LungFlag, a Machine-Learning (ML) Personalized Tool for Assessing Pulmonary Complications a Community Setting, Demonstrates Comparable Performance in Flagging Non-Small Cell Lung Cancer (NSCLC) Regardless of Sex or Race

Eran N Choman[1], David Morgenstern, PhD[2]

[1] Medial EarlySign, 6 Hanagar st, Hod Hasharon, Israel.
[2] Roche Diagnostics Operations, 9115 Hague Rd, Indiananpolis, IN USA.

Background

Clinical trials targeting heavy smokers have shown to reduce mortality[1] and paved the way for the USPSTF to recommend lung cancer screening based on a combination of age, smoking history and current or recent smoking status. However, less than 27% of Americans diagnosed with lung cancer meet the original USPSTF criteria for screening, leaving many high-risk individuals unable to access screening[2].

Using a risk-based selection score resulted in higher sensitivities compared to criteria using dichotomized age and smoking history[3]. Application of ML-based risk models to cancer screening cohorts has shown increase in screening efficiency expressed by increased compliance, resources yield and increased detection of early stages in particular[4, 5]. The classifiers developed for that purpose used training data that are random samples of the screen-eligible population, with same distribution of the key data elements[6, 7].

Therefore, ML-based models may be vulnerable to sex- and race-based bias arising from historical bias in access to health care as well as biased training data. Demonstrating fairness in the predictions of ML-based models is a prerequisite to their acceptance by clinicians, patients and policy-makers. We assessed the clinical performance of LungFlag based on sex and race, two key demographic subgroups at risk for disparate outcomes due to bias.

Methods

The LungFlag machine learning model is a retrospectively validated ML model[8] that was trained on US-based population from an integrated health network serving a diverse population, randomly selected to have good representation of the whole network population and included about 200,000 individuals including over 6,500 NSCLC cases. LungFlag uses existing routine outpatient lab measurements, smoking history, comorbidities, and demographic data to flag high-risk individuals.

LungFlag demonstrated good ability to identify individuals who are at elevated risk for lung cancer and other pulmonary conditions. When evaluating the performance we compared 2 models - LungFlag and the adaptation the commonly used PLCOm2012 to work on EMR data (mPLCOm2012).

In order to evaluate the fairness of the models, we calculated sensitivity and odds ratio in sex and race subgroup at an overall positivity rate of 3%. We chose a case-control design based on a large US-based community and outpatient dataset including 39,135 case patients with NSCLC and 212,454 contemporaneous NSCLC-free controls. We included ever-smokers, ages 45-80, with available lab measurements from 3-12 months before diagnosis, and a minimum follow-up of 24 months. Sub-populations with less than 1% representation from the total population were excluded.

Results

The population characteristics is detailed below:

Table 1: Population Characteristics

Parameter	Case Patients	Control Subjects
Total	39,135	212,454
Age: Avg (SD)	68 (10.4)	62 (9.8)
Sex: F : M	19,555 (50%) : 19,580 (50%)	112,720 (53%) : 99,734 (47%)
Race: W : B	33,724 (89%) : 4,058 (11%)	179, 387 (89%) : 19,554 (11%)
Heavy Smokers*	19%	10%
COPD	41%	15%
NSCLC Stages: %0-II : %III-IV	29% : 71%	

* USPSTF criteria

The comparison between the sub-populations presented by sensitivity and specificity indexes is detailed below:

Table 2: Sex Fairness – 3% Positivity Rate

Sub-Population	mPLCOm2012			LungFlag		
	Specificity	Sensitivity	Odds Ratio	Specificity	Sensitivity	Odds Ratio
Female	97.1 [97.0 - 97.2]	16.4 [15.9 – 16.9]	6.6 [6.1 – 7.1]	97.2 [97.1 - 97.3]	23.1 † [22.6 - 23.7]	10.4 † [9.8 – 11.2]
Male	97.0 [96.9 – 97.1]	15.8 [15.3 – 16.3]	6.1 [5.6 – 6.5]	96.8 [96.7 - 96.9]	23.4 † [22.8 - 24.0]	9.2 † [8.7 – 9.9]

† Statistically significant difference LungFlag versus mPLCOm2012

No statistically significant difference was observed in the specificity and sensitivity of the two models to flag individuals that were diagnosed with NSCLC based on sex. LungFlag outperformed mPLCO2012 with a statistically significant improvement in sensitivity of 41% for Females and 48% for Males.

Table 3: Race Fairness – 3% Positivity Rate

Sub-Population	mPLCOm2012			LungFlag		
	Specificity	Sensitivity	Odds Ratio	Specificity	Sensitivity	Odds Ratio
White	96.9 [96.8 - 97.0]	16.6 [16.2 – 17.0]	6.2 [5.8 – 6.6]	96.9 [96.8 - 97.0]	23.5 † [23.1 - 24.0]	9.6 † [9.1 – 10.2]
Black	97.7 * [97.5 - 97.9]	13.3 [12.2 – 14.4]	6.5 [5.4 – 7.8]	97.1 [96.8 - 97.3]	23.7 † [22.5 - 25.0]	10.4 † [8.8 – 12.0]

* USPSTF criteria
† Statistically significant difference LungFlag versus mPLCOm2012

No statistically significant difference was observed in the specificity and sensitivity of the LungFlag model according to race. Statistically significant differences in specificity and sensitivity rates were demonstrated in the mPLCOm2012 model to flag individuals that were diagnosed with NSCLC in the Black versus White sub-population. LungFlag outperformed mPLCO2012 with statistically significant improvement in sensitivity of 42% for White and 78% for Black.

Conclusions

We determined that in a large, community-based retrospective dataset the LungFlag model demonstrated fairness with respect to sex and race, showing similar clinical sensitivity while the mPLCOm2012 model demonstrated statistically significant differences between the sub-populations. Additionally, LungFlag demonstrated statistically significant improvement over mPLCO2012 (41%-78%) for the sex- and race-based sub-populations. Further assessment in prospective studies and in additional racial sub-populations is recommended to support this conclusion.

References

[1] National Lung Screening Trial Research T, Aberle DR, Adams AM, et al. 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409. 10.1056/NEJMoa1102873

[2] Pinsky PF, Berg CD. 2012. Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered? J Med Screen 2012;19:154-6. 10.1258/jms.2012.012010

[3] Julia Walter, et al. 2023. Comparison of the sensitivity of different criteria to select lung cancer patients for screening in a cohort of German patients, Volume12, Issue7, April 2023, Pages 8880-8896.

[4] Daniel Underberger, et al., 2022. Collaboration to Improve Colorectal Cancer Screening Using Machine Learning. NEJM Catal Innov Care Deliv 2022;3(4).

[5] Ran Goshen, et al., 2018. Computer-Assisted Flagging of Individuals at High Risk of Colorectal Cancer in a Large Health Maintenance Organization Using the ColonFlag Test. JCO Clin Cancer Inform. 2018 Dec;2:1-8.

[6] Gavin C. Cawley, et al. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research 11 (2010) 2079-2107.

[7] Bianca Zadrozny, et al.2004. Learning and evaluating classifiers under sample selection bias. First international conference on Machine learning, July 2004.

[8] Gould MK, Huang BZ, Tammemagi MC, et al. 2021. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. AJRCCM 204(4).