LLM-Assisted Annotation and Biomarker Feature Extraction to Support Future Multimodal Models Predicting Lung Cancer Immunotherapy Resistance.

Eric Guan[1], Nitya Lagadapati[1], Sean Li[1], Weiyao (Stella) Tao[1], Marcus Torres-Negron[1], Isha Yeleswarapu[1], Amanda Kwok[1], Sidharth Raghavan[1], Iyinoluwa Tugbobo[1], Beiya Xu[1], Joseph L. Greenstein, PhD[1], Casey Overby Taylor, PhD[1], Siamak Ardekani, MD PhD[1], Joseph Murray, MD PhD[2,3]

[1]Department of Biomedical Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA
[2]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3]Caris Life Sciences, Phoenix, AZ, USA

Team Axolotl

Introduction

- 1 in 5 cancer deaths in the US are due to lung cancer. [1]
- Non-small cell lung cancer (NSCLC) accounts for 87% of lung cancers. [2]
- While certain immune checkpoint inhibitors have proved useful, around 64% of patients experiences resistance to these treatments when they are used as a second-line treatment. [3]
- Many predictors like comorbidities or biometrics are insufficient alone.
- Radiographic images and imaging reports, and results from blood tests like comprehensive metabolic panel (CMP) and cell blood count with differential (CBC with diff) have information rich data that could be useful in machine learning models but contain excessive noise in raw form.

Objectives

- **Objective 1:** Develop and evaluate an LLM-based pipeline to identify NSCLC metastatic features by optimizing prompting strategies and compare model outputs to expert manual annotations.
- **Objective 2:** Develop and evaluate a machine learning model using laboratory-derived biomarker features to predict immunotherapy response in NSCLC patients.

Data Overview

- **Objective 1:** 360 radiographic reports (text-based) without the impressions. Ground-truth JSON format interpretations for all 360 reports were curated by an expert thoracic oncologist. (Fig. 1)
- **Objective 2:** Longitudinal CBC with differential (n = 3201; 117 patients) and CMP (n = 4513; 124 patients) records. (Fig. 2)
- **Overlap:** Shared cases (n=98) across datasets enables cross-comparison.

Figure 1

CT scan of the chest WITHOUT intravenous contrast, using low dose lung cancer screening protocol.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs and Airways: The lungs and airways are normal with no focal abnormality demonstrated.

Pleura: The pleural spaces are clear.

Base of neck, mediastinum and heart: The thyroid gland is normal. No mediastinal, hilar or axillary lymphadenopathy is seen. The heart and pericardium are within normal limits.

Soft tissues: Normal.

Abdomen: This study was performed without contrast and with lower than standard dose. These factors reduce the sensitivity for detection of small lesions in the upper abdomen. Given these technical limitations, no focal lesion is seen within the visualized liver, spleen, pancreas, kidneys and adrenal glands.

Bones: The visualized bony thorax is within normal limits.

IMPRESSION: No pulmonary nodules are identified.

Lung-RADS Category: 1 - Negative
Explanation of the Lung-RADS categories can be found at: http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/LungRADS/AssessmentCategories

RECOMMENDATION: Continue annual screening with low-dose Chest CT in 12 months

Figure 2

	CMP	CBC w/ Diff
Total Parameters	32	47
Total Unique Patients	124	117
Baseline records	378	344
On-treatment records	4135	2856
Average record / patient	36.4	27.4
Parameters with values acquired for all records	4	8
Parameters with values acquired for 50% of records	20	22

Conclusions

- Increased variability at higher temperatures does not necessarily translate to reduced semantic accuracy.
- Large language models like GPT, LLaMA, and Claude have high potential to assist in the structured extraction of data from imaging report texts with similar levels of inaccuracies across tested parameters, and BERTScores in the mid-to-upper range of semantic similarity.
- Unsupervised clustering of longitudinal CBC and CMP data identified clinically meaningful patient subgroups with distinct immunotherapy response profiles, with baseline liver function markers and on-treatment neutrophil dynamics emerging as the strongest cluster-defining features, supporting the feasibility of blood-test-based resistance prediction in NSCLC and warranting prospective validation.
- Future work will train and cross-validate an XGBoost classifier on the identified candidate features.

