Dynamic Lung Cancer Risk Modeling in Non or Light-Smokers: Leveraging Radiomics and Machine Learning

Morteza Salehjahromi[1] , Rongqin Zheng[2], Hui Li[1], Eman Showkatian[1], Mohamed Qayati[1], Ziping Li[2], Sheeba J. Sujit[1], Muhammad Aminu[1], Maliazurina B. Saad[1], Lingzhi Hong[1,2], John D. Hazle[1], David Jaffray[1,4], John V. Heymach[2], Jianjun Zhang[3,2,5,6], Jia Wu[1,2,4]

1. Department of Imaging Physics, MD Anderson Cancer Center, Houston, TX USA
2. Department of Thoracic/Head and Neck Medical Oncology, MD Anderson Cancer Center, Houston, TX, USA
3. Department of Genomic Medicine, MD Anderson Cancer Center, Houston, TX, USA
4. Institute for Data Science in Oncology, MD Anderson Cancer Center, Houston, TX, USA
5. Lung Cancer Genomics Program, MD Anderson Cancer Center, Houston, TX, USA
6. Lung Cancer Interception Program, MD Anderson Cancer Center, Houston, TX, USA

Introduction

Lung cancer, predominantly linked to smoking, also significantly affects non-smokers, who account for about 20% of lung cancer deaths in the U.S. Current screening guidelines exclude non-smokers from low-dose CT screenings, highlighting a crucial research gap. This study introduces a novel dataset of 122 light or non-smoking patients diagnosed with lung cancer, analyzed through their 622 CT scans to predict lung cancer risk by examining extracted lung features. While smoking is a well-known risk factor, non-smokers are also at substantial risk. Lung cancer in non-smokers typically manifests as adenocarcinoma, which develops more slowly, providing a critical window for early diagnosis and effective treatment. The absence of routine low-dose CT screening for never smokers in USPSTF guidelines underscores the need for enhanced research. This research represents possibly the first extensive radiomic analysis focusing exclusively on a non-smoker cohort.

Methods and Predictive Modeling

The study curated a unique cohort of 122 non-smoking lung cancer patients, analyzing longitudinal CT scans taken up to 14 years before diagnosis, focusing primarily on the last two years. Feature extraction, including radiomic analysis, was used and fed to the random survival forest (RSF) method within a nested cross-validation framework to assess lung cancer development risk. By extracting features from both the lung nodules and surrounding pulmonary structures, the study evaluates these characteristics to predict lung cancer risks. The predictive accuracy of the model is quantified using the concordance index (C-index), assessing the likelihood and timing of lung cancer development. The variations of important predictive features are illustrated through box plots that detail the selection frequency and rank across different time frames. We also showcased a comparative visualization of the model's predictive performance against the established Brock University model.

Future Directions

The findings underscore the potential of radiomic analysis in identifying lung cancer risk among non-smokers, a group currently overlooked in screening protocols. This could lead to more personalized diagnostic and treatment strategies tailored for non-smokers.