Natural Language Processing-Optimized Case Selection for Real-World Evidence Studies

Jacob E. Koskimaki,[1] Jenny Hu,[2] Yiduo Zhang,[2] Jose Mena,[1] Nehanda Jones,[1] Elizabeth Lipschultz,[1] Vivek Prabhakar Vaidya,[4] Gabriel Altay,[3] Vance Erese,[3]
Krishna Kumar Swaminathan,[4] Emma Mendonca,[4] Tarun Dutt,[4] Kuldeep Singh,[4] Tian King,[3] Vinay Lakkimsetty,[4] Hussein Al-Olimat,[3] Brittany Manning,[1] George Komatsoulis,[1] Simon Chu,[3] Jeff Ottens[3]

[1]CancerLinQ at the American Society of Clinical Oncology, Alexandria, VA, USA; [2]AstraZeneca LP, Gaithersburg, MD, USA; [3]Tempus Labs, Inc., Chicago, IL; [4]ConcertAI, King of Prussia, PA

Abstract #1556

Introduction

- Accurate data completeness is essential for quality care improvement and research on de-identified patient records.
- Much clinical information remains in unstructured notes or documents and requires methods such as natural language processing (NLP) and curation to process as structured data.
- CancerLinQ is a big data and health information technology platform from the American Society of Clinical Oncology (ASCO) to collect systematically analyze RWE[1] with over 90 active sites and data from 12 source electronic health record (EHR) systems.

- The **objectives** of this study were:
  - To develop clinical criteria for lung cancer cohorts for NLP model development and subsequent curation review
  - To evaluate performance of NLP models employed on a subset of CancerLinQ lung cancer data
  - To evaluate the success of cohort identification using NLP-derived outputs from unstructured clinical content through manual curation review and validation

Objective 1: to develop clinical criteria for lung cancer NLP studies

- A subset of nearly 60,000 lung cancer cases from CancerLinQ containing unstructured content were sent to Tempus and ConcertAI for NLP model development.
- AstraZeneca developed six clinical cohorts of research interest around a range of variables including: EGFR status, cancer stage, histology, radiation therapy, surgical resection and oral medications, which are often incomplete.[2]
- Cohorts of interest were based on clinical criteria and data likely to reside in unstructured notes (**Table 1**).
- **Figure 1** highlights the workflow for NLP model development and validation of predictions through curation.

Key Takeaways

Accurate data completeness is essential for quality care improvement and research studies on de-identified patient records.

NLP models had high precision and recall values for clinical criteria such as EGFR status, stage, histology, radiation, surgical resection and oral medications in lung cancer.

NLP-optimized approaches can significantly improve data completeness and cohort identification over native EHR or curated data alone.

Contact Email: jacob.koskimaki@cancerlinq.org

Table 1: Clinical criteria defining variables of interest for six lung cancer cohorts. Cohorts focused on NSCLC and SCLC, EGFR status, squamous and non-squamous histology, surgical resection, radiation and oral medications of interest.

Cohort	Cohort Description
1a	NSCLC, stage I, II, III, *EGFR*+, complete resection
1b	NSCLC, non-squamous, stage I, II, III, *EGFR* wild type/unknown, complete resection
2a	NSCLC, stage III, unresected, curative radiation to the chest total dose ≥ 50 Gy, received Imfinzi
2b	NSCLC, stage III, unresected, curative radiation to the chest total dose ≥ 50 Gy, did not receive Imfinzi
3	SCLC, received Imfinzi
4	NSCLC, received Tagrisso as first-line treatment

Figure 2: Analytic workflow for NLP processing shown for *EGFR* biomarker identification (model B1-L). The model employs a combination of subtasks and methods including document processing, elastic search queries, rule-based filtering, probabilistic classification and patient-level gene classification.

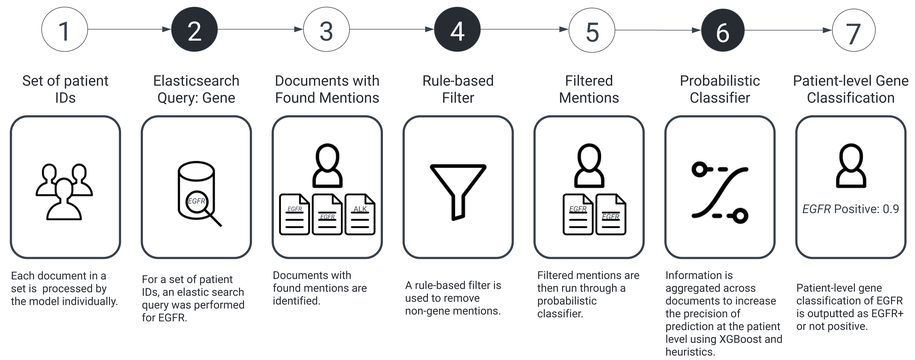

Table 3: NLP-Predicted values for EGFR Status across the Total Lung Cancer Patient Cohort. EGFR status was predicted by NLP on all available data.

