

OA08 MOVING BEYOND NEW DRUGS: WHAT ELSE MATTERS?
THURSDAY, SEPTEMBER 09, 2021 - 09:30-10:30

OA08.04 Validation of Scalable, Automated Data Extraction in an Advanced Lung Cancer Patient Population

M. Gauthier¹, J. Law¹, L. Le², J.J.n. Li¹, S. Zahir¹, M. Sung¹, C. Pettengell³, L. Darwish³, S. Aviv³, R. Chu¹, A. Sacher¹, G. Liu¹, P. Bradbury¹, F.A. Shepherd¹, N. Leigh¹

*1*Department of Medical Oncology, Princess Margaret Cancer Centre, Toronto/ON/CA, *2*Department of Biostatistics, Princess Margaret Cancer Centre, Toronto/ON/CA, *3*Pentavere Research Group Inc, Toronto/ON/CA

Introduction: Manual extraction from electronic health records (EHRs) is currently the standard approach for accessing real-world healthcare data but can be time consuming and challenging to maintain over time. Automated data extraction using natural language processing (NLP) is emerging as a viable method of data extraction from structured and unstructured fields of EHRs. While speed of NLP-based data extraction is established, some question the validity of the extracted data. This study compares the accuracy of, and concordance between, manual and NLP-extracted data from EHRs of patients with advanced lung cancer (aLC). **Methods:** EHRs of 1209 patients with aLC were screened using the AI engine, DARWEN™, to identify a subset of 333 patients diagnosed and treated with systemic therapy at Princess Margaret Cancer Centre in Toronto between January 2015 and December 2017. Full feature models were run on all 333 patients to extract data from EHRs, from which 100 patients were randomly selected for manual data extraction by two trained abstractors to validate against NLP-extracted data. An expert adjudicator reviewed inconsistencies between manual and NLP-extracted results and was referenced as the gold standard when calculating accuracy and concordance. **Results:** NLP-extracted data from EHRs proved to be accurate and concordant with manual extraction methods (Table 1). Features with lower syntactic and semantic variation such as patient demographics (i.e., age and sex), characteristics (i.e., histologic subtype and comorbid conditions), and treatment details were reported with high accuracy and concordance. These tend to be the cases where manual reviewers would agree. Conversely, features with richer syntactic and semantic variation requiring deeper clinical interpretation had slightly lower accuracy by NLP extraction and, typically, manual review. By nature of the varying ways that biomarker testing and reporting is documented, extracting this data can be challenging. While NLP detection of biomarker testing was highly accurate and concordant, detection of results was more variable. NLP out-performed manual extraction in identifying metastatic sites with the exception of lung and lymph node metastases, which was due to analogous terms used in radiology reports that were not applied to variable definitions used to train DARWEN™.

Table 1. Accuracy and concordance between manual and NLP data extraction.

	Accuracy (%)		Concordance (%)
	NLP	Manual	
Date of birth	100	99	99
Sex	100	100	100
Date of Stage IV diagnosis (+/- 30 days)	94.0	83.0	77.0
ECOG PS at Stage IV diagnosis	93.0	78.0	71.0
Smoking status	88.0	94.0	82.0
Histologic subtype	98.0	98.0	96.0
First line treatment type	95.0-99.0	96.0-100	92.0-99.0
Treatment type (Any line)	94.0-99.0	84.0-98.0	83.0-96.0
Biomarker Testing Performed	98.0-99.0	97.0-100	96.0-98.0
Biomarker Status (Positive or Negative)	86.2-100	94.7-100	86.2-100
Metastatic Sites of Disease	66.0-99.0	71.0-100	58.0-99.0
Immunosuppressive medications	80.0-100	86.0-100	76.0-100
Comorbidities	96.0-100	96.0-100	93.0-100

Conclusion: The use of NLP technology in oncology provides opportunity for real-world evidence studies at a larger scale than ever before. NLP was not only faster than manual extraction but, for many features, was also more accurate than a traditional manual approach, demonstrating the advances of modern NLP techniques as a scalable alternative to manual extraction.