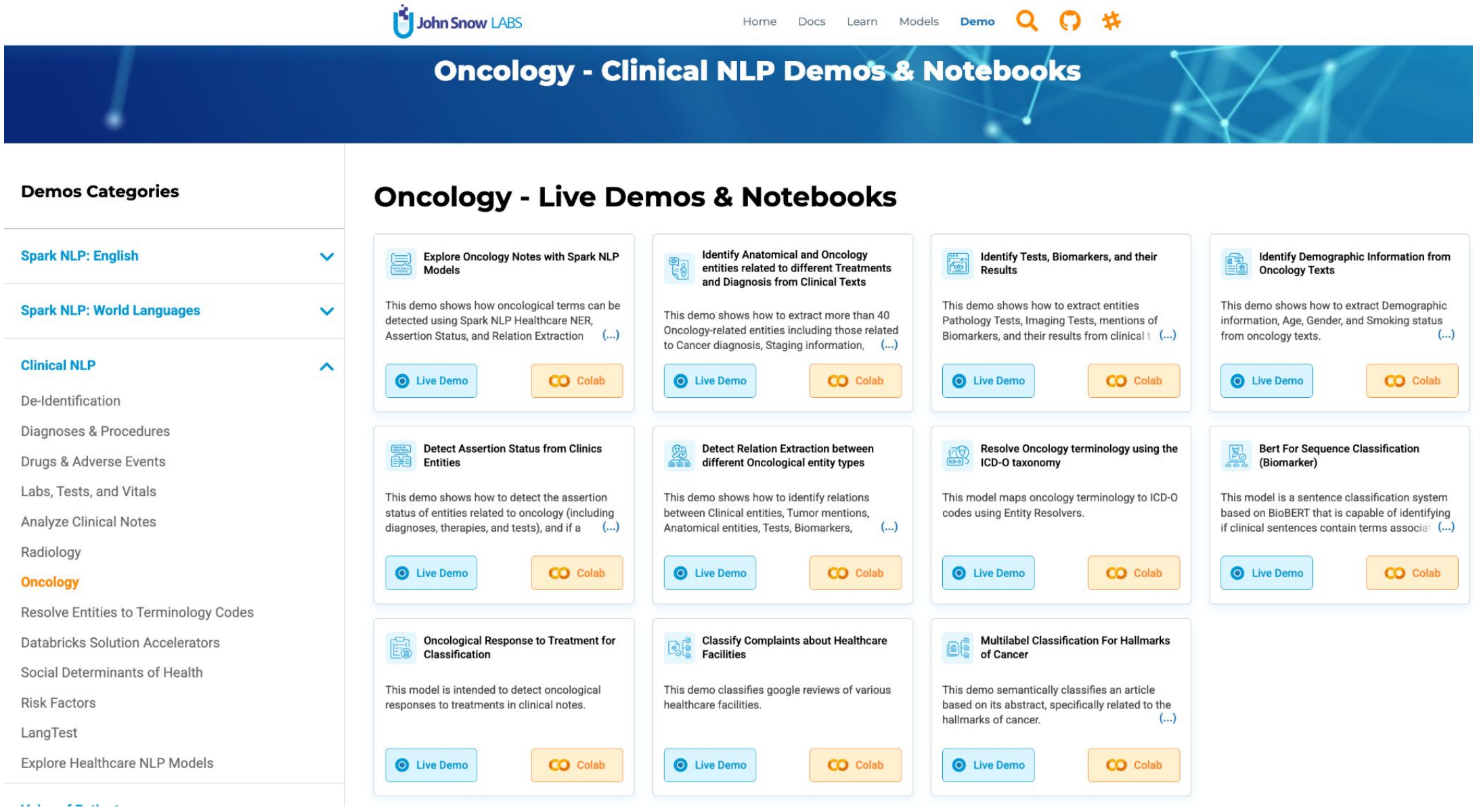
This is a 52-year-old **AGE** inmate with a 5.5 **MEASUREMENTS** cm **UNITS** diameter nonfunctioning mass **SYMPTOM** in his **GENDER** right **DIRECTION** adrenal **BODYPART** shown by CT of **IMAGINGTEST** abdomen **BODYPART** . During the umbilical hernia repair **PROCEDURE** , the harmonic scalpel **MEDICAL_DEVICE** was utilised superiorly **DIRECTION** and laterally **DIRECTION** .

