# Comparing LLM-based to Expert-Curated Extraction for Biomarker Attributes in Lung and Breast Cancer

3036

Sheenu Chandwani, MPH, PhD[1] | Payal Keswarpu, MBBS, MD[2] | Ashwani Yadav, MSc[2] | Vivek Prabhakar Vaidya, BSc[2] | Jiby Joseph, MD, MS[1] | Ruth Pe Benito, MPH[1]
[1]ConcertAI, LLC, Cambridge, MA, USA, [2]ConcertAI, LLC, Bengaluru, India.

ConcertAI

## BACKGROUND & OBJECTIVE

Biomarker data is essential for real-world evidence (RWE) and health economics and outcomes research (HEOR) in oncology. However, biomarker information is frequently embedded within unstructured pathology reports, clinical narratives, and physician notes, creating substantial challenges for scalable and standardized data capture.

- Biomarker documentation is heterogeneous across tumor-types, assays and clinical workflows.
- Key attributes (e.g., name, result category, variant type, genomic alteration, exon number) are often inconsistently represented
- Manual abstraction is resource-intensive and difficult to scale for large RWE/HEOR studies

Study Objective: To evaluate the performance of a multi-agent LLM/SLM-based pipeline for automated extraction of oncology biomarker attributes from unstructured EHR data, benchmarked against expert-curated reference annotations.

## METHODS

**Data Source**
EHR notes from the US-representative ConcertAI network were accessed for patients with lung cancer and breast cancer.

**AI Extraction Framework**
A multi-agent suite of decoder-only LLM and SLM models was developed and trained using the following approaches:

**1. Biomarker Named Entity Recognition**
Fine-tuned small language models (SLMs) identify biomarker entities in clinical text.

**2. Contextual Attribute Extraction**
Fine-tuned large language models (LLMs) extract multimodal attributes and classify assertion state e.g., positive, negative, pending, hypothetical.

**3. Multi-Agent Integration**
Multiple agents incorporate longitudinal clinical context to interpret complex statements across large note contexts.

**Attributes Extracted**
The following attributes were extracted and standardized:
- **Biomarker Name**: The specific biomarker tested (e.g., EGFR, KRAS, ER, PD-L1)
- **Categorical Result**: Positive, negative, or indeterminate result
- **Exon Number**: Specific exon location of the alteration (where applicable)
- **Genomic Alteration**: Nature of the genetic change (e.g., mutation, expression, deletion, amplification, gene rearrangement, copy number variation, single nucleotide variation)
- **Variant Type**: Amino acid-level alteration descriptor (e.g., V600, G12C, T790m)

**Biomarkers Evaluated**
Model outputs were validated across 11 biomarkers:
- **Lung cancer biomarkers**: EGFR, ALK, KRAS, ROS1, TMB, PD-L1
- **Breast cancer biomarkers**: ER, PR, HER2, ESR1, PIK3CA

**Validation Approach**
Performance was evaluated at the biomarker-record and –patient level against expert-curated reference annotations. Precision, recall and F1-score were calculated independently for each extracted attribute. Attribute matching required agreement on biomarker identity and corresponding extracted value.

## RESULTS

**Table 1. Real-World EHR Case Example – How Biomarker Information Appears in Practice**

| | EHR Excerpt (De-Identified) | Challenge |
|---|---|---|
| 01 | ***"The Guardant 360 test identified two patient mutations: ESR1 and PIK3CA"*** | Entity disambiguation – multiple biomarkers in one sentence, each needing independent extraction |
| 02 | ***"Testing for HER2/neu, BRCA1/2, PIK3CA, and ESR1 negative"*** | Negation scope – does "negative" apply to all four biomarkers or only ESR1 |
| 03 | ***"Demonstrates ESR1 mutation. Recommending elacestrant, approved for ESR1-mutated metastatic breast cancer"*** | Patient Finding vs infotext – first ESR1 (green) is the patient's result; second (red) is drug approval background |
| 04 | ***"Mutations in ESR1 possibly also point towards this resistance"*** | Assertion state – speculative phrasing ("possibly" must not be treated as a confirmed positive result |

