

Relevance and accuracy of ChatGPT-generated NGS reports with treatment recommendations for oncogene-driven NSCLC.

Zac Hamilton, Noor Naffakh, Natalie Marie Reizine, Frank Weinberg, Shikha Jain, Vijayakrishna K. Gadi, Christopher Bun, Ryan Huu-Tuan Nguyen; University of Illinois Chicago, Chicago, IL; Department of Hematology Oncology University of Illinois of Chicago College of Medicine, Chicago, IL; CancerIQ, Chicago, IL

Background: Next-generation sequencing (NGS) is a routine clinical practice in advanced NSCLC. NGS reports are information-dense and clinical interpretation remains a challenge. ChatGPT is a large language model (LLM) AI chatbot that can generate text in response to user-generated prompts. We sought to assess the clinical relevance and accuracy of ChatGPT-generated NGS reports with first-line (1L) treatment recommendations for NSCLC patients with targetable driver oncogenes. **Methods:** Eight driver oncogenes with FDA-approved targeted treatment for 1L stage IV NSCLC were identified in the latest NCCN Clinical Practice Guidelines available to the AI model (version 5, September 2021). The prompt, "Create a next-generation sequencing report with a list of first-line treatment options for a patient with stage IV non-small cell lung cancer with an [oncogenic driver]." was run in a separate "new chat" 10 times for each driver oncogene (n=80). Each ChatGPT output was recorded and scored. The Relevance Score (RS) awarded 1 point for every NCCN preferred option and 0.5 points for each "other recommended" treatment listed in the AI-generated output, divided by the maximum possible score for the driver oncogene. Spurious recommendations were awarded 0 points. The Accuracy Score (AS) represents reported treatment options listed in NCCN over the total number of treatments in a report. Percentage of reports listing an NCCN-preferred 1L therapy, a clinical trial as an option, and character and word count were also captured. **Results:** The average length of the AI-generated NGS reports was 117 words (range: 44 – 232). The median number of treatments recommended was 5 (range: 3 – 8). An oncogenic driver-specific preferred 1L treatment was included in 55 reports (68.8%), and a recommendation to explore clinical trials was listed in 43 reports (53.8%). The RS for the total sample was 0.59 (95% CI: 0.52 – 0.65), and the AS was 46.0% (95% CI: 40.2% – 51.8%). **Conclusions:** ChatGPT can rapidly generate concise NGS reports with treatment options for NSCLC with driver oncogenes. Recommendation relevance was moderate, and accuracy was limited with high variability across oncogenes. Overall, ChatGPT recommendations were promising given the complexity of the task with no prompting or training provided to the AI. As LLM AI platforms mature, they may generate more relevant and accurate NGS reports, offering a potentially valuable tool for NGS report annotation for clinicians, and increased accessibility for patients. Research Sponsor: RHN is a recipient of the Robert A. Winn Diversity in Clinical Trials Career Development Award, funded by Bristol Myers Squibb Foundation.

| Oncogene | RS | Std Dev | AS (%) | Std Dev (%) |
|-------------------------|------|---------|--------|-------------|
| EGFR exon 19 del. | 0.49 | 0.20 | 59.9 | 18.6 |
| EGFR exon 21 L858R mut. | 0.46 | 0.15 | 57.0 | 11.4 |
| ALK rearrangement | 0.75 | 0.13 | 73.6 | 21.9 |
| ROS1 rearrangement | 0.86 | 0.13 | 70.2 | 17.5 |
| BRAF V600E mut. | 0.67 | 0.00 | 20.3 | 2.8 |
| NTRK1/2/3 gene fusion | 1.00 | 0.00 | 48.7 | 11.1 |
| Metex14 skipping mut. | 0.10 | 0.19 | 7.3 | 13.4 |
| RET rearrangement | 0.37 | 0.15 | 31.2 | 11.4 |
| Total | 0.59 | 0.31 | 46.0% | 26.6% |