Entity Resolution

ICD10CM, Snomed, RxNorm, CPT-4, ICD10CPS, RxCUI, ICDO, UMLS, ATC, HPO, …

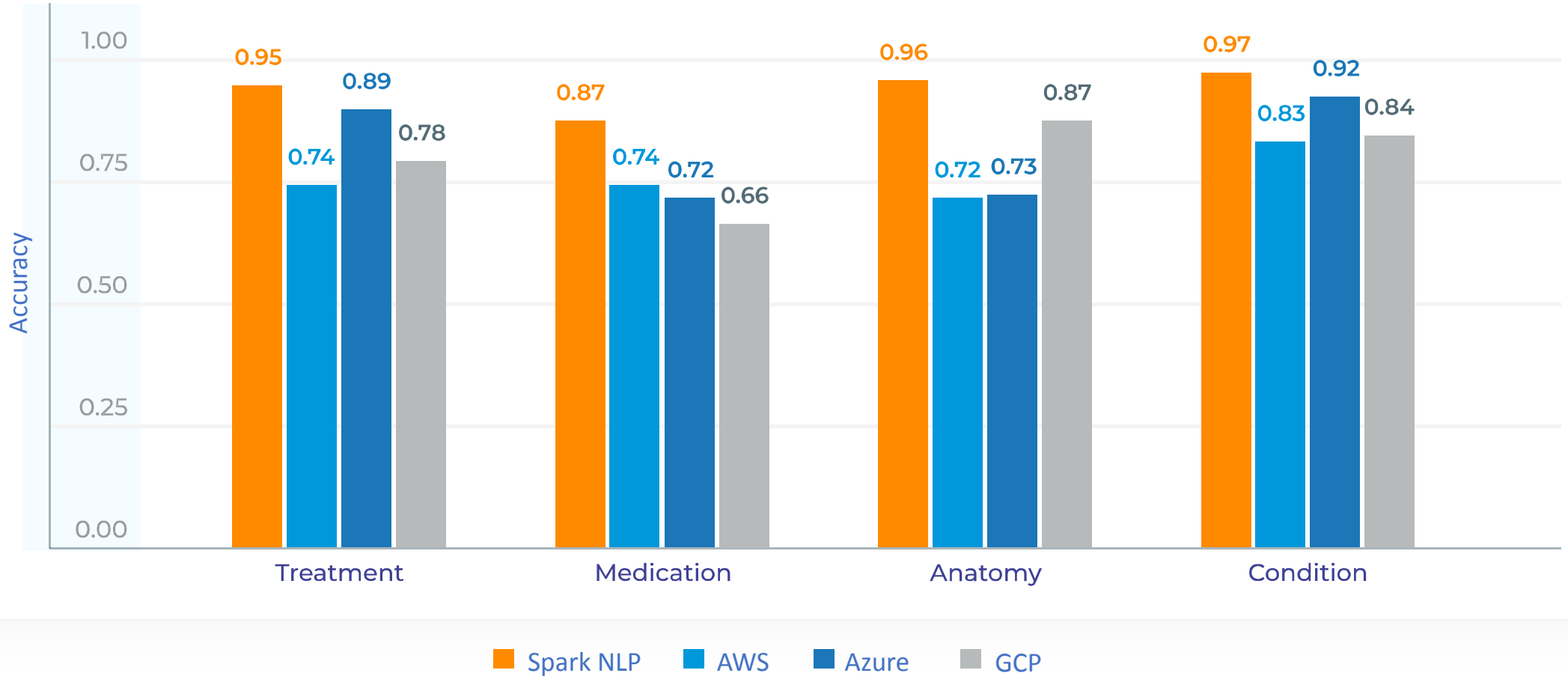
Term	Vocab	Code	Explanation (ground truth)
CT	CPT-4	76497	Unlisted computed tomography procedure
CT of abdomen	CPT-4	74150	Computed tomography, abdomen; without contrast material
umbilical hernia repair	CPT-4	49587	Repair umbilical hernia, age 5 years or older; incarcerated or strangulated
nonfunctioning mass, right adrenal	ICD10CM	D35.01	Benign neoplasm of right adrenal gland

