Validation of LungFlag™ Prediction Model Using Electronic Medical Records (EMR) On Taiwan Data

Eran N Choman[1], Alon Lanyado[1], Eitan Israeli, PhD[1], Nicolò Olghi[2], Yue Jin, PhD[3], Shang-Yun Liu, PhD[4], Milan Obradovic, PhD[2], Pan-Chyr Yang, MD, PhD[5]

[1] Medial EarlySign, 6 Hanagar st, Hod Hasharon, Israel.
[2] F. Hoffmann-La Roche Ltd, Basel, Switzerland.
[3] F. Hoffmann-La Roche Ltd, Santa Clara, USA.
[4] F. Hoffmann-La Roche Taipei, Taiwan.
[5] National Taiwan University Hospital and College of Medicine, Taipei, Taiwan.

Sponsored By F. Hoffmann-La Roche Ltd

Background

Lung cancer (LC) is the leading cause of cancer death according to the WHO. In Taiwan, LC accounted for 9,629 deaths in 2020 and contributed one-fifth of total cancer deaths. More than 50% of those are among non-smokers. To reduce mortality rate, starting 2022, the Ministry of Health and Welfare launched the LC Early Detection Program to provide biennial low-dose computed tomography (LDCT) screening for high-risk groups based on smoking and age criteria. Risk-scores are considered to present better classification power than criteria in identifying high-risk populations, two of these are the dedicated questionnaire based PLCOm2012 and machine learning (ML) structured EMR based LungFlag. The latter was developed using data from Kaiser Permanente Southern California, USA [1-4].

Methods

We examined if EMR data in Taiwan could be used to better identify individuals at higher risk of LC using risk scores. We extracted data of individuals with diagnosed LC (cases) and randomly selected individuals without records of LC (controls) from National Taiwan University Hospital (NTUH) medical records. We tested and compared the LungFlag model to the PLCOm2012 model - both applied to the extracted EMR data which included individuals with ages 40 to 80 between 2010-2022 and the minimal mandatory medical information necessary for the models captured in EMR (e.g., demographics, diagnosis, smoking information). Model performance was evaluated by AUC, Sensitivity, Odds-Ratio (OR) indexes for multiple false positive rate (FPR) cut-offs (top 3% and 10%) using bootstrap methodology.

Results

Total data of 65,882 individuals was extracted, 60,751 met the age criteria (10,557 Cases, 50,194 Controls). Average pack-years for ever smokers were 38 and 30 for cases and controls, respectively.
Calculation of PLCOm2012 and LungFlag was performed on the diagnosis date for cases, and on a randomly selected date for controls requiring at least 2 years of follow-up period to reduce the potential false negatives.
Both LungFlag and PLCOm2012 outperform the USPSTF criteria in selecting individuals at risk from the ever-smoker population. LungFlag demonstrated statistically significant superiority over PLCOm2012 with OR at 3% FPR of 11.7 vs. 5.9 among ever-smokers and 9.0 vs. 4.3 among USPSTF eligible, respectively. Furthermore, LungFlag had an OR of 9.9 at FPR of 3% for never smokers and 2.4 for individuals without smoking status.

Conclusions

LungFlag was demonstrated to be able to work on Taiwanese EMR data, presented superiority over PLCOm2012. Both models outperformed the USPSTF selection criteria with statistical significance. The results were comparable to the previous validation studies. LungFlag can potentially support Taiwanese screening program in identifying elevated risk populations among ever-smokers. LungFlag has shown classification power also for never-smokers and more model retraining for never smokers using local data should be considered to examine if its classification power could be further increased.

Comparison of Models Performance in detection of Lung cancer at False Positive Rates of 3% or 10% from each sub-population by LungFlag vs PLCOm2012 suggests superiority of the LungFlag model

Sub-Population	Controls	Cases	Model	AUC	False Positive Rate 3%		False Positive Rate 10%	
					Sensitivity %	Odds Ratio	Sensitivity %	Odds Ratio
Ever smokers	4,986	911	$PLCO_{2012}$	0.697 [0.678 - 0.716]	15.3 [12.3 - 18.2]	5.9 [4.5 - 7.2]	33.2 [30.0 - 36.8]	4.5 [3.9 - 5.2]
			LungFlag	0.769 † [0.750 - 0.788]	26.5 † [23.0 - 31.0]	11.7 † [9.7 - 14.5]	46.6 † [42.8 - 50.7]	7.9 † [6.7 - 9.3]
USPSTF	1,035	342	$PLCO_{2012}$	0.702 [0.670 - 0.732]	11.6 [7.5 - 16.8]	4.3 [2.6 - 6.5]	27.2 [22.0 - 33.0]	3.4 [2.5 - 4.4]
			LungFlag	0.762 † [0.733 - 0.791]	21.6 [15.0 - 28.6]	9.0 † [5.8 - 13.0]	41.8 † [34.3 - 47.7]	6.5 † [4.7 - 8.2]
Never Smokers	19,372	2,207	LungFlag	0.649 [0.638 - 0.661]	13.6 [12.5 - 15.0]	9.9 [8.3 - 12.1]	26.6 [25.0 - 28.5]	3.4 [3.1 - 3.8]
Unknown smokers	26,589	7,439	LungFlag	0.581 [0.574 - 0.588]	6.9 [6.6 - 7.2]	2.4 [2.3 - 2.5]	* NA	* NA

* NA - Most patients only have sex and age information – causing same calculated score. Therefore, unable to select ~10% of the population.

† Statistically significance difference between LungFlag and PLCOm2012 (P<0.05)

References

1. Junjie Huang et al., Cancer Incidence and Mortality in Asian Countries: A Trend Analysis. Cancer Control. 2022 Jan-Dec; 29: 10732748221095955.
2. Vivian W. Y. Lee et al., Burden of smoking in Asia-Pacific countries. Tob Induc Dis. 2021; 19: 28.
3. Fei Zhou et al., Lung cancer in never smokers—the East Asian experience. TLCR , Vol 7, No 4 (August 28, 2018)
4. 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).