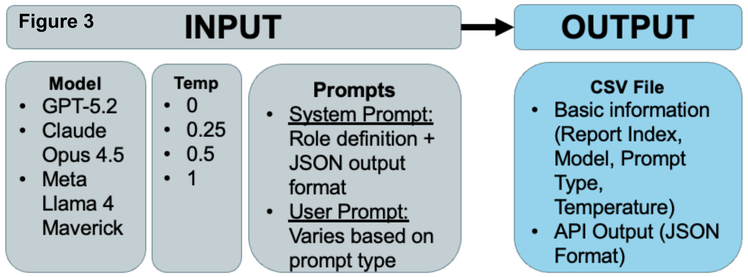
Overall, these results show that routinely collected clinical data can be leveraged to enable scalable, data-driven prediction of treatment outcomes for those with NSCLC.

[1] American Cancer Society. Lung cancer statistics. 2025. [2] American Cancer Society. What is lung cancer? 2025.
[3] Hiltbrunner S, et al. Nat Commun. 2023;14:5154. [4] Dietrich N, et al. AJR Am J Roentgenol. 2025;225:e2532956

Methods

LLM-Assisted Annotation

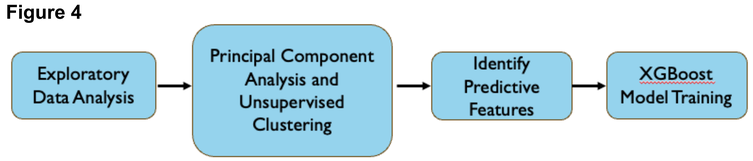
Figure 3



- Tested all combinations of input parameters; prompts included zero-, one-, and few-shot (varying #examples) and chain-of-thought (stepwise extraction). [4]
- Primary disease site and metastatic status, sites, and stages were extracted.
- BERTScore, a pretrained model that compares text using context, was used to compare API output of 45 reports to ground truth classifications; scores range from 0 to 1, with higher values indicating greater semantic similarity.

Biomarker Feature Extraction

Figure 4

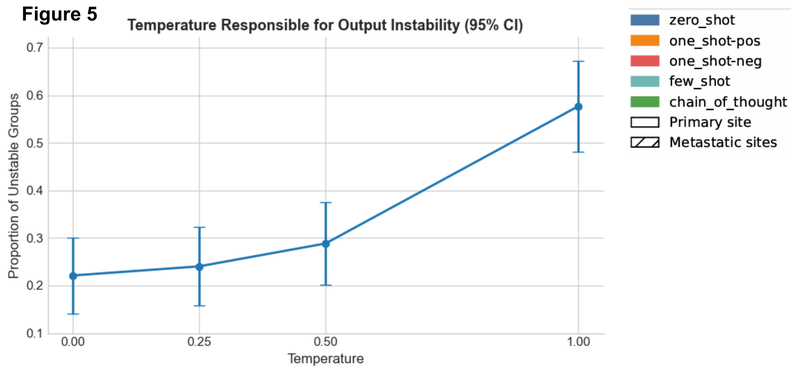


- Missingness, parameter distributions, and baseline vs. on-treatment shifts were assessed. Durable clinical benefit (DCB), a binary variable based on whether the patient is alive and cancer progression free at 6 months, was defined as the outcome variable.
- Principal Component Analysis and unsupervised clustering were performed on features across CBC, CMP and Clinical data to identify patient subgroups.
- The derived features and clusters were then evaluated for associations with DCB to identify clinically relevant predictors of immunotherapy response.

Results

LLM-Assisted Annotated Dataset

Figure 5



- 26% of prompt-model type groups produced multiple unique outputs when only temperature varied. Out of these groups, higher temperature (T=1) led to greater deviation from the group consensus. (Fig. 5)
- BERTScore evaluations for non-binary extractions (primary disease sites & metastatic disease sites) showed no statistically significant differences in performance across temperatures, models, and prompt-types. (Fig. 6/7)
- Across all models, prompt types, and temperatures, mean BERTScore F1 was 0.77 for primary disease site and 0.65 for metastatic disease sites, indicating moderate-to-high semantic similarity between LLM outputs and expert-annotated ground truth. (Fig. 6/7)