Out-of-the-Box Oncology NLP Models

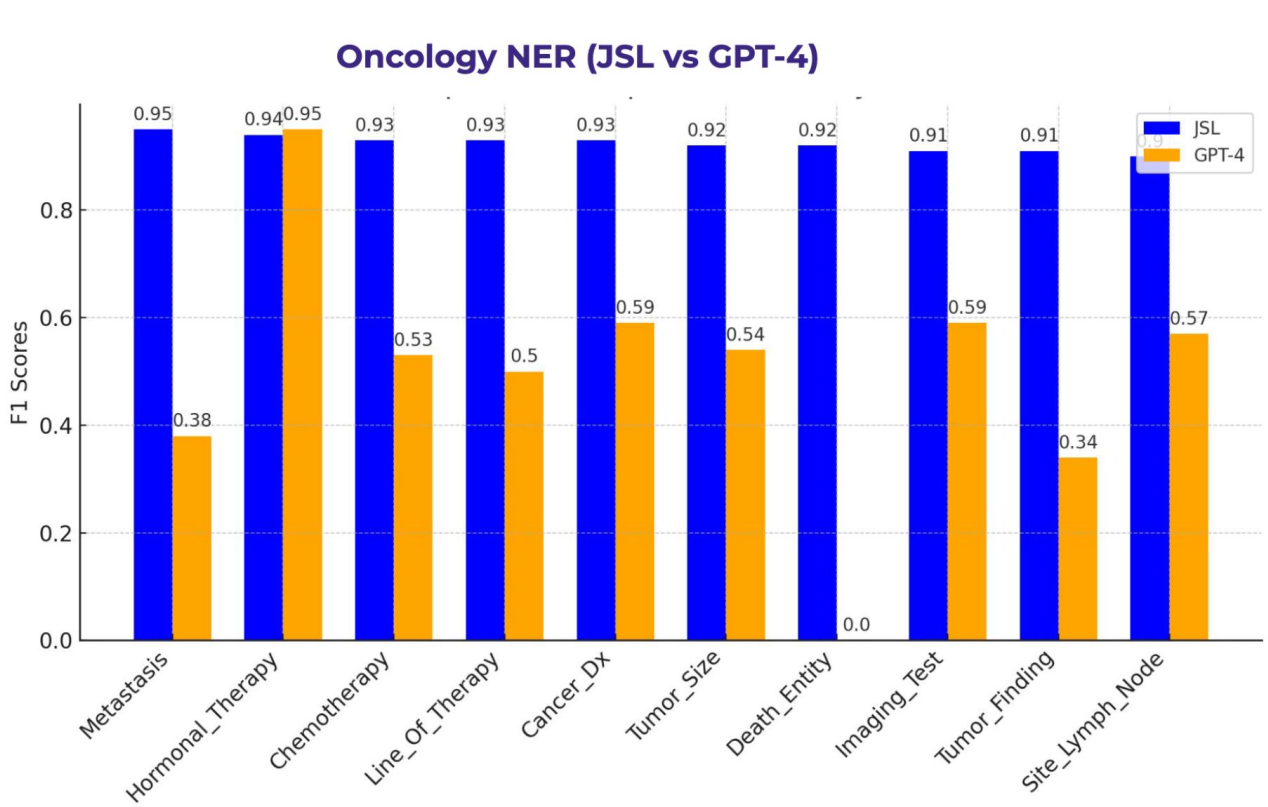


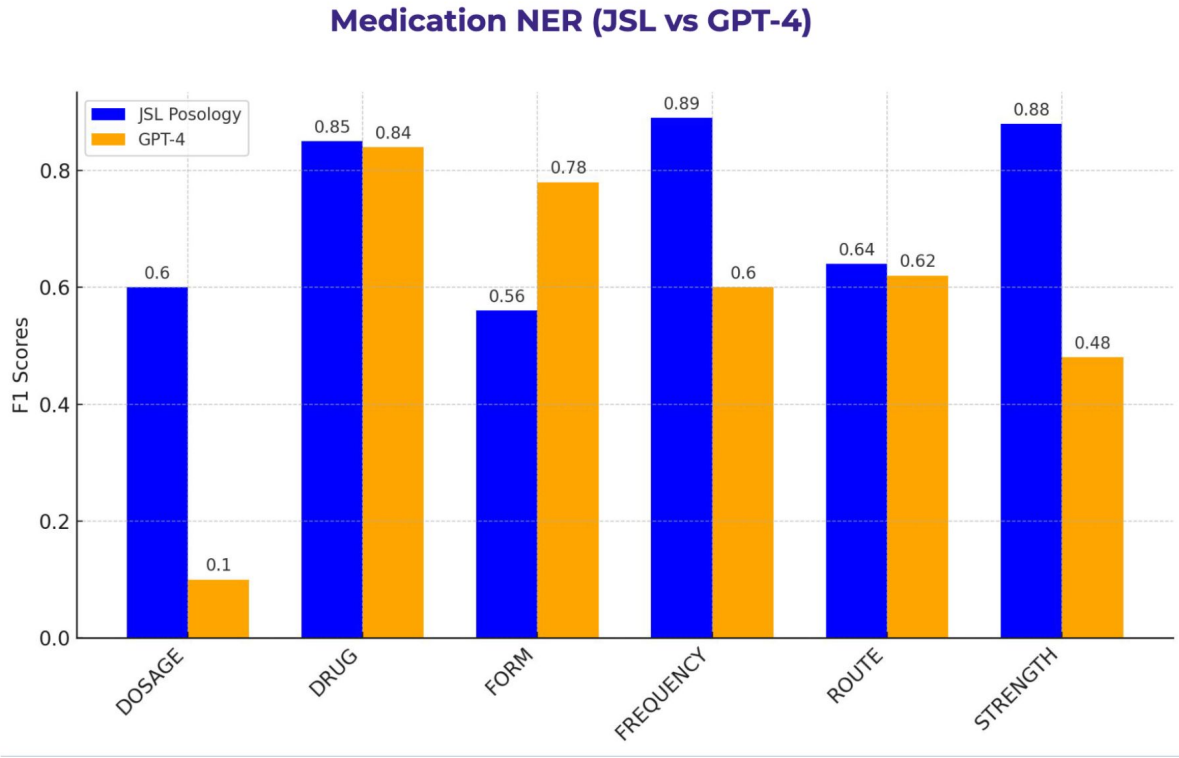
4-6x Fewer Errors than AWS, Azure, & GCP

www.johnsnowlabs.com/comparison-of-key-medical-nlp-benchmarks-spark-nlp-vs-aws-google-cloud-and-azure/



