Are Artificial Intelligence-Assisted Evidence Reviews Health Technology Assessment-Ready? A Targeted Literature Review Using the ELEVATE-GenAI Quality Framework

MSR149

Sheena Singh,[1] Wenxi Tang,[2] Andrew Easton,[1] Lin Zhan[2]

[1]Cytel, London, United Kingdom; [2]BeOne Medicines, San Carlos, CA, USA

CONCLUSIONS

- AI-assisted evidence synthesis appears capable of supporting HEOR workflows, but most published applications have not yet achieved reporting standards that would support routine HTA use
- Critical gaps remain in calibration and uncertainty, deployment and efficiency, fairness and bias, robustness testing, and security and privacy
- Adoption of ELEVATE-GenAI may provide a practical pathway toward regulatory-grade reporting and improve confidence in AI-enabled evidence generation

INTRODUCTION

- Large language models (LLMs) are increasingly applied to evidence synthesis for health technology assessment (HTA)
- HTA decisions require transparency, reproducibility, and methodological rigor
- LLM use should be evidence-based, proportionate, and subject to human oversight to align with HTA principles
- Reporting standards are therefore critical if artificial intelligence (AI)-generated evidence is to support reimbursement decisions
- Whether current applications meet regulatory-grade reporting standards remains uncertain
- The ELEVATE-GenAI checklist provides a practical framework for reporting LLM use in health economics and outcomes research (HEOR) (**Table 1**)[1]
- Unlike broader AI reporting guidelines, ELEVATE-GenAI focuses on issues particularly relevant to evidence synthesis, HTA, and decision-making research
- It was developed by an ISPOR Working Group on Generative AI through a literature review and expert consensus process
- Improved adherence to the framework outlined in ELEVATE-GenAI may enhance reporting transparency, reproducibility, and reliability
- The objective of this study is to characterize the use of AI in oncology HEOR through a targeted literature review (TLR) and to evaluate the HTA readiness of LLM applications using the ELEVATE-GenAI framework

Table 1. Shortened Version of the ELEVATE-GenAI Checklist

Domain	Key Consideration
Model characteristics	Is the model clearly described (version, access, training)?
Accuracy	Are outputs validated against human or reference standards?
Comprehensiveness	Are all relevant elements of the task addressed?
Factuality	Are outputs fact-checked and source verified?
Reproducibility & generalizability	Can methods be replicated and applied elsewhere?
Robustness	Does performance remain stable under varying inputs?
Fairness & bias	Are potential demographic or representational biases assessed?
Deployment & efficiency	Are implementation requirements and performance reported?
Calibration & uncertainty	Is uncertainty quantified and managed appropriately?
Security & privacy	Are data protection, governance, and compliance addressed?

METHODS

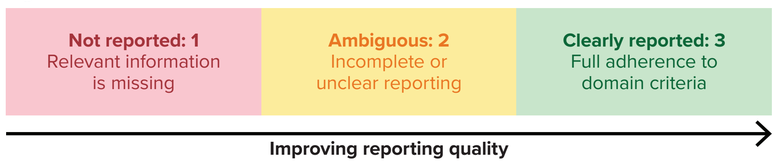
- Searches were conducted from January 2023 to December 2025 and included structured searches of MEDLINE and Embase
- Supplementary searches of key oncology and HTA conferences from 2023 to 2025, and of the websites of 17 HTA bodies in Asia, Europe, Oceania, and North America were conducted to identify additional information
- Participants, Concept, Context (PCC) eligibility criteria were used to identify studies evaluating AI applications in oncology evidence synthesis (**Table 2**)
- The ELEVATE-GenAI checklist was applied to eligible full-text publications
- A 3-point scale was used to assess the reporting quality of each of the 10 domains: clearly reported (CR, 3 points), ambiguous (2 points), and not reported (NR, 1 point) (**Figure 1**)
  - These scores reflect reporting completeness, not study quality or AI performance

