Automating Pennsylvania Act 112 Follow-Up Recommendation Detection in Radiology Reports Using Large Language Models

Satvik Tripathi, Researcher, Department of Radiology , University of Pennsylvania
Shawn Lyo, MD; Tessa Cook, MD, PhD, CIIP, FSIIM, FCPP

Introduction

Pennsylvania Act 112 requires patient notification when imaging reveals a finding that requires follow-up imaging within 90 days. While structured reporting macros exist to flag these cases, their inconsistent usage by radiologists can lead to missed notifications. This creates a potential compliance gap. Large language models (LLMs) have demonstrated strong capabilities in understanding medical text and context, suggesting they could reliably identify reports that meet Act 112 notification criteria, regardless of whether the macro was used. However, their effectiveness in this specific regulatory compliance use case has not been systematically evaluated.

Hypothesis

A large language model can accurately identify radiology reports that meet Act 112 notification criteria, independent of macro usage.

Methods

We collected 1,000 abdominal imaging reports from our radiology information system: 500 reports with documented follow-up recommendations within 90 days and 500 with either no follow-up or recommendations beyond 90 days, based on structured macro documentation. We developed a system using Azure OpenAI GPT-4, ensemble prompting and universal self-consistency techniques, to analyze report text, which was stripped of the structured macros, and classify whether each case met Act 112 notification criteria. The model's classifications were compared against the ground truth established by the structured macro documentation.

Results

The LLM achieved an F1-score of 0.72, driven by a relatively higher precision (83%) compared to recall (64%) (Figure 1). Notably, the model's follow-up recommendation rate did not vary significantly based on the actual follow-up intervals specified in the macro (Figure 2).

Conclusion

Our findings demonstrate that GPT-4-based large language models can effectively identify radiology reports requiring Act 112 follow-up notification, achieving high precision. However, discrepancies in recall suggest opportunities for refinement.

Automating this process with LLMs could enhance compliance with local policies. Future efforts will focus on optimizing sensitivity and validating the approach across other subspecialties and report types.

Figure(s)

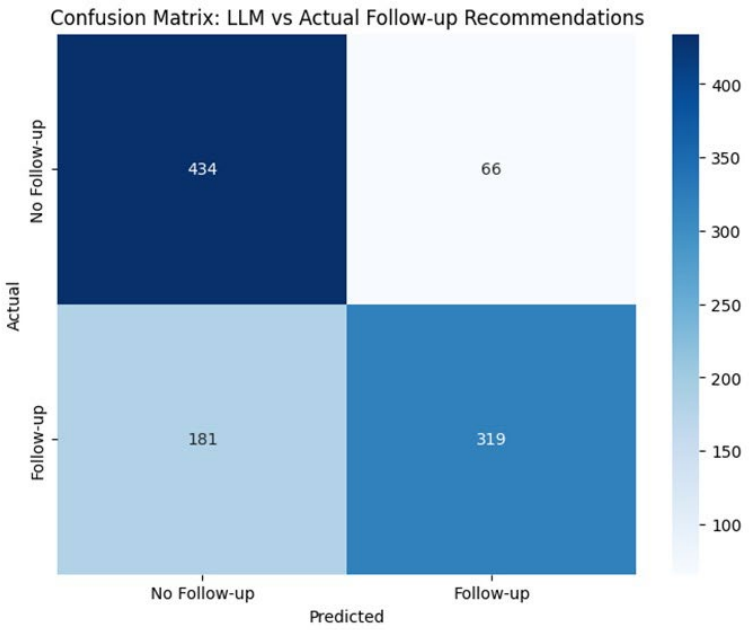

Figure 1. Figure 1. Confusion Matrix of LLM Performance in Identifying Act 112 Follow-up Cases. The LLM demonstrated an overall F1-score of 0.72 which was driven by a relatively higher precision (0.83) compared to recall (0.64).

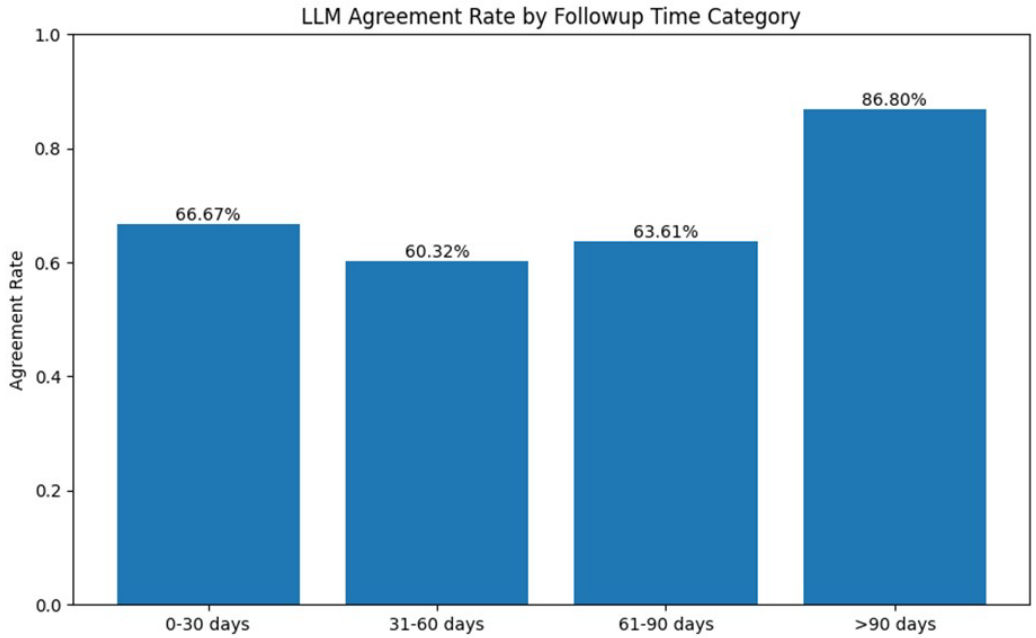

Figure 2. Figure 2. Bar chart showing the agreement rate between LLM predictions and actual follow-up recommendations across different time intervals. The LLM shows highest agreement (86.80%) for cases requiring follow-up >90 days, while performing less consistently for shorter follow-up intervals (60.32-66.67%) without significant difference between the time intervals requiring follow-up notification.

Keywords

Administration & Operations; Applications; Artificial Intelligence/Machine Learning; Emerging Technologies; Quality Improvement & Quality Assurance