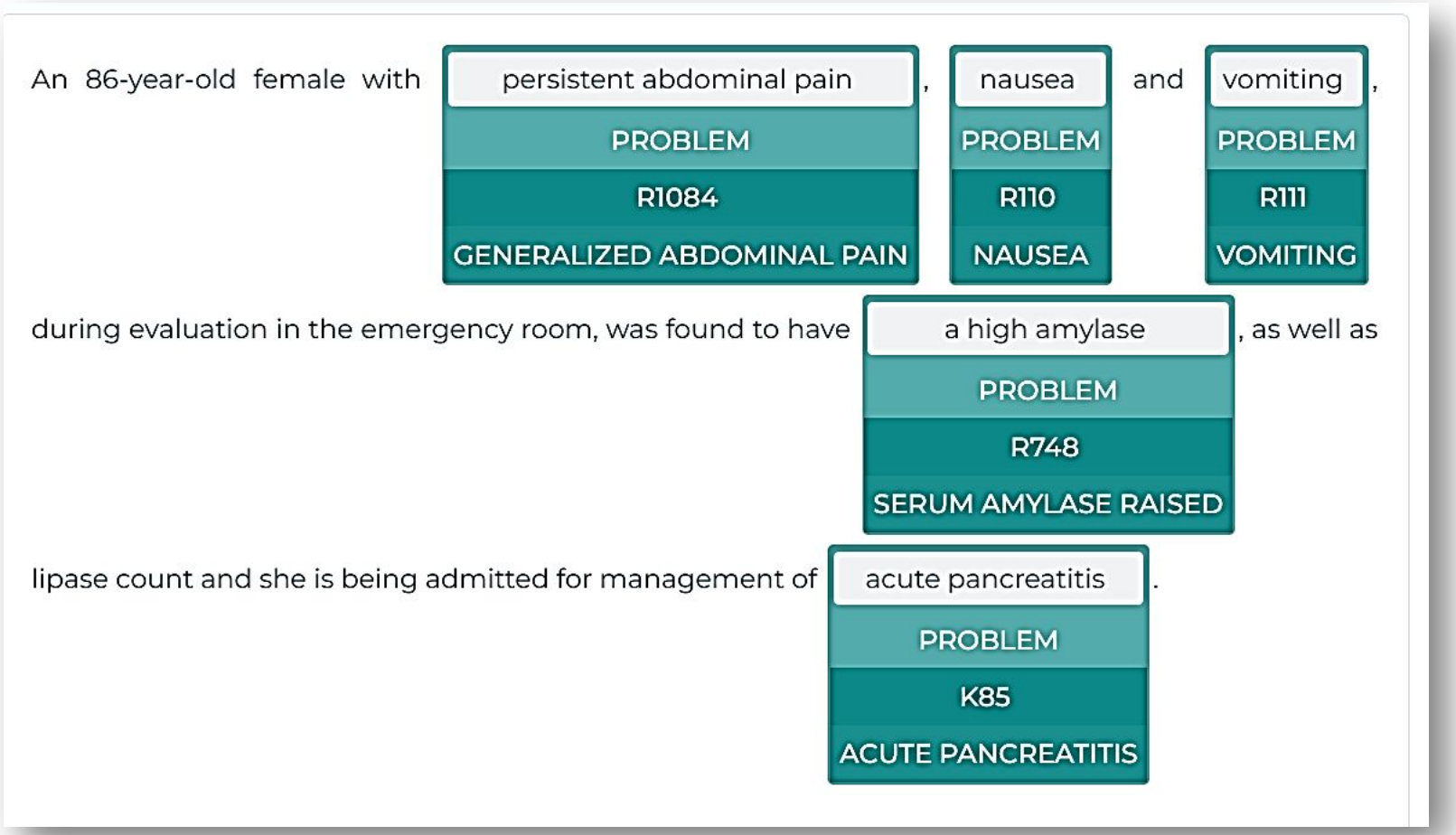
Benchmark: NER Metrics





		Spark NLP Clinical Models			AWS Medical Comprehend			GCP Healthcare API		
Entity	Sample	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Problem	4891	0.726	0.585	**0.648**	0.539	0.478	0.507	0.850	0.516	0.642
Test	5903	0.782	0.662	**0.717**	0.594	0.703	0.644	0.576	0.461	0.512
Drug	10284	0.946	0.882	0.913	0.815	0.910	0.860	0.962	0.885	**0.922**
Avg. F1				**0.759**			0.670			0.692

