A large language model based approach for extraction of molecular biomarker information from pathology reports: a nationwide study of *EGFR* and *KRAS* testing rates in patients with lung cancer in the Netherlands

Vincent D. de Jager[1], Michiel van der Ree[2], Betzabel Cajiao Garcia[1], Chantal Kuijpers[3], Anthonie J. van der Wekken[4], Léon van Kempen[1,5], Ed Schuuring[1], Stefan Willems[1]

[1]Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[2]Center for Information Technology, University of Groningen, Groningen, the Netherlands
[3]The Dutch Nationwide Pathology Databank (Palga), Houten, the Netherlands
[4]Department of Pulmonary Diseases and Tuberculosis, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[5]Department of Pathology, University Hospital Antwerp, University of Antwerp, Edegem, Belgium

Background: National cancer registries and pathology report databases contain valuable real-world information. However, manual assessment is currently needed for extraction of molecular biomarker data from pathology reports, which is labor-intensive, costly and causes significant time delay for data-analysis. We aimed to develop a large language model (LLM) based method to extract predictive biomarker testing data for *EGFR* and *KRAS* from pathology reports of patients with lung cancer.

Methods: Patient cohorts and pathology reports were derived from the Netherlands Cancer Registry and the nationwide pathology database. Manually-captured data of *EGFR* and *KRAS* in 3,887 patients diagnosed with metastatic non-small cell lung cancer (NSCLC) in 2019 were used for model training, testing and validation. Variables included *EGFR*/*KRAS* testing, use of next-generation sequencing, and test results. The final model was applied to pathology reports of 4,122 patients diagnosed with metastatic NSCLC between July 2022 and June 2023. To determine model accuracy, manual annotation was performed for 410 random cases.

Findings: In the testing set of the 2019 cohort, the model yielded F1 scores ≥0.98 across all variables. In the 2022-2023 testing set, F1 scores ≥0.95 were achieved. None of the manually annotated positive molecular test results were missed. Manual re-checking of one model-reported, false-positive *KRAS* mutation revealed incorrect manual annotation. Standardized notation of reported mutations was correct in 98.7% (for *KRAS*) and 100.0% for *EGFR*). In the entire 2022-2023 cohort, model output revealed testing rates of 88.1% for *KRAS* and 86.4% for *EGFR*. The model reported positive *KRAS* and *EGFR* test results in 40.5% and 11.5% of tested patients, respectively.

Interpretation: It is possible to train and use an LLM based model to accurately extract biomarker testing results from pathology reports. This application enables rapid, low-cost assessment of biomarker testing rates and incidence rates of specific mutations based on nationwide-collected pathology reports.