Identifying Non-Small Cell Lung Cancer Patients From a Cohort of Heterogeneous Lung Cancer Patients Using Boosted Trees on Electronic Health Records Data

Authors: Chandrashekaraiah, Prajwal[1] ; Rudeen, Karl[1] ; Agrawal, Smita[1] ; Thiruvenkadam, Sheshadri[1] ; Vaidya, Vivek P[1] ; Narayanan, Babu O[1]

Institution: [1]Concerto HealthAI

Objectives: Ability to distinguish between subtypes of lung cancer (LC) is important for clinical outcomes and cost analysis, but this information is seldom captured in the structured electronic health record (EHR) data. The objective of this study is to develop and validate an artificial intelligence model to identify non-small cell lung cancer (NSCLC) patients from a cohort of heterogeneous LC patients using de-identified retrospective EHR data.

Methods:

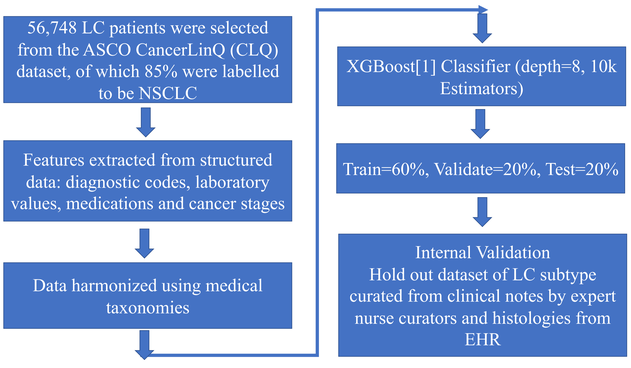

Fig 1: Flowchart explaining the methodology used for modeling, testing and validation

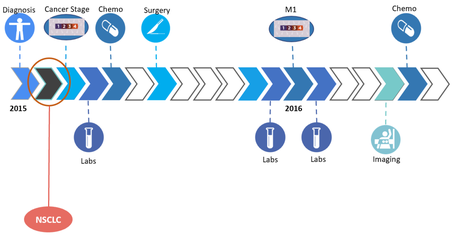

Fig 2: Depiction of a patients journey as captured in the EHR and the features used to build the model

Results: On the test set, the model had an AUC-ROC of **0.93** and overall accuracy of 93%. For identifying NSCLC patients, the precision was 0.93 with recall 0.99 The PPV of the model was 0.93 and NPV was 0.91 (Fig 3). This model compares favorably against a previously developed medications and tests-based NSCLC case finding algorithm using claims data which had an AUC of 0.88 [2]. Features that strongly predicted NSCLC status included genetic testing, surgery, administration of targeted therapies, immunotherapies and chemotherapies including pemetrexed, gemcitabine, taxanes, platins and vinorelbine etc. whereas administration of drugs such as etoposide, topotecan and irinotecan we strong predictors of non-NSCLC status (Fig 4).

Predicted label

True label	Non-NSCLC	NSCLC
Non-NSCLC	1027	700
NSCLC	96	9527

	Precision	Recall
Non-NSCLC	0.91	0.59
NSCLC	**0.93**	**0.99**

Overall Accuracy = 93%, AUC = 0.93

Fig 3: Results of model performance on test set

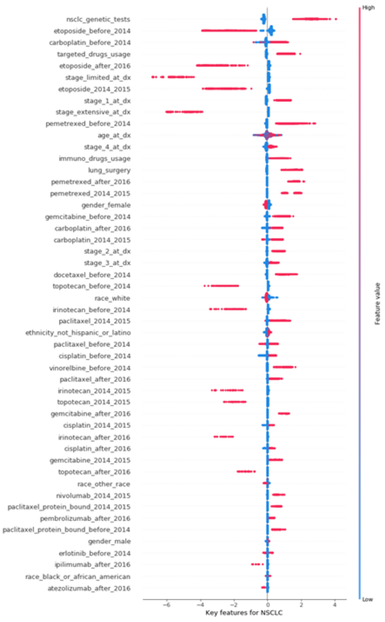

Fig 4: SHAP plot depicting top 50 features of the model

Conclusions: We have developed a model which can identify NSCLC patients from a heterogeneous population of LC patients with a high precision and recall. This could save substantial time and effort by quickly identifying patients for retrospective outcomes and cost studies as compared to expert manual curation.

References:

1. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016)
2. Ralph M. Turner, Yen-Wen Chen, Ancilla W. Validation of a Case-Finding Algorithm for Identifying Patients with Non-small Cell Lung Cancer (NSCLC) in Administrative Claims Databases. Front. Pharmacol., 30 November 2017