Benchmark: Extracting ICD-10-CM Codes



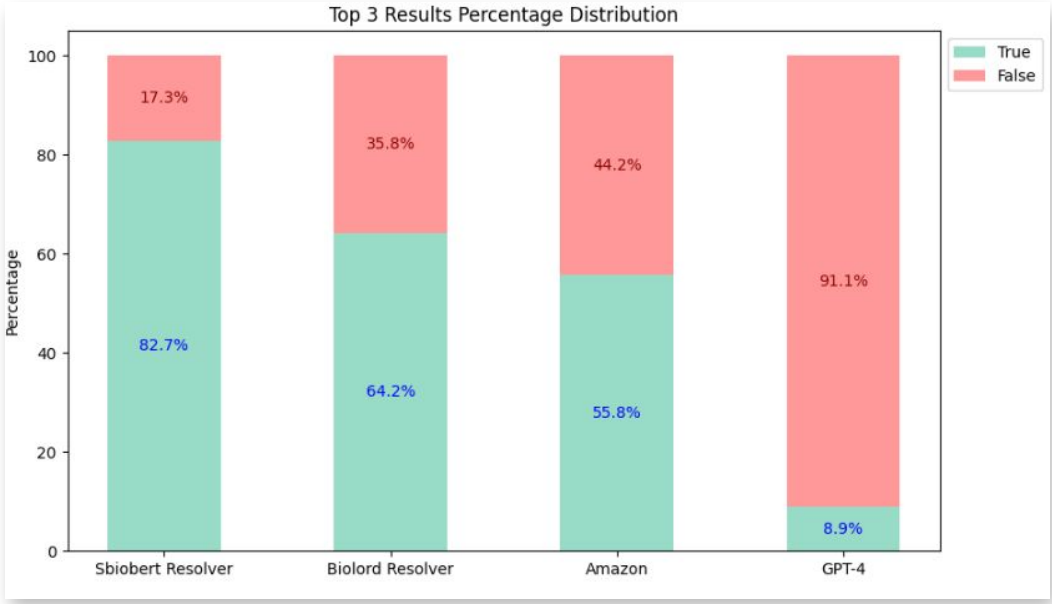
Extracting ICD-10-CM codes is done with a **76% success rate vs. 26% for GPT-3.5 and 36% for GPT-4.**

https://www.johnsnowlabs.com/comparing-spark-nlp-for-healthcare-and-chatgpt-in-extracting-icd10-cm-codes-from-clinical-notes/

Benchmark: Extracting RxNorm Codes

John Snow LABS

The patient is a 40-year-old white male who presents with a chief complaint of "chest pain". The patient is diabetic and has a prior history of coronary artery disease. The patient presents today stating that his chest pain started yesterday evening and has been somewhat intermittent. He has been advised Aspirin (DRUG, 1191, ASPIRIN) 81 milligrams QDay. Humulin N (DRUG, 92880, HUMULIN N) . insulin (DRUG, 5856, INSULIN) 50 units in a.m. Hydrochlorothiazide (DRUG, 5487, HYDROCHLOROTHIAZIDE) 50 mg QDay. Nitroglycerin (DRUG, 4917, NITROGLYCERIN) 1/150 sublingually PRN chest pain.



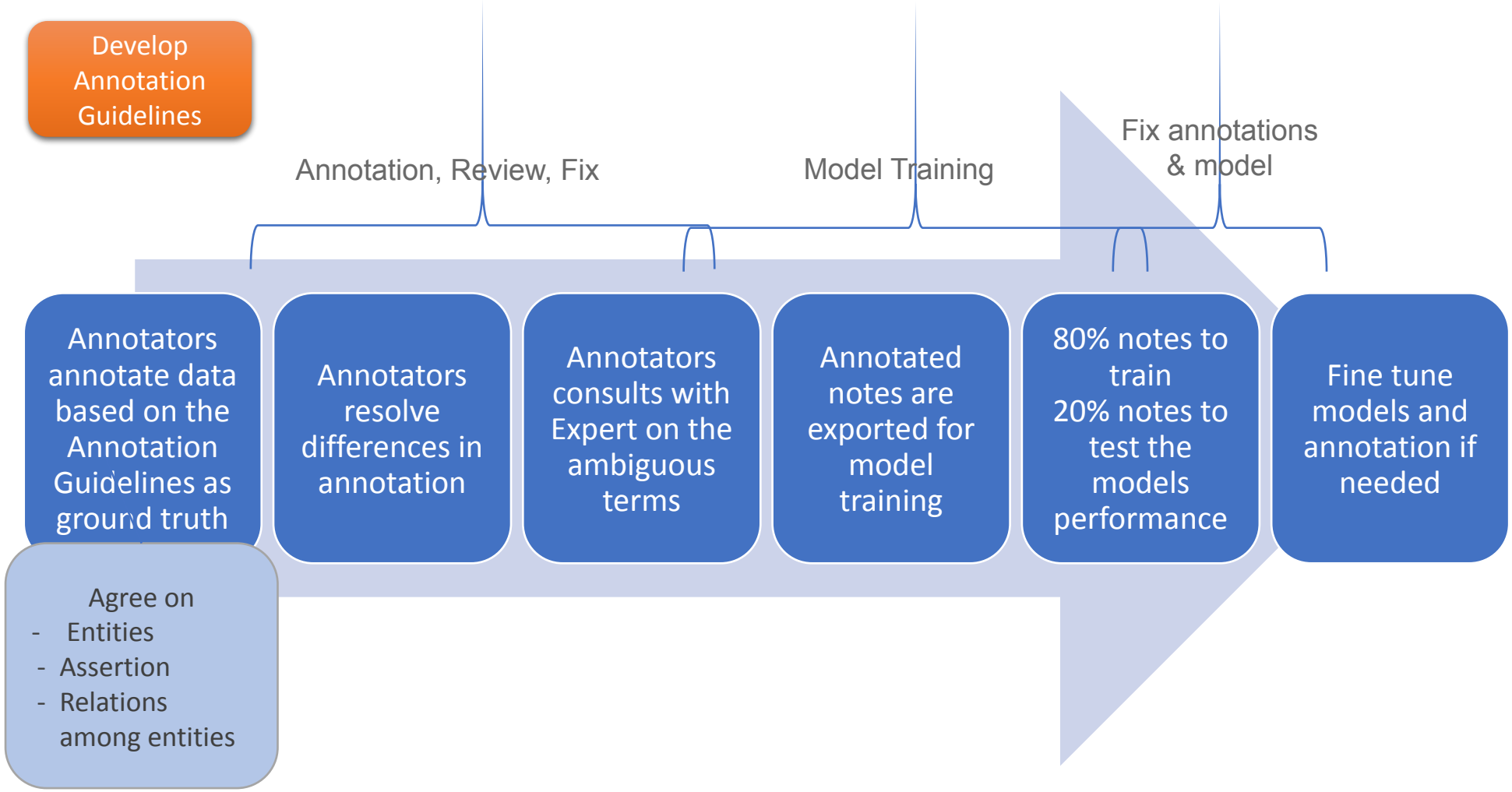
	Top-3 Accuracy	Top-5 Accuracy	Cost
Healthcare NLP	82.7%	84.6%	$4,500
Amazon Comprehend Medical	55.8%	56.2%	$24,250
GPT-4 (Turbo)	8.9%	8.9%	$44,000
GPT-4o	8.9%	8.9%	$22,000

