View all Parexel's posters at ISPOR 2025

Large Language Model (LLM) as Catalyst for Rapid and Efficient Critical Appraisal of Epidemiological Research

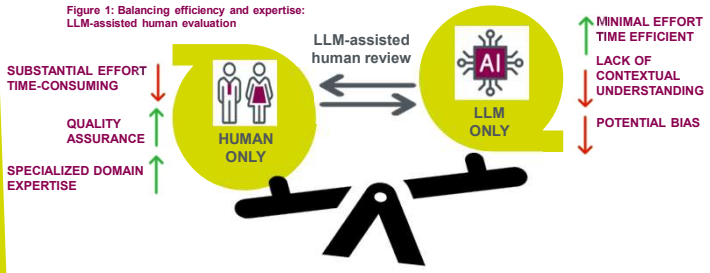
#EPH156

S. Budhia[1], G.S. Mangat[2], A. Jain[2], and S. Sharma[3]
Parexel International, [1]London, United Kingdom; [2]Mohali, India; 3Chandigarh, India

Background

- Artificial intelligence (AI) is advancing rapidly in healthcare due to its ability to analyze vast amounts of data and provide insights that support evidence-based decision-making. More recently, LLMs have shown substantial promise due to their ability to learn and adapt to various linguistic patterns without extensive specialized training. Furthermore, they have been applied to literature reviews to screen scientific articles and extract information.
- Despite these advances, research on LLMs' capabilities in reading and critically appraising scientific papers is limited. Appraising and analyzing scientific articles is a challenging and time-consuming task for researchers, particularly when the articles are lengthy and complex, and the volume of literature is substantial. This poses a significant barrier to efficiently extracting insights and building confidence in the scientific data presented.
- The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist provides a standardized framework for evaluating key elements of observational studies and sufficient information for critical evaluation of epidemiological data. This guideline is comprised of 22 items that authors must adhere to before submitting their manuscripts for publication. While comprehensive, using STROBE as a critical appraisal tool requires substantial time investment from human reviewers when conducting epidemiological literature reviews to understand the burden of disease, natural history, disease surveillance, unmet need, etc.
- As the volume and complexity of epidemiological research continue to expand, the traditional (human reviewer only) approach to critical appraisal becomes increasingly unsustainable (Figure 1).
  - While AI significantly reduces appraisal time and offers an efficient alternative to traditional methods, it introduces unique challenges in the lack of contextual understanding, inability to make nuanced interpretations, and potential bias. A hybrid approach—combining AI's speed with human experts' critical judgment—creates an optimal workflow that maximizes efficiency while maintaining the quality standards essential for reliable critical appraisal (Figure 1).

Figure 1: Balancing efficiency and expertise: LLM-assisted human evaluation



- In this study, we evaluated the use of LLM-assisted human evaluation vs. human review only to assess the quality of epidemiological literature using the STROBE checklist.

Objectives

- To compare the performance of LLM-assisted human review vs. human review only in applying the STROBE checklist for epidemiological studies:
  - To evaluate whether LLMs can efficiently reduce the workload of human reviewers in epidemiological study assessment.
  - To assess the accuracy (alignment) between LLM-assisted reviews and human-only reviews.
  - To evaluate the completeness and thoroughness of LLM appraisals across the 22 STROBE checklist items.
  - To identify specific STROBE checklist items where LLMs perform consistently well or struggle.

Methodology

Design of the study

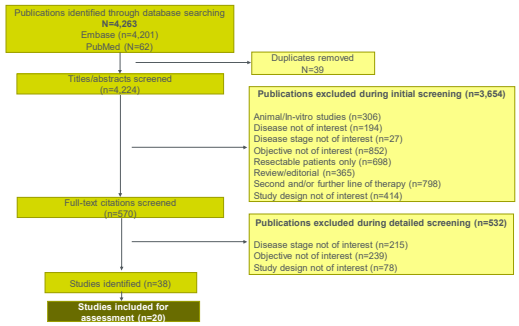
- This study uses a methodological research design to evaluate the comprehension capabilities of an LLM tool using the STROBE checklist.