Figure 6

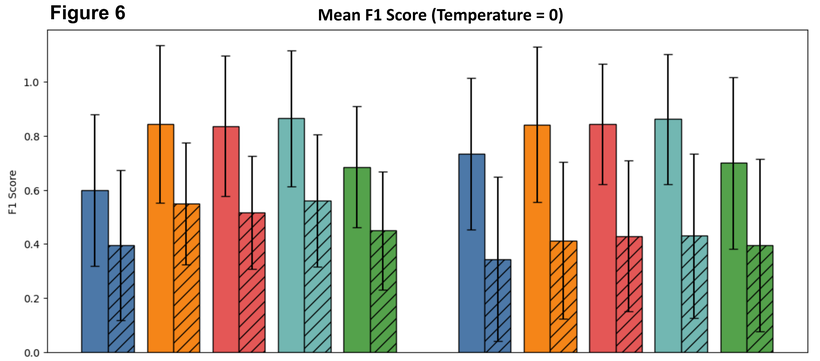
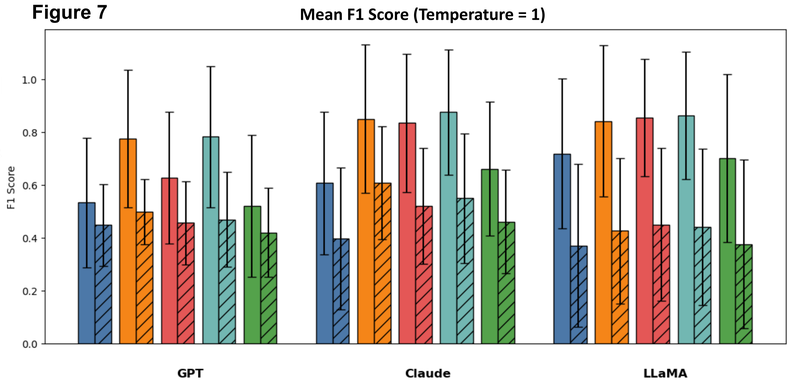


Figure 7



Biomarker Identification and Clusters

Figure 8

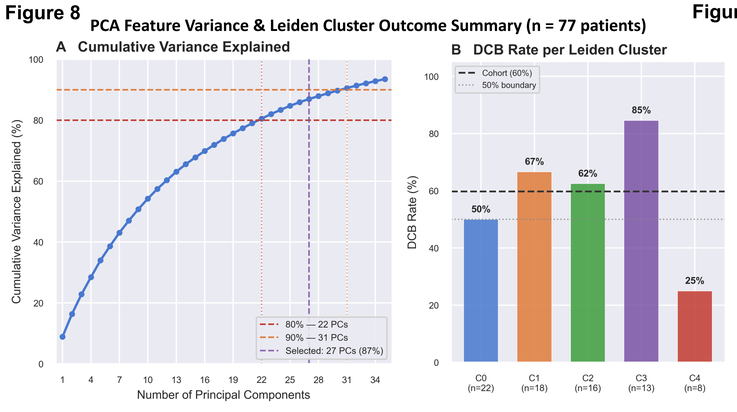
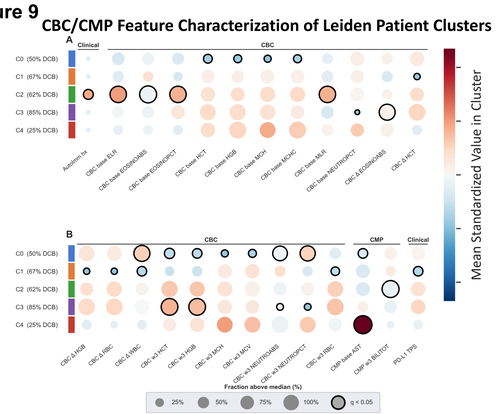


Figure 9



- PCA and unsupervised Leiden clustering of 77 NSCLC patients (128 CBC/CMP/clinical features) identified 5 subgroups with DCB rates spanning 25%–85%, demonstrating meaningful immunotherapy response stratification. (Fig. 8)
- The highest-DCB cluster (85%, n=13) showed significantly lower on-treatment neutrophil counts (q<0.05), consistent with favorable immunotherapy response biomarkers; the lowest-DCB cluster (25%, n=8) showed elevated baseline AST (r=0.926, q=0.002), implicating hepatic status as resistance predictor. (Fig. 9)