Extracting RxNorm codes is done with a **82.7% success rate vs. 55.8% for Amazon and 8.9% for GPT-4.**

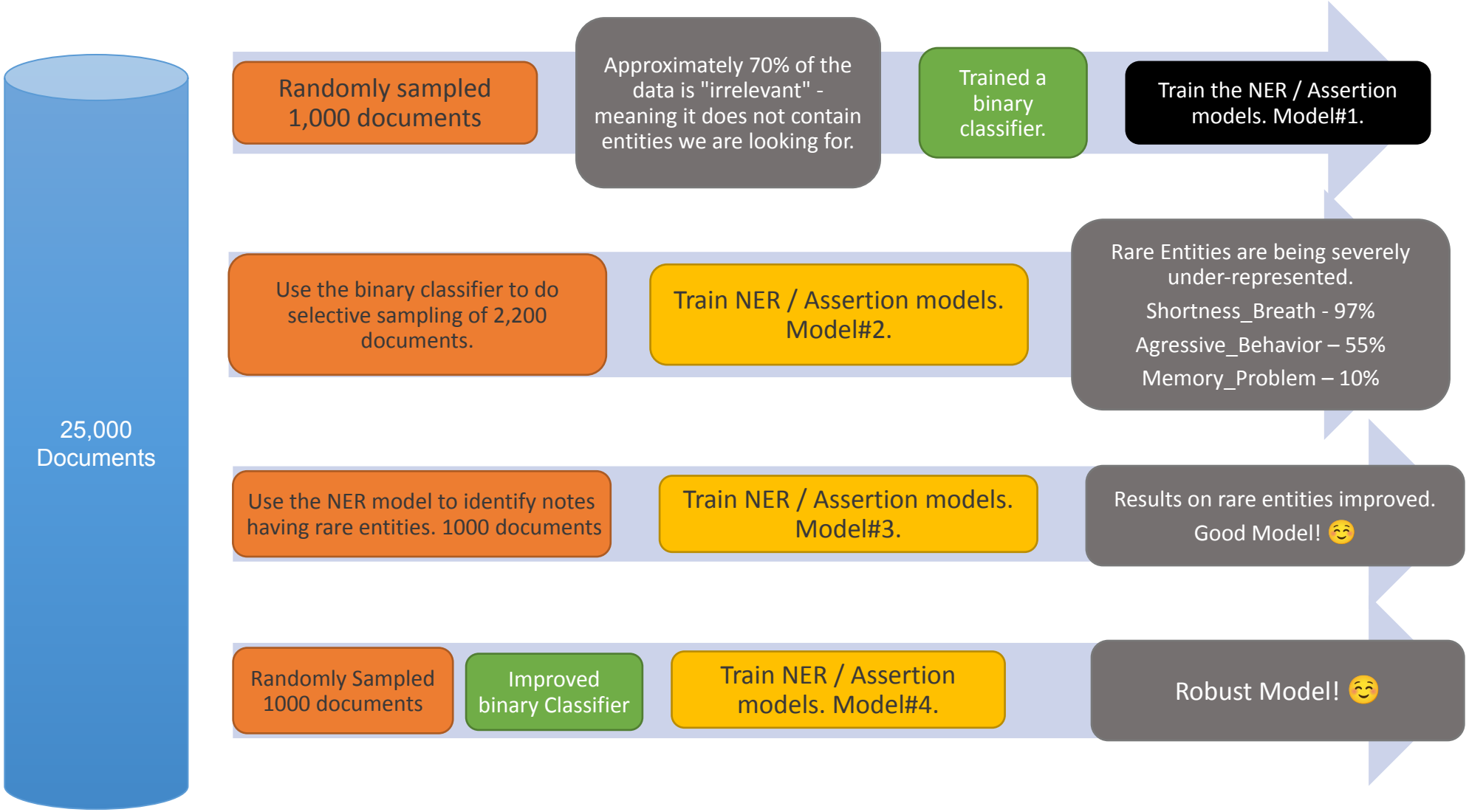
Also **5x times cheaper!**

[RxNorm Code Mapping with NLP: Comparative Analysis between the tools by John Snow Labs, Amazon, and GPT-4]

