HealthManagement
Promoting Management and Leadership

Using ChatGPT to Structure Data from Mechanical Thrombectomy Radiology Reports

The use of mechanical thrombectomy for acute ischemic stroke, especially in cases of large vessel occlusion, has become standard practice. This procedure involves various key metrics such as time intervals, imaging scores, and the number of manoeuvres performed, which are crucial for research and quality control. Traditionally, extracting this data from medical reports is a manual and time-consuming process, prone to errors. Recent advancements in language models like GPT-4 and its predecessor, GPT-3.5, offer promising solutions. These models have demonstrated the ability to generate standardised reports from free-text radiology reports accurately and efficiently, reducing the risk of errors. GPT-4 has shown particular effectiveness in maintaining accuracy while simplifying reports. While previous studies have utilised similar language models for extracting data from radiology reports, there's been limited exploration in the context of neuro-interventional procedures like mechanical thrombectomy. A recent study published in Radiology aims to assess whether GPT-4 can effectively extract data from free-text neuroradiology reports related to mechanical thrombectomy for acute ischaemic stroke, thereby facilitating the creation of comprehensive databases. Additionally, the study seeks to determine if the earlier version, GPT-3.5, can serve as a viable alternative to GPT-4, considering its cost-effectiveness.

Retrieving and Processing Thrombectomy Reports for Medical Data Extraction

Reports from patients who underwent mechanical thrombectomy for ischaemic stroke between November 2022 and September 2023 were collected from a single institution. Inclusion criteria were patients over 18 with confirmed large or medium vessel occlusion, and exclusion criteria were incomplete reports or absence of occlusion confirmation. A dataset of at least 100 reports was aimed for based on prior studies. An additional 30 reports from a German institution between September 2016 and December 2019 were obtained to assess model generalizability. A German prompt was created and tested on 20 reports. Reports were processed by ChatGPT to generate CSV tables with procedural details, then manually reviewed by an interventional neuroradiologist. Only exact matches between model and expert entries were considered correct. Format errors (e.g., incorrect punctuation) were distinguished from content errors (e.g., incorrect information). For categories with high error rates, a more detailed prompt was used for reanalysis.

Comparative Analysis of GPT-4 and GPT-3.5 Performance

A total of 107 reports from patients who underwent mechanical thrombectomy were initially collected, with 100 patients ultimately included in the study. Both GPT-4 and GPT-3.5 successfully processed all reports. GPT-4 achieved a higher accuracy rate (94.0%) compared to GPT-3.5 (63.9%). The accuracy varied across categories, with GPT-4 showing high accuracy for most variables. Agreement between GPT-4 and the neuroradiologist was very good ($\kappa = 0.93$), while GPT-3.5 demonstrated moderate agreement ($\kappa = 0.59$). Most errors from both models were due to content rather than format. A more detailed prompt improved accuracy, particularly for GPT-4, reducing errors from 39 to 19 for a specific category. GPT-3.5 had a higher overall error rate, with a majority being content errors.

Implications and Limitations of Language Model Usage for Medical Data Extraction

In this retrospective study, the efficacy of GPT-4 and GPT-3.5 in extracting data from free-text reports on mechanical thrombectomy for ischaemic stroke patients was assessed. GPT-4 demonstrated higher accuracy compared to GPT-3.5, with correct data extraction rates of 94.0% and 63.9%, respectively. The study validated these findings with external reports, showing similar trends. Previous studies have also shown the effectiveness of GPT-4 in various medical text mining tasks, supporting the usability of large language models (LLMs) for data extraction from free-text reports. While GPT-4 outperformed GPT-3.5, certain data points showed poor results, indicating the need for human supervision. Optimisation of prompts could potentially reduce incorrect data entries. However, the study acknowledges several limitations, including its retrospective nature, limited external validation, language-specific prompts, and the need for ongoing model updates and human oversight.

Despite these limitations, the study concludes that GPT-4 could serve as an alternative to manual data extraction, potentially improving efficiency in retrieving procedural data from radiology reports. However, human supervision remains essential due to the occurrence of errors.

Source: RSNA Radiology

Image Credit: iStock

Published on : Thu, 25 Apr 2024