Low-cost Feature Extraction from Clinical Notes Using Hummingbird: ECOG Performance Status of Lung Cancer Patients

Ishtiyaque Ahmad[1], Vivek Rudrapatna[2], Trinabh Gupta[1]

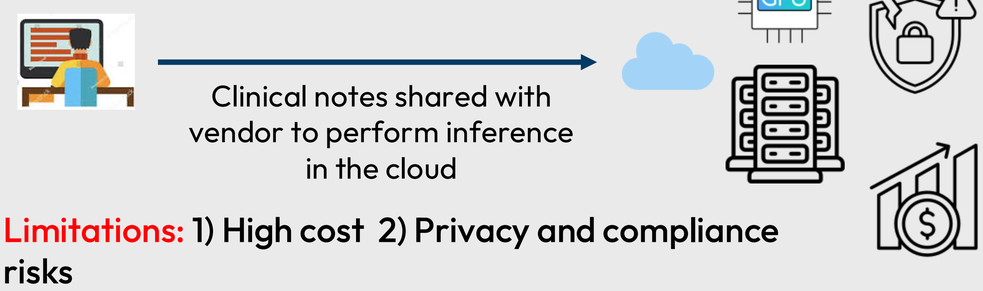
[1]DataUnite, Inc., Cupertino, CA, USA; [2]University of California San Francisco, CA, USA

Background

- Extracting structured information from clinical notes remains challenging.
- Recent LLM advancements improve extraction quality, but lead to:
  1. **Privacy & compliance risks:** due to sending data to service providers (OpenAI, etc.)
  2. **High cost:** due to GPU infrastructure or expensive API calls

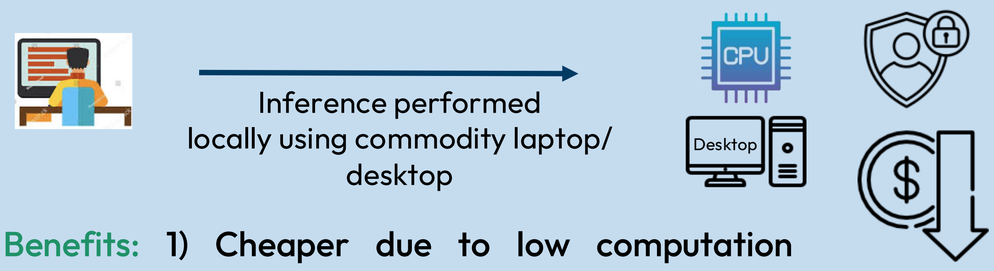
Hummingbird, a lightweight CPU-based NLP framework, runs on commodity hardware within hospital firewalls, offering a scalable and privacy-preserving alternative.

LLM Approach



Limitations: 1) High cost 2) Privacy and compliance risks

This work (Hummingbird)



Benefits: 1) Cheaper due to low computation involved 2) secure as no data needs to be sent to the cloud

Objective

Evaluate Hummingbird, a lightweight CPU-based NLP framework, for extracting Eastern Cooperative Oncology Group (ECOG) performance status from oncology notes, enabling large-scale characterization of functional status and trial-relevant populations in real-world lung cancer care, and compare its cost with GPT-5.4-mini.

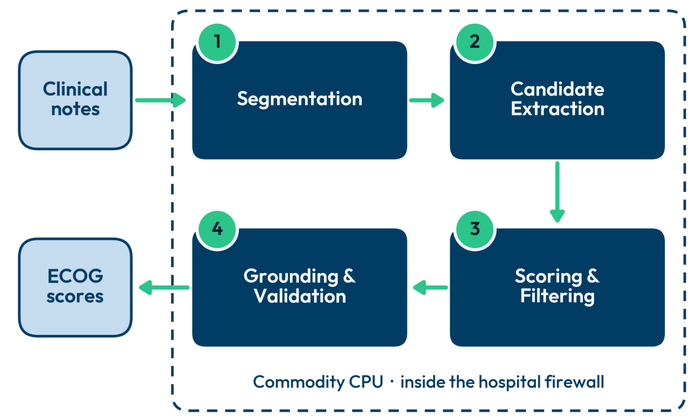
Methods

- **Study Cohort:** Adult patients (≥18 years) with a first primary lung cancer diagnosis (ICD-10-CM C34.x; excluding C78.0x) between January 1, 2020 and January 31, 2026, identified from the UCSF deidentified EHR.
- **Notes:** Lung oncology notes containing the pattern “ECOG”.
- **Model Approach:** Hummingbird was deployed on a CPU-only infrastructure. ECOG extraction was formulated as a multi-class classification task (ECOG 0 to 4, and no documented ECOG).
- **Performance Evaluation:** 200 manually annotated notes served as the gold standard.
- **Cost Analysis:** Extraction costs were estimated by projecting GPT-5.4-mini API token usage and by measuring Hummingbird CPU utilization and projecting it to equivalent AWS CPU costs.



Hummingbird architecture

Consists of 4 stages, each of which runs on a commodity CPU machine with a small amount of memory.



Results

Extraction Performance: Hummingbird was evaluated against manual review on 200 annotated notes sampled from the cohort.

Performance Metric	Value
Accuracy	98.5%
Weighted F1-score	98.6%

Cost Analysis: GPT-5.4-mini cost projected from API token usage; Hummingbird cost projected from measured CPU utilization to AWS CPU pricing.

Model	Extraction cost (full cohort)	Relative cost
GPT-5.4-mini	$119.30	55x
Hummingbird (this work)	**$2.16**	**1x**

Real-World ECOG Characterization

Applied to the full cohort of 5,923 patients contributing 39,765 lung oncology notes containing the pattern “ECOG”:

- 51% of patients had at least one documented ECOG assessment; among them, 79.9% had a baseline ECOG of 0 to 1 and 20.1% had ECOG ≥2.
- 2,420 patients had at least one follow-up ECOG assessment, with a median of 7 assessments (IQR 3 to 16) over a median follow-up of 8.6 months.
- Among patients with serial assessments, 44.4% experienced worsening of ECOG score.

Conclusion and next steps

Hummingbird extracted ECOG performance status from notes with high accuracy while enabling large-scale characterization of functional status in a real-world lung cancer population. With 55x lower cost than frontier-model APIs, Hummingbird provides a scalable, privacy-preserving approach for generating clinically meaningful real-world evidence.

Future work: Validate this approach across multiple clinical variables and institutions to assess generalizability.

Contact: Ishtiyaque Ahmad, ishtiyaque@dataunite.co

DataUnite