Polygenic Risk Scores Improve Non-Small Cell Lung Cancer Prediction in Real-World US Data

Shashi Khan, Vikash Verma, Louis Brooks Jr, Marissa Seligman, Abhimanyu Roy, Abhinav Nayyar, Ankit Arora, Ram Mishra, Pallavi Mohanty, Sudhanshu Chawla, Rishav Singla, Anuj Gupta, Vishan Khatavkar

Background

- Non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases and remains the leading cause of cancer-related mortality, largely due to diagnosis at advanced stages.[1]
- Existing NSCLC risk stratification frameworks primarily rely on demographic and clinical factors such as age, smoking status, and symptoms, which incompletely capture underlying biological susceptibility.
- Polygenic risk scores (PRS) aggregate information across multiple genetic variants to quantify disease susceptibility; however, their application in NSCLC risk prediction using real-world U.S. data remains limited.

Objective

- The objective of this study was to evaluate whether integrating a polygenic risk score -based molecular burden score with clinical and demographic factors improves discrimination of NSCLC in U.S. real-world clinicogenomic data.

Methodology

- Continuous enrollment for ≥12 months pre- and post-index ensured Optum eligible patients. Clinical variables, including demographics, comorbidities, and somatic alterations from routine clinical NGS and EHR were incorporated into a multivariable logistic-regression based classification model to distinguish NSCLC from other malignancies in the follow-up period.
- This retrospective cohort study used Optum® Market Clarity linked with the Optum® Clinicogenomics Database, integrating de-identified claims, EHR, and NGS data for Optum-eligible patients.
- NSCLC patients were identified using ICD-9/10 codes supplemented by NSCLC-specific clinical text, and compared with a mutually exclusive All-Cancer cohort without NSCLC.

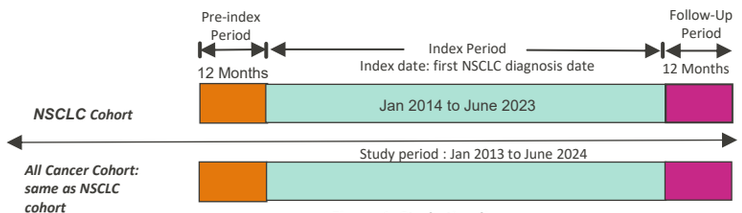

Figure 1. Study timeline

Schematic of study design showing index date definition, 12-month pre-index baseline period, and 12-month post-index follow-up window for NSCLC and All-Cancer cohorts.

- Forty-two NSCLC-relevant genomic alterations across **EGFR, KRAS, TP53, PIK3CA, IDH1/2, RB1, and related genes** were evaluated.
- A PRS-like molecular burden score was constructed using patient-level variant allele frequencies to represent aggregated somatic mutation burden.
- The PRS was included as a continuous predictor alongside clinical variables.
- Multivariable logistic regression modeled NSCLC versus All-Cancer status. The model performance was assessed using AUC, precision, recall, and confusion matrix metrics.

Results

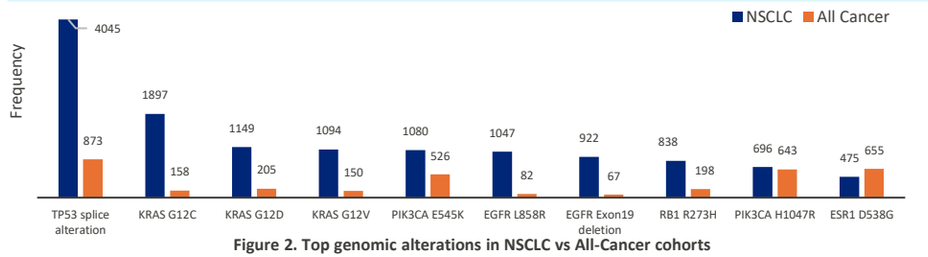

Figure 2. Top genomic alterations in NSCLC vs All-Cancer cohorts

Frequency of the most common genomic alterations across NSCLC and All-Cancer cohorts, highlighting enrichment of TP53 splice alterations, KRAS G12C, and EGFR activating mutations in NSCLC.