Model Training Process overview



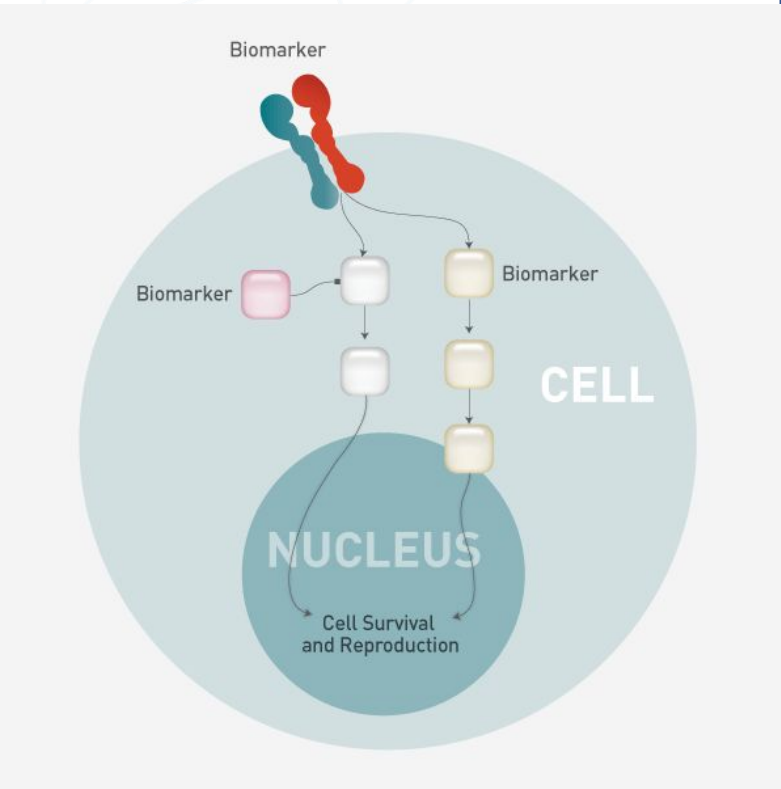
Model Training and Evaluation Process



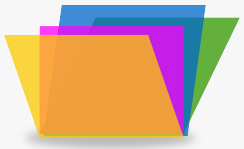
Use Case II - Biomarker Table Extraction

Biomarker and Biomarker Result Table Generation from Oncology Notes

Oncology researchers often need an efficient approach to extract and organize biomarker and biomarker result information from specific sections of oncology notes focused on biomarker analysis. This data is essential for both research and analysis. However, manually retrieving this information from lengthy oncology notes is time-consuming, labor-intensive, and prone to human error.



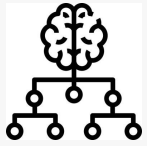
Use Case II - Biomarker Table Extraction



