Developing an EHR-Based Multi Feature Machine Learning Model to Identify Lung Cancer Subtype

MSR69

Authors: S. CHANDWANI[1], V. PRIYA[2], V. VAIDYA[2]

[1]ConcertAI, LLC, Cambridge, MA, USA, [2]ConcertAI, LLC, Bengaluru, India

INTRODUCTION

- Delineating lung cancer subtypes into NSCLC vs SCLC is critical for clinical research due to their distinct biology, therapeutic options, and biomarker profiles.
- Real world data (RWD) sources, such as EHRs, claims, and registries have well documented challenges in differentiating NSCLC vs SCLC as this information is often buried in unstructured EHR notes, has coding insufficiencies, or there is a lag in recency.
- Relying on expert determined pathologic evidence of subtype from EHRs is a resource intensive approach that can put high operational constraints in generating scalable datasets.

OBJECTIVE

The goal of this research was to establish a scalable ML-based model that leverages multiple clinical features from the EHR to identify lung cancer subtype.

METHOD

- **Data source**: EHR notes from the US representative ConcertAI network were accessed for patients with a C34 code for lung cancer.
- **Training of the ML algorithm**:
  - A randomly selected training dataset comprising of 7,914 lung cancer patients was established.
  - Snippets from EHR notes of the training set were labelled into NSCLC or SCLC based on exact tumor name or synonyms alongside supporting evidence such as stage mentions (extensive, limited for SCLC), or histology (eg: adenocarcinoma for NSCLC).
- **Modelling**:
  - Abstraction: Small Language Models (SLMs) are used to extract direct and indirect evidence of lung cancer subtype.
  - Assertion: Large Language Models (LLMs) are used to semantically and temporally assert and filter evidence.
  - Integration:
    - ML model is applied at patient level to integrate evidence and resolve contradictions; if unresolved, no prediction is made.
    - XGBoost1 (eXtreme Gradient Boosting) is utilized as it naturally handles data missingness and sparsity.
    - It applies Sparsity- Aware Split Finding: Optimal default direction found by trying both directions in a split and choosing the one which proposes a maximum gain.
- **Validation:**
  - Performance metrics were generated by comparing a randomly selected sample of 50 patients predicted as NSCLC and SCLC each (validation set) to expert-determined subtype from the EHR.
  - Clinical relevance was tested by comparing first-line systemic anti-cancer treatment and key biomarker status distribution in AI-curated and expert-curated datasets.

RESULTS

- For the training set, the model predicted 67.5% (5,346) NSCLC, 17.7% (1,398) SCLC, and 14.8% (1,170) other patients.
- The test set comprised of 102,029 patients, and model predicted 84.4% (86,151) NSCLC, 12.6% (12,862) SCLC, and 3% (3,016) other.

Figure 1. Lung Cancer Subtype Predictive Modelling Approach from Oncology EHR

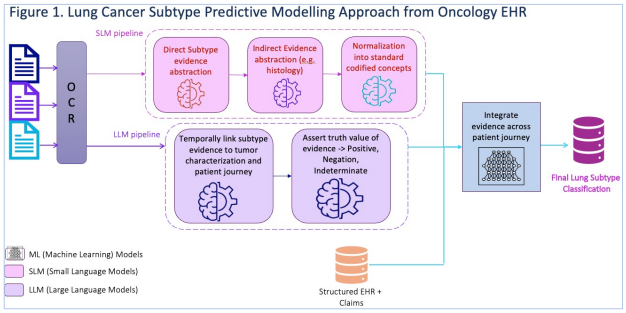

Figure 2. Key **Biomarker** Status for SCLC and NSCLC Predicted Subtypes

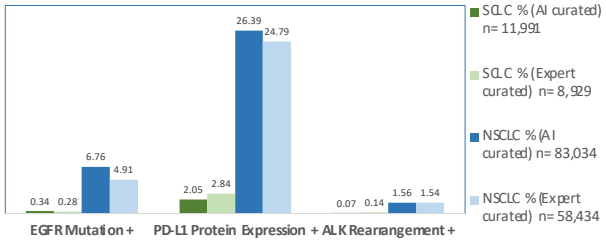

Table 1. Performance of ML-Based Lung Cancer Subtype Compared to Expert Determination