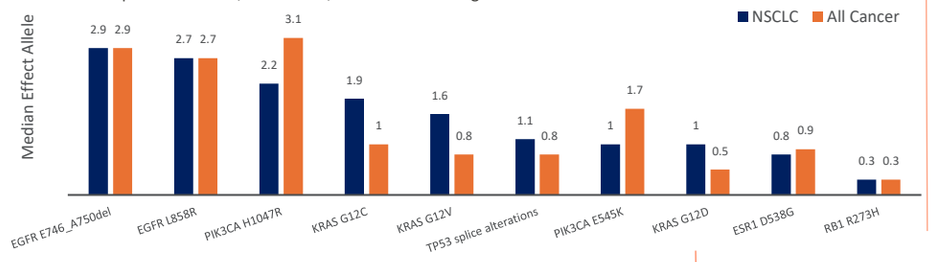

Figure 3. Median PRS across selected genomic alterations

Comparison of mean PRS values for key genomic alterations between NSCLC and All-Cancer cohorts, demonstrating higher PRS values for EGFR and KRAS alterations in NSCLC.

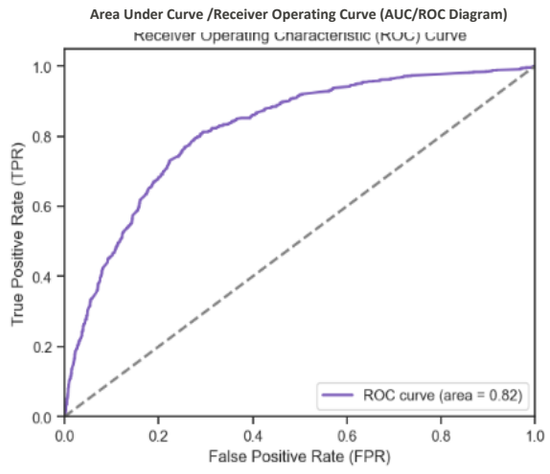

Figure 4. Model performance

ROC curve showing discrimination of the PRS-based model between NSCLC and All-Cancer cohorts (AUC = 0.82).

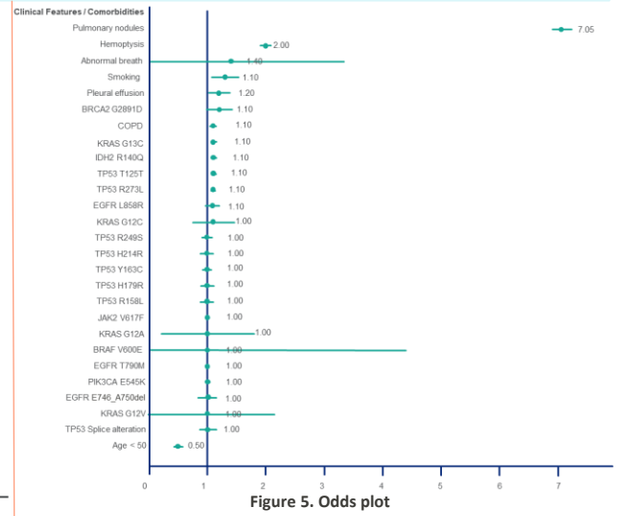

Figure 5. Odds plot

Odds ratios (95% CI) for selected features from the multivariable logistic regression model distinguishing NSCLC from All-Cancer cohorts.

- Of 43,104 initially identified patients with NSCLC, 6,227 met the eligibility criteria, while 12,454 patients formed the mutually exclusive All-Cancer comparator cohort (Figure 1).
- TP53 splice alterations were frequently observed genomic events in NSCLC, followed by KRAS G12C, KRAS G12D, and KRAS G12V. EGFR mutations, including L858R and exon 19 deletions, along with PIK3CA hotspot mutations (E545K and H1047R), were more prevalent in the NSCLC cohort compared with the All-Cancer cohort (Figure 2).
- Molecular burden scores in terms of median PRS were highest for EGFR exon 19 (E746_A750del) deletions, EGFR L858R, KRAS G12C, and TP53 splice alterations in the NSCLC cohort, with more limited differentiation observed for other alterations (Figure 3).
- The multivariable logistic regression–based classification model demonstrated good performance in distinguishing NSCLC from other malignancies (AUC/ROC = 0.82; Figure 4).
- Clinical features, particularly pulmonary nodules and hemoptysis, contributed most strongly to model discrimination, while aggregated genomic features provided complementary discriminatory value beyond clinical and demographic variables (Figure 5).

Conclusions

- In this real-world clinicogenomic analysis, an integrated logistic regression model effectively differentiated NSCLC from other malignancies. These results support clinicogenomic classification approaches and motivate future evaluation of germline risk scores for prospective NSCLC risk assessment.

Reference: 1. Leiter A, et. al., The global burden of lung cancer: status and future trends. Nature reviews Clinical oncology. 2023, 624-39.

Optum