O-8

Evaluation of Large Language Model-Based Extraction of Neuroendocrine Neoplasm Variables from Pathology Reports

Jasmin M. Francisco[1], Anwen Huang[2], Alan Paciorek[1,3], Phi Le[2], Bryan Khuong Le[1], Emily K. Bergsland[1,2], Aman Chauhan[1,2], Li Zhang[1,3]

[1]University of California, San Francisco (UCSF) Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA; [2]Department of Medicine, University of California, San Francisco; [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA

BACKGROUND

Manual extraction of clinical information at scale remains challenging, especially for heterogeneous conditions like neuroendocrine neoplasms (NENs). This process often requires searching through unstructured clinical records, which is labor-intensive. Recent studies suggest that large language models (LLMs) can effectively automate data extraction for clinical research. This study aims to evaluate LLM performance for outcomes research in NENs by comparing LLM-extracted data with a reference dataset curated by human abstractors.

METHODS

We developed a data extraction pipeline using OpenAI's GPT-4o LLM through a secure, IRB-approved platform authorized for use with protected institutional and clinical data. A total of 364 pathology reports were downloaded from the electronic medical record (EMR). For each report, the LLM was instructed to extract five diagnostic variables relevant to NENs: biopsy site (primary or metastatic), differentiation (well, moderately, or poorly-differentiated), Ki-67 index, histological grade (G1-G3), and stage (localized, regional, or distant). LLM prompt instructions were designed to mirror the extraction guidelines used during the manual extraction process. To account for output variability, the LLM processed each report three times, and the most frequently recorded output was retained as the final output for each variable for comparison.

RESULTS

We compared the accuracy of LLM-generated outputs with manually extracted outputs. Missing values were coded as "Unknown", and concordance included the cases in which both results were "Unknown". The LLM achieved the highest accuracy for differentiation (89%, 95% CI 85-91%), followed by Ki-67 index (85%, 95% CI 81-89%), biopsy site (85%, 95% CI 81-88%), and grade (83%, 95% CI 79-86%), but performance was considerably lower for stage (51%, 95% CI 46-56%).

CONCLUSIONS

The LLM demonstrated promising performance for extracting key diagnostic variables from pathology reports, though further optimization is needed, particularly for more complex variables, such as stage. Discordance across differentiation, Ki-67 index, biopsy site, and grade were primarily explained by the different pathology reports used for each patient during the manual extraction process. Lower accuracy observed for stage reflects the complexity of stage assignment, which requires information beyond pathology reports. Future work will focus on refining prompt design and incorporating additional clinical data sources to improve extraction accuracy and broaden the scope of automated data abstraction. Overall, our findings suggest a potential use of LLMs in extracting variables from EMRs to facilitate research on NENs.

ABSTRACT ID #36534