Lung Subtype	Precision	Recall	Specificity
SCLC	0.92	0.92	0.96
NSCLC	0.96	0.87	0.93

Table 2a. Systemic Anti-Cancer Regimen Distribution in First Line for Advanced SCLC Subtype Predicted vs. Expert Curated

Predicted SCLC, Line 1, n= 5,880		Expert curated SCLC, Line 1, n= 3,991	
Regimen name (Top 5)	**%**	**Regimen name (Top 5)**	**%**
carboplatin/cisplatin + etoposide	37.4%	atezolizumab + carboplatin/cisplatin + etoposide	44.2%
atezolizumab + carboplatin/cisplatin + etoposide	33.9%	carboplatin/cisplatin + etoposide	33.7%
carboplatin/cisplatin + durvalumab + etoposide	4.5%	carboplatin/cisplatin + durvalumab + etoposide	6.3%
atezolizumab + carboplatin/cisplatin + etoposide+ lurbinectedin	1.5%	lurbinectedin	1.8%
atezolizumab	1.3%	atezolizumab	1.6%

Table 2b. Systemic Anti-Cancer Regimen Distribution in First Line for Advanced NSCLC Subtype Predicted vs. Expert Curated

Predicted NSCLC, Line 1, n= 25,121		Expert curated NSCLC, Line 1, n= 30,129	
Regimen name (Top 7)	**%**	**Regimen name (Top 7)**	**%**
carboplatin/cisplatin + pembrolizumab + pemetrexed	14.9%	carboplatin/cisplatin + pembrolizumab + pemetrexed	15.4%
pembrolizumab	10.9%	carboplatin/cisplatin + docetaxel/paclitaxel	12.1%
carboplatin/cisplatin + docetaxel/paclitaxel	6.2%	pembrolizumab	10.7%
osimertinib	6.2%	carboplatin/cisplatin + etoposide	9.7%
carboplatin/cisplatin + pemetrexed	5.7%	carboplatin/cisplatin + pemetrexed	8.6%
carboplatin/cisplatin + docetaxel/paclitaxel + pembrolizumab	4.9%	osimertinib	7.0%
carboplatin/cisplatin + durvalumab + docetaxel/paclitaxel	4.3%	carboplatin/cisplatin + docetaxel/paclitaxel + pembrolizumab	5.2%

CONCLUSIONS

- High performing predictive model that delineates NSCLC and SCLC lung subtypes was established with precision and recall metrics of 87% to 96% by leveraging multiple features from structured and unstructured EHR of oncology patients using SLMs, LLMs, and XGBoost1 (eXtreme Gradient Boosting) model.
- Clinical alignment with real-world treatment patterns and biomarker distribution between SCLC and NSCLC was seen in both expert-curated and AI-curated scaled datasets, supporting its utility for studying disease phenotype and associated treatment patterns and outcomes.

REFERENCES

1. Patel, S. P, et al. (2024). Validation of an updated algorithm to identify patients with non–small cell lung cancer in electronic health record and claims data. Journal of Oncology Informatics, 12(3), 145–156.
2. Wang, L., Luo, Y., & Friedman, C. (2019). Natural language processing for populating lung cancer clinical research data. Journal of the American Medical Informatics Association, 26(6), 521–530.
3. Turner, R. M., Lloyd, C., & Lasky, D. (2017). Validation of a case-finding algorithm for identifying patients with non–small cell lung cancer in administrative claims databases. Pharmacoepidemiology and Drug Safety, 26(9), 1103–1110.

CONTACT

Sheenu Chandwani, PhD, schandwani@concertai.com