Table 2. PCC Eligibility Criteria

Domain	Inclusion Criteria	Exclusion Criteria
Participants	No restriction	NA
Concept	Published data reporting the use of AI across any HEOR domains in oncology. Data may report advantages or disadvantages or performance metrics (eg, accuracy, sensitivity, scalability). Data may be qualitative or quantitative.	Data relating to HEOR activities not involving AI-assisted technology.
Context	Use of AI in HEOR functions/domains, ie, evidence synthesis/reviews, medical writing, real-world evidence, comparative effectiveness, HTA submission (including EU-JCA), market access, economic evaluations, reimbursement, value dossiers.	Data relating to AI-assisted technology in functions other than HEOR, eg, cancer screening or detection.
Types of Evidence Source	• Peer reviewed articles and abstracts • Conference abstracts • Annual reports and other reports of HTA bodies • Hand searches of bibliographies of relevant studies, reviews, and commentaries • Websites of AI-assisted tools and technologies	Other study designs, eg, clinical trials, case reports.

Abbreviations: AI, artificial intelligence; EU-JA, European Union Joint Clinical Assessment; HEOR, health economics and outcomes research; HTA, health technology assessment; NA, not applicable; PCC, Participants, Concept, Context.

Figure 1. ELEVATE-GenAI Assessment Scale

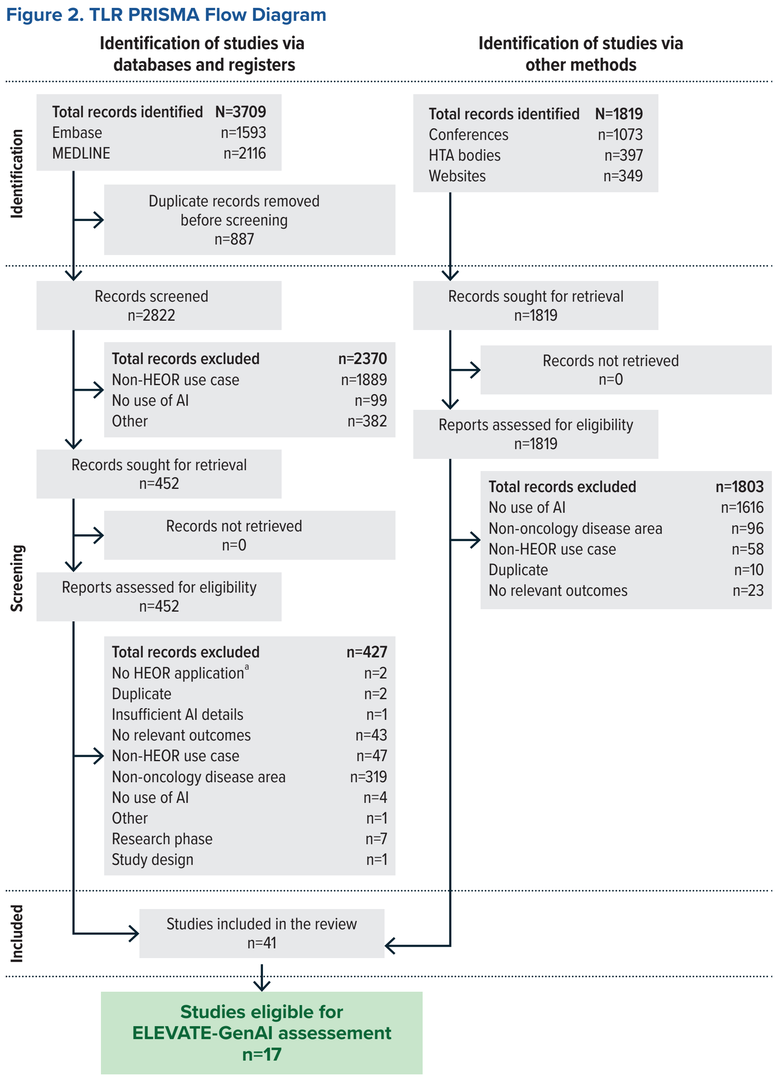


Note: Scores reflect reporting completeness, not study quality or AI performance.

