Evaluation of large language model (LLM)-based clinical abstraction of Electronic Health Records (EHRs) for Non-Small Cell Lung Cancer (NSCLC) patients

Kabir Manghnani[1], Katie Mo[1], Kunal Nagpal[1], Xifeng Wang[1], Kaitlynn Cunnea[1], Bridget Bax[1], Michael Bodker[1], Arpita Saha[1], Chelsea Osterman[1], Riccardo Miotto[1], Chithra Sangli[1]
[1]Tempus AI, Inc., Chicago, IL

Abstract Presentation #11160

INTRODUCTION

Abstraction is a critical step for converting clinical data from unstructured EHRs into a structured format suitable for real-world data analyses. Typically this is a manual, labor-intensive activity requiring substantial training. While prior work has shown that abstraction by humans is reliable (Mo et al.), advances in LLMs may improve the efficiency of abstraction. We aim to measure the performance of LLMs in abstracting a diverse set of oncology data elements.

SUMMARY

- A cost-efficient LLM system was utilized for clinical data abstraction. **Queries for each field were iteratively refined** on a development cohort.
- **LLM-based abstraction shows high agreement** with human abstractors across a variety of critical abstraction fields on an evaluation cohort.
- The use of LLMs **may significantly reduce the burden of human abstraction** and allow for large-scale curation of oncology records.
- Challenges in handling nuanced contexts underscore the need for **careful refinement and evaluation prior to widespread use**.

METHODS

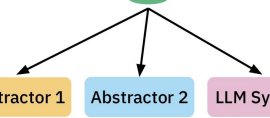
Overview

Cohort Selection (N=222 patients)

- Sequenced by Tempus
- NSCLC
- Advanced or metastatic diagnosis between Jan 2018-Oct 2020
- Received treatment within 90 days of advanced or metastatic diagnosis
- Age at advanced or metastatic diagnosis ≥18 years old

Unstructured Patient Records

- n=248 pages per case (mean)



Abstraction

- Abstractors were blinded to study participation

LLM System

A two-stage LLM system balancing cost and comprehensiveness was used to abstract clinical elements for demographics, diagnosis, third-party lab biomarker testing, and first line (1L) treatment.

Stage 1: Cost-efficient Abstraction via Retrieval-Augmented Generation (RAG): This stage first extracted 16 snippets of 512 characters each from the patient record. The snippets were chosen based on semantic similarity to the abstraction query via an open-source sentence transformer model (bge-small-v1.5). These snippets were then combined with the abstraction query and input into the GPT-4o LLM to generate the abstracted value.

Stage 2: (As necessary) Comprehensive Record Processing via Long-Context LLM: Stage 2 was initiated if the Stage 1 process failed to yield an answer (when GPT-4o signaled its inability to respond based on the provided snippets). In these cases, the entire patient record was concatenated into a single text input. This input, along with the abstraction query, was input into a long-context LLM (Gemini-Pro-1.5) to re-attempt the abstraction.

Query Development and Validation: Abstraction queries specific to each data field were iteratively refined on a dedicated development cohort (n=90 patients). The development cohort was distinct from the evaluation cohort used to generate the performance results reported.

Reliability Analysis

Gwet's agreement coefficient (AC) was the primary measure of agreement between the LLM and each abstractor.

RESULTS

Figure 1. LLM and Abstractor Agreement using Gwet's Agreement Coefficient (AC)
Agreement was calculated when both the LLM and abstractor provided non-null values.

	Abstractor A vs LLM			Abstractor B vs LLM		
	N	AC	95% CI	N	AC	95% CI
Demographic						
Birth date, within ±30 days	216	1	(1,1)	216	1	(1,1)
Sex	205	0.97	(0.94,1)	205	0.96	(0.92,1)
Race	124	0.97	(0.94,1)	122	0.98	(0.96,1)
Smoking status	190	0.98	(0.96,1)	191	0.98	(0.95,1)
Diagnosis						
Stage at primary diagnosis	194	0.92	(0.88,0.96)	185	0.93	(0.89,0.97)
Stage at advanced diagnosis	203	0.95	(0.92,0.99)	199	0.96	(0.94,0.99)
Histology	208	0.98	(0.96,1)	208	0.97	(0.94,0.99)
Year of advanced diagnosis	181	0.95	(0.92,0.98)	177	0.93	(0.89,0.97)
Third Party Biomarker						
EGFR mutation	79	1	(1,1)	70	1	(1,1)
ALK fusion	69	0.98	(0.95,1)	62	1	(1,1)
ROS1 mutation	64	1	(1,1)	58	1	(1,1)
PD-L1 status	114	0.92	(0.85,0.99)	109	0.95	(0.89,1)
BRAF mutation	45	1	(1,1)	40	1	(1,1)
RET mutation	20	1	(1,1)	17	0.87	(0.66,1)
NTRK1 fusion	4	1	(1,1)	3	1	(1,1)
NTRK2 fusion	1	1	(1,1)	0	-	-
NTRK3 fusion	3	1	(1,1)	1	1	(1,1)
First Line Treatment						
Anti-cancer Agents	184	0.86	(0.80,0.91)	174	0.85	(0.80,0.91)
Initiation date, within ±30 days	191	0.84	(0.78,0.91)	179	0.81	(0.74,0.88)

Key: AC 0.9-1 (Excellent) | AC 0.8-0.89 (Very Good)

Figure 1. The LLM demonstrated high agreement with each abstractor (≥0.81 across all categories). Agreement was highest in demographic and diagnosis domains and lower for 1L treatment domain, which require deeper understanding of a patient's temporal journey.

Figure 2. LLM Completeness
Percent of cases when LLM yields a non-null value when both abstractors provided a non-null value

Domain	Weighted average
Demographic	91.8%
Diagnosis	92.5%
Third Party Biomarker	84.0%
First Line Treatment	88.8%
Overall	90.3%

Figure 2. The LLM system overall yielded abstracted values for 90.3% of elements where both abstractors provided non-missing values.

Figure 3. LLM Predicts More
Percent of cases when LLM yields a non-null value when both abstractors provided null values

Domain	Weighted average
Demographic	7.5%
Diagnosis	2.8%
Third Party Biomarker	38.5%
First Line Treatment	2.9%
Overall	31.4%

Figure 3. When neither abstractor provided values, the LLM sometimes provided outputs. Discrepancies were primarily driven by nuances in abstraction rules. The LLM often included Tempus-tested biomarkers, while abstractors were more rigorous in abstracting only third-party biomarker results.

REFERENCES

- Mo, K., Wang, X., Cunnea, K., Bax, B., Berezina, M., Osterman, C., Miotto, R., and Sangli, C. (2025). "Assessing the reliability, accuracy, and utility of clinical abstraction methods from unstructured Electronic Health Records (EHRs)" ASCO Annual Meeting, Chicago, IL, 2025. Abstract e23311.

ACKNOWLEDGMENTS

We thank Amrita A. Iyer, Ph.D from the Tempus Science Communications team for poster review.

Correspondence: chithra.sangli@tempus.com