Validation of Natural Language Processing (NLP) for Automated Ascertainment of EGFR and ALK Tests in SEER Cases of Non-Small Cell Lung Cancer (NSCLC).

Background: The Surveillance, Epidemiology, and End Results (SEER) registries lack information on the Epidermal Growth Factor Receptor (EGFR) mutation and Anaplastic Lymphoma Kinase (ALK) gene rearrangement test results. With the goal of enabling population-based outcomes research in molecularly selected NSCLC subgroups, we conducted a validation study of NLP for ascertainment of EGFR and ALK testing from electronic pathology reports (e-paths) of patients included in the Seattle-Puget Sound (SPS) and Kentucky Cancer (KCR) SEER registries.

Methods: We obtained 4,278 and 1,041 e-paths pertaining to 1,634 and 565 patients with stage IV non-squamous NSCLC diagnosed from 1/1/2011 to 12/31/2013 and included in the SPS and KCR registries, respectively. Two oncologists independently reviewed all reports to generate a gold-standard dataset. We used 855 of the SPS reports to train hybrid rule-based and machine learning algorithms for detection of test status (reported vs. not reported), and test result if reported (positive vs negative) for EGFR mutational analysis and ALK testing by FISH, IHC, or gene sequencing. In the remaining 3,423 SPS reports, we conducted a 5-fold cross-validation analysis to estimate the internal NLP sensitivity, specificity, positive predictive value, and negative predictive value for test status and results, respectively. We used a hierarchical rules system to assess the NLP accuracy at the patient level. For external validation, we repeated all analyses in the KCR dataset.

Results: In the SPS internal validation report sample, the validity metrics ranged from 97% to 99% for EGFR and ALK test status, and from 95% to 100% for EGFR and ALK test results, respectively. In the KCR external validation report sample, the metrics ranged from 74% to 96% for EGFR and ALK test status, and 2% to 100% for test results, respectively. At the patient level, the NLP accuracy for EGFR and ALK was 95% and 96%, and 70% and 72% in the SPS and KCR cohorts, respectively.

Conclusions: NLP is a valid method for determining EGFR and ALK test status and results for patients included in SEER registries with access to e-path, but the algorithms likely need to be registry-specific.