RESULTS

- In total, 41 oncology evidence synthesis studies were identified. Of these, 17 LLM-based studies were eligible for ELEVATE-GenAI assessment (**Figure 2**)

Figure 2. TLR PRISMA Flow Diagram

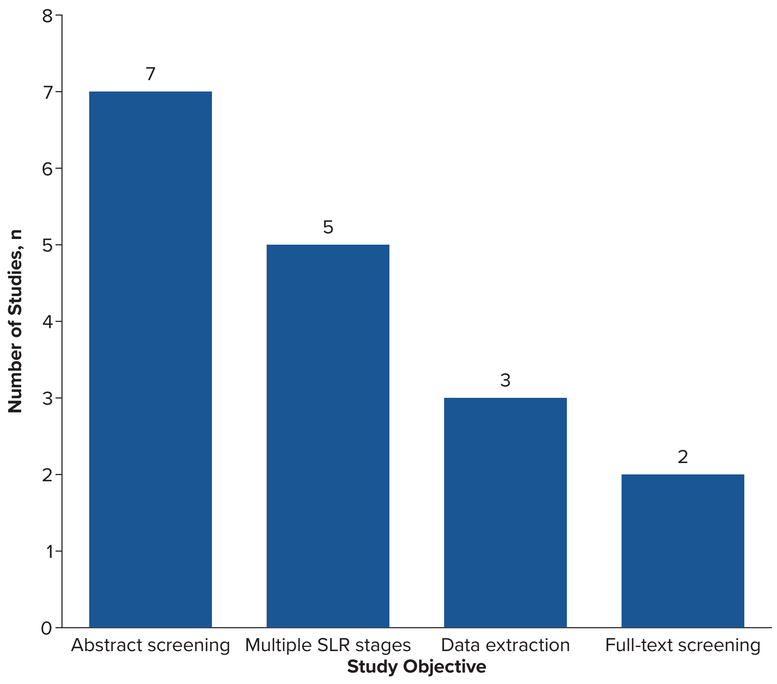


[a]Conceptual/policy only.
Abbreviations: AI, artificial intelligence; HEOR, health economics and outcomes research; HTA, health technology assessment; TLR, targeted literature review.

Study Characteristics

- Most studies were published in 2025 (14/17), with the remaining studies published in 2023, 2024, and 2026
- Studies represented a global evidence base, spanning North America (7/17), Europe (4/17), Asia (4/17), Oceania (1/17), and multinational collaborations (1/17)
- All studies evaluated AI-assisted literature review workflows, with abstract screening being the most frequently assessed task (**Figure 3**). Several studies also evaluated AI across multiple review stages, including full-text screening, data extraction, and evidence synthesis
- All studies benchmarked AI performance against human reviewers or human-generated reference standards

Figure 3. Tasks Evaluated in Included Studies



Abbreviation: SLR, systematic literature review.

ELEVATE-GenAI Score Summary

- Overall scores ranged from 50% to 83%, with a median of 75% (**Figure 4**)
- Accuracy assessment was the best-reported domain, clearly reported in 88% of studies (**Figure 5**)
- Model characteristics, factuality verification, and reproducibility were generally well reported
- Major reporting gaps were observed for calibration and uncertainty, deployment and efficiency, fairness and bias, robustness testing, and security and privacy
- Fairness and bias monitoring were not reported in any study
- Examples of common reporting deficiencies and corresponding best-practice reporting recommendations across these domains are presented in **Table 3**
- The potential implications of key reporting gaps for HTA decision-making and evidence generation are summarized in **Table 4**, highlighting how inadequate reporting may affect transparency, reliability, reproducibility, and confidence in AI-assisted research findings

