Smoking history extracted from electronic health record (EHR) using machine-learning methods exhibited distinct tumor mutational patterns in patients with lung cancer

PrimaryAuthor: Semanti Mukherjee, PhD[1]
Secondary Authors Andrew Schroeder MS[1];Subrata Chatterjee, Phd[1]; John Cadley, MS[1]; Christina Falcon[1]; Justin Lee, MD PhD[1] ; Miika Mahine, PhD[1]; Chaitanya Bandlamudi PhD[1]; Yelena Kemel; David Solit, MD[1]; Mark Kris, MD[1]; David Jones, MD[1]; Michael Berger, PhD[1]; Adam J Schoenfeld, MD[1]; Fernanda Palubriaginof, MD PhD[1]; Jonine Bernstein, PhD[1]; Nikki Schultz, PhD[1]; Charles M. Rudin, MD PhD[1]; Kenneth Offit, MD[1];
Affliation: [1]Memorial Sloan Kettering Cancer Center, New York, NY

Though smoking is a major risk factor for lung cancer, it has been a challenge to collect patients' smoking history information accurately from the EHR due to data inconsistency and incompleteness. To address these challenges, we utilized a weak supervision methodology to automatically annotate smoking status of patients with lung cancer. We assessed the performance and correlated smoking behavior with tumor characteristics.

We analyzed 6,355 patients with lung cancer who underwent tumor profiling with MSK-IMPACT. In total, 14,555 unstructured clinical notes were extracted from EHR at the Memorial Sloan Kettering Cancer Center. The weak supervision methodology used a generative model for intermediate labels that were subsequently tuned by a ML classifier to generate the final labels. Clinical notes from randomly sampled set of 564 patients were used for performance assessment. The rest of the patients were split into training and validation datasets used for model training and hyperparameter tuning. Pack years were also extracted from clinical notes using Regex. We next correlated smoking metrics with tumor characteristics including tumor mutation burden (TMB) and chromosomal instability, as inferred by the fraction of genome altered (FGA). Multivariate analysis was conducted after controlling for age at sequencing, gender, histological subtypes, and ancestry for primary and metastatic tumor samples separately.

The weak supervision classifier had almost perfect performance for 2-label classification model (ever smokers and never smokers) with macro F1-score: 97.7%,; balanced accuracy : 97.1%, 97.1%, precision:98.4%, 98.4% and recall: 99.5%,94.6% respectively. For 3-label classification model (never smoker, former smoker, and current smoker), with macro F1-score : 79.8%; balanced accuracy: 97.1%, 86.7%, 71.2%, precision: 93.9%, 90.1%, 61.7%, recall: 96.1%, 93.3%, 46.0% respectively. Analyzing genomic data, we observed that smoking status(smoker vs. never smoker) and pack-years was associated with higher TMB in both primary and metastatic tumor samples ($p<2e-16$ for both). FGA was significantly associated with smokers compared to never smokers in primary tumors samles ($p=2.8e-4$). Among smokers diagnosed with lung adenocarcinoma, significantly high FGA in primary tumor samples was observed in males compared to females after adjusting for pack-years and clinical variables ($p= 3.3e-3$).

We demonstrated high performance of the weak supervision framework for automated curation of smoking history from EHR for a large lung cancer dataset. The genomic results confirmed distinct mutational patterns associated with smoking behavior in patients with lung cancer. We are currently exploring multimodal approaches by including chest CT images and "time of quitting" to improve performance of the 3-class model.