Selection of studies

- We conducted a targeted literature search using both EMBASE and PubMed databases. For EMBASE, we utilized Emtree terminology, while for PubMed, we employed MeSH terms; both searches were supplemented with free text terms to ensure comprehensive retrieval. Our population of interest comprised adult patients with locally advanced unresectable esophageal cancer, with a global geographic scope.
- The search strategy combined indexed and free terminology for disease ("esophageal cancer" OR "esophageal neoplasm" OR "gastroesophageal junction cancer") with terms indicating disease stage ("stage 2" OR "stage 3" OR "advanced" OR "locally advanced" OR "non-metastatic" OR "unresectable" OR "inoperable"). We further refined our search by incorporating epidemiological terms ("epidemiology," "incidence," OR "prevalence") and observational study design filters. To enhance search sensitivity and precision, we employed various synonyms and proximity operators.
- Following completion of our targeted review, we identified 38 relevant studies, from which we selected 20 for our STROBE checklist assessment comparing LLM-assisted human review vs. human review only (Figure 2).
  - We prioritized studies published as journal articles with clearly defined epidemiological objectives, comprehensive reporting, adequate sample sizes, etc. Additionally, we favored more recent publications to ensure the assessment reflected current epidemiological methods and reporting practices. This structured selection process ensured that our comparative evaluation of LLM-assisted human review vs. human review only was conducted using high-quality epidemiological research representative of current standards in the field (Figure 2).

Methodology….

Figure 2: Flow of citations



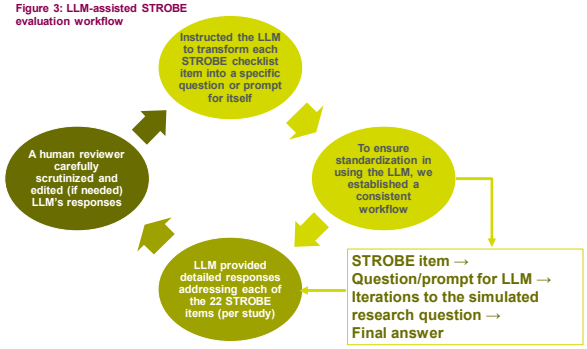
Benchmark development (Human reviewer only)

- An experienced human reviewer conducted a comprehensive critical appraisal of all 20 studies included in the analysis. This methodological approach involved a highly qualified individual with 8 years of advanced expertise in literature review methodology, statistical analysis, and epidemiological methods, who independently evaluated each study against the established STROBE assessment criteria. This provided a robust and reliable reference standard against which the LLM-assisted human evaluation responses were compared.

LLM-assisted human review for comparison

- The prompts were designed to elicit comprehensive analyses from LLM, rather than simple yes/no answers. The prompt package included appraisal guidance, a response template, and study documents for quality assessment. Outcomes measured were accuracy (LLM-assisted human evaluation vs. human-only alignment), completeness (thoroughness of item appraisal), and time efficiency.

Figure 3: LLM-assisted STROBE evaluation workflow



Results

Table 1: STROBE criteria compliance heat map

STROBE Criteria	Compliance Level
1: Title and abstract	100%
2: Background/rationale	100%
3: Objectives/hypotheses	100%
4: Study design elements	100%
5: Setting/dates	100%
6: Participant selection	100%
7: Variable definitions	100%
8: Data sources/measurement	100%
9: Bias addressing	90%
10: Study size rationale	100%
11: Quantitative variables handling	40%
12: Statistical methods compliance assessment	60%
13: Participant flow	90%
14: Descriptive data reporting matrix	100%
15: Outcome events reporting	100%
16: Results quantification matrix	90%
17: Other analyses	100%
18: Key results summary	100%
19: Limitations discussion	100%
20: Cautious interpretation	100%
21: Generalizability	0%
22: Funding source and role	80%

Green: 100% compliant
Yellow: 80-90% compliant
Purple: 40-60% compliant
Red: 0% compliant

Compliance level calculated across all 20 studies

- LLM-assisted human review significantly reduced mean appraisal time to 14 minutes per study compared to 27 minutes for human reviewers alone—a substantial 48% reduction.
- High compliance was achieved between methods, with LLM-assisted human review and human-only review demonstrating ~88% agreement across all STROBE checklist items. This level of agreement suggests that LLM assistance maintains evaluation quality while delivering significant time savings.
- The LLM consistently struggled with generalizability assessments, an area requiring nuanced human judgment. This limitation highlights the continued importance of human expertise for evaluating external validity and contextual applicability of research findings to broader populations.
- Performance limitations were observed for items 11 (quantitative variables handling) and 12 (statistical methodology), where the LLM provided partial responses requiring additional prompting. These items involve complex methodological considerations that appear to challenge current LLM capabilities without supplementary guidance.

Conclusion

- LLMs show promising capability to expedite critical appraisal processes but cannot replace human expertise. We advocate for a two-tiered approach: deploying LLMs as preliminary reviewers, followed by essential human expert engagement on complex interpretative elements. This workflow strategically conserves time while directing valuable human expertise toward the more nuanced and contextual aspects of literature evaluation. Importantly, the human reviewers in this workflow must possess specific domain knowledge and critical appraisal skills to validate, contextualize, and refine the LLM outputs effectively.