**Table 2. F1 score comparing AI model to expert extracted biomarker attributes in lung and breast cancer at the segment level**

| Biomarker | Biomarker Name | Categorical Result | Variant Type | Genomic Alteration | Exon Number |
|---|---|---|---|---|---|
| EGFR | 0.99 | 0.92 | 0.93 | 0.81 | 0.90 |
| ER | 0.99 | 0.97 | — | NA | NA |
| PR | 0.96 | 0.94 | — | NA | NA |
| HER2 | 0.98 | 0.95 | — | NA | NA |
| ALK | 0.99 | 0.95 | 0.62 | NA | NA |
| KRAS | 1.00 | 0.62 | 0.97 | 1.00 | NA |
| ROS1 | 1.00 | 1.00 | 0.57 | NA | NA |
| PD-L1 | 0.99 | 0.84 | 0.88 | NA | NA |
| ESR1 | 1.00 | 0.88 | 0.80 | NA | NA |
| PIK3CA | 1.00 | 0.97 | 0.89 | NA | NA |
| TMB | 1.00 | 0.86 | — | NA | NA |

**Table 3. F1 score comparing AI model to expert extracted biomarker attributes in lung and breast cancer aggregated at the patient level**

| Biomarker | Biomarker Name | Categorical Result | Variant Type | Genomic Alteration | Exon Number |
|---|---|---|---|---|---|
| EGFR | 0.92 | 0.94 | 0.93 | 1 | 1 |
| ER | 0.99 | 0.96 | NA | NA | NA |
| PR | 0.95 | 0.97 | NA | NA | NA |
| HER2 | 0.97 | 0.90 | NA | NA | NA |
| ALK | 1.00 | 1.00 | 1.00 | NA | NA |
| KRAS | 1.00 | 0.67 | 1.00 | 1 | NA |
| ROS1 | 1.00 | 1.00 | 1.00 | NA | NA |
| PD-L1 | 1.00 | 0.92 | 1.00 | NA | NA |
| ESR1 | 1.00 | 0.86 | 1.00 | NA | NA |
| PIK3CA | 1.00 | 1.00 | 0.93 | NA | NA |
| TMB | 1.00 | 0.82 | NA | NA | NA |

## RESULTS

**Table 5. Clinical AI Challenge and Solution Approach**

| Clinical Scenario | Example from EHR | What AI Must Recognize | AI Approach Required |
|---|---|---|---|
| Negative Result | ***"Testing showed ESR1 negative"*** | ESR1 was tested and the result was negative. | **Negative Detection** |
| Indirect Positive | ***"Guardant revealed an ESR1 mutation"*** | A positive ESR1 alteration is implied through physician note shorthand. | **Inferential Reasoning** |
| Treatment Linked | ***"ESR1 mutated. Recommending elacestrant"*** | The biomarker result is positive and reinforced by a related therapy recommendation. | **Contextual Inference** |
| Third-Party Result | ***"Son is PTEN positive"*** | PTEN positivity applies to a family member, not the patient. | **Patient Identification** |
| Pending/ Hypothetical | ***"Waiting for Guardant ESR1"*** | ESR1 testing is planned or pending; no result has been established. | **Assertion Status Handling** |

## KEY TAKEAWAYS

Multi-agent LLM/SLM pipelines achieved high accuracy for scalable extraction of oncology biomarker attributes from unstructured EHR documents, with strongest performance for core biomarker identification (F1 up to 1.00) and more variable performance for complex genomic contextualization tasks.

## CONCLUSIONS

- Multi-agent decoder-only models demonstrated strong performance for automated extraction of multidimensional oncology biomarker attributes from unstructured EHR.
- Biomarker name (F1>=0.92) showed the highest accuracy demonstrating robust core entity recognition, and achieved at least an 82% F1 score across majority other attributes across all biomarkers tested.
- Complex attributes showed more variable performance, notably KRAS categorical result. This reflects a real-world documentation pattern in which KRAS positivity is implied by variation notation — e.g., "KRAS G12C" — rather than stated explicitly as a positive result. While a clinician interprets this as unambiguously positive, the model must learn to infer categorical status from implicit evidence.
- These findings highlight the potential of AI-driven auto-curation to scale biomarker characterization for RWE/HEOR while significantly reducing manual review burden.
- Future work will include broader validation at the patient and population level and correlation with clinical expectations in the patient treatment journey.
- Model improvements will target implicit positivity recognition for mutation-variant–based biomarkers (e.g., KRAS G12C) where variant type extraction can serve as a high-confidence proxy for categorical result derivation.

## REFERENCES

1. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. Proc EMNLP. 2022.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–180.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29:1930–1940.