EGFR Result	Total (N=54,722)
Positive, n (%)	2,368 (4.3%)
Not positive, n (%)	22,843 (41.7%)
No EGFR results, n (%)	29,511 (53.9%)

Table 4: Rate of NLP Cohort Predictions Successfully Validated by Curators.

Cohort	Number of cases available from NLP-assisted identification methods	Number of cases sent to Tempus and ConcertAI for curation**	Number of cases returned to CancerLinQ with curated content	Percent of successfully-curated cases
1a	408	408	341	83.6%
1b	4313	1500	1285	85.7%
2a	852	620	466	75.2%
2b	3050	750	724	96.5%
3	559	500	402	80.4%
4	971	812	647	79.7%
Total	10153	4590	3865	84.2%

***Cases sent for additional curation were selected in part for minimum target cohort numbers*

- Previous work showed cases sent for curation without NLP processing first only 9.4% qualified for the same clinical cohorts (41,186 cases sent for curation with 3,878 curated and qualifying for a clinical cohort).
- NLP-derived data improved the overall cohort prediction to 84.2% of selected cases.

Figure 1: Process for NLP model development and curation review using CancerLinQ data. Cases were sent to Tempus and ConcertAI for NLP model development. CancerLinQ determined which cases qualify for cohorts of interest, which were independently validated by curators.

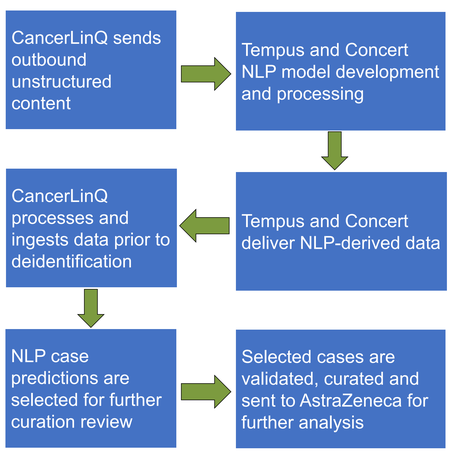

Table 2: Description of NLP models, tasks performed, output, precision, recall and F1 statistics. Models were trained on previously-curated data and generally had acceptable values across domains.

Model ID	Model Output	Variable	Precision	Recall	F1
T1-L	Diagnosis	NSCLC	1	1	1
		SCLC	1	0.98	0.99
T2-B	Metastatic Sites	No	0.98	0.94	0.96
		Yes	0.84	0.94	0.89
H1-L	Histology	Non-squamous	0.98	0.95	0.96
		Squamous	0.82	0.92	0.87
B1-L	*EGFR*	Not positive	0.95	0.98	0.96
		EGFR positive	0.88	0.8	0.84
S1-L	Stage	Stage group I	0.93	0.91	0.92
		Stage group II	0.91	0.8	0.85
		Stage group III	0.94	0.92	0.93
		Stage group IV	0.88	0.95	0.91
PL-1	Modality	Radiation No	0.89	0.84	0.86
		Radiation Yes	0.95	0.97	0.96
P2-L	Surgery	Surgery No	0.96	0.93	0.94
		Surgery Yes	0.9	0.94	0.92
M1-L	Medication	Imfinzi	0.87	0.85	0.86
		Tagrisso	0.93	0.93	0.93

Objective 2: to evaluate performance of NLP models on CancerLinQ data

- NLP models were developed by Tempus and ConcertAI to predict key clinical variables across domains.
- **Figure 2** illustrates the analytic workflow for processing documents and returning NLP-derived content for a key biomarker, *EGFR* status.
- Additional models were developed similarly to EGFR status.
- Models generally had acceptable precision and recall values based on training data and comparisons to previously curated results (precision from 0.88 to 1 and recall from 0.8 to 1 across models and domains) (**Table 2**).
- Once fully developed and validated, NLP models were applied to the full production set of nearly 60,000 lung cancer patients.

Objective 3: to evaluate success of cohort identification using NLP-derived outputs through curation

- The percentage of successfully-curated cases was determined per cohort by dividing the number of curated cases by the number of cases sent for curation and predicted by NLP.
- Nearly half of all lung cancer cases were identified as positive (4.3%) or not positive (41.7%). The remainder did not have EGFR results (53.9%) (**Table 3**).
- **Table 4** shows the number of NLP predictions by cohort, the number sent for curation, and the number successfully curated and independently validated by curators.

Conclusions

- Accurate data completeness is central for quality care improvement and research on de-identified patient data.
- NLP models successfully identified candidate cases in cohorts of interest that would not be available through structured, native EHR data alone.
- NLP models had high precision and recall values based on previously curated training data.
- NLP-derived variables were confirmed through the process of curation with a high total success rate (84.2%), where additional data elements were selected for manual curation review.

References

1. Potter et al., *JCO Clin Cancer Inform.* 2020 Oct;4:929-937.
2. Schorer AE, Moldwin, R, Koskimaki J, *JCO Clin Cancer Inform.* 2022 Jan

Acknowledgments

This study was funded by AstraZeneca. The authors would like to thank the patients, their families and caregivers, and all investigators involved in this study.