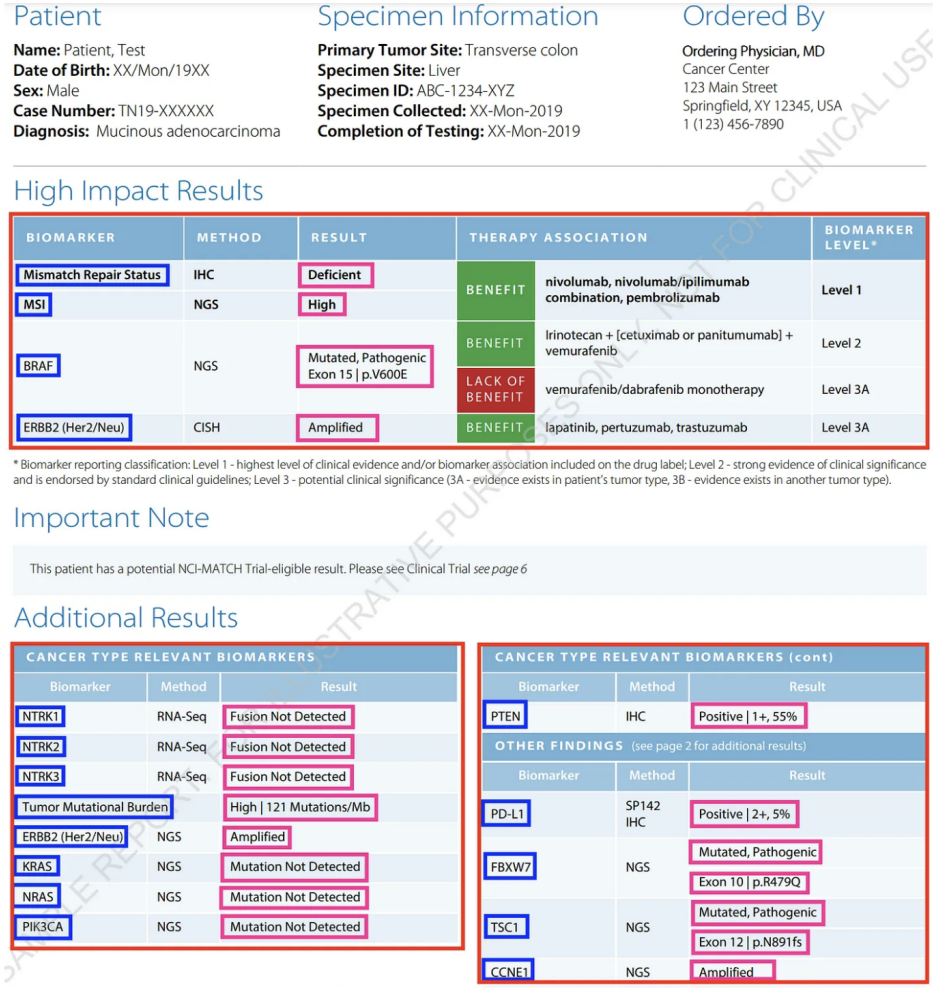
Data

Sequence Classification (Sequences contain Biomarker)

Entity Extraction (Biomarker-Biomarker Result)



Relation Extraction (Biomarker-Biomarker Result)



Patient

Name: Patient, Test
Date of Birth: XX/Mon/19XX
Sex: Male
Case Number: TN19-XXXXXX
Diagnosis: Mucinous adenocarcinoma

Specimen Information

Primary Tumor Site: Transverse colon
Specimen Site: Liver
Specimen ID: ABC-1234-XYZ
Specimen Collected: XX-Mon-2019
Completion of Testing: XX-Mon-2019

Ordered By

Ordering Physician, MD
Cancer Center
123 Main Street
Springfield, XY 12345, USA
1 (123) 456-7890

High Impact Results

BIOMARKER	METHOD	RESULT	THERAPY ASSOCIATION		BIOMARKER LEVEL*
Mismatch Repair Status	IHC	Deficient	BENEFIT	nivolumab, nivolumab/ipilimumab combination, pembrolizumab	Level 1
MSI	NGS	High			
BRAF	NGS	Mutated, Pathogenic Exon 15 | p.V600E	BENEFIT	Irinotecan + [cetuximab or panitumumab] + vemurafenib	Level 2
			LACK OF BENEFIT	vemurafenib/dabrafenib monotherapy	Level 3A
ERBB2 (Her2/Neu)	CISH	Amplified	BENEFIT	lapatinib, pertuzumab, trastuzumab	Level 3A

* Biomarker reporting classification: Level 1 - highest level of clinical evidence and/or biomarker association included on the drug label; Level 2 - strong evidence of clinical significance and is endorsed by standard clinical guidelines; Level 3 - potential clinical significance (3A - evidence exists in patient's tumor type, 3B - evidence exists in another tumor type).

Important Note

This patient has a potential NCI-MATCH Trial-eligible result. Please see Clinical Trial *see page 6*

Additional Results

CANCER TYPE RELEVANT BIOMARKERS

Biomarker	Method	Result
NTRK1	RNA-Seq	Fusion Not Detected
NTRK2	RNA-Seq	Fusion Not Detected
NTRK3	RNA-Seq	Fusion Not Detected
Tumor Mutational Burden		High | 121 Mutations/Mb
ERBB2 (Her2/Neu)	NGS	Amplified
KRAS	NGS	Mutation Not Detected
NRAS	NGS	Mutation Not Detected
PIK3CA	NGS	Mutation Not Detected

CANCER TYPE RELEVANT BIOMARKERS (cont)

Biomarker	Method	Result
PTEN	IHC	Positive | 1+, 55%
OTHER FINDINGS (see page 2 for additional results)		
Biomarker	**Method**	**Result**
PD-L1	SP142 IHC	Positive | 2+, 5%
FBXW7	NGS	Mutated, Pathogenic Exon 10 | p.R479Q
TSC1	NGS	Mutated, Pathogenic Exon 12 | p.N891fs
CCNE1	NGS	Amplified

relations	entity1_begin	entity1_end	chunk1	entity1	entity2_begin	entity2_end	chunk2	entity2
is_finding_of	968	975	negative	Biomarker_Result	981	1010	thyroid transcription factor-1	Biomarker
is_finding_of	968	975	negative	Biomarker_Result	1016	1023	napsin A	Biomarker
is_finding_of	1040	1047	positive	Biomarker_Result	1053	1054	ER	Biomarker
is_finding_of	1040	1047	positive	Biomarker_Result	1060	1061	PR	Biomarker
is_finding_of	1068	1075	negative	Biomarker_Result	1081	1084	HER2	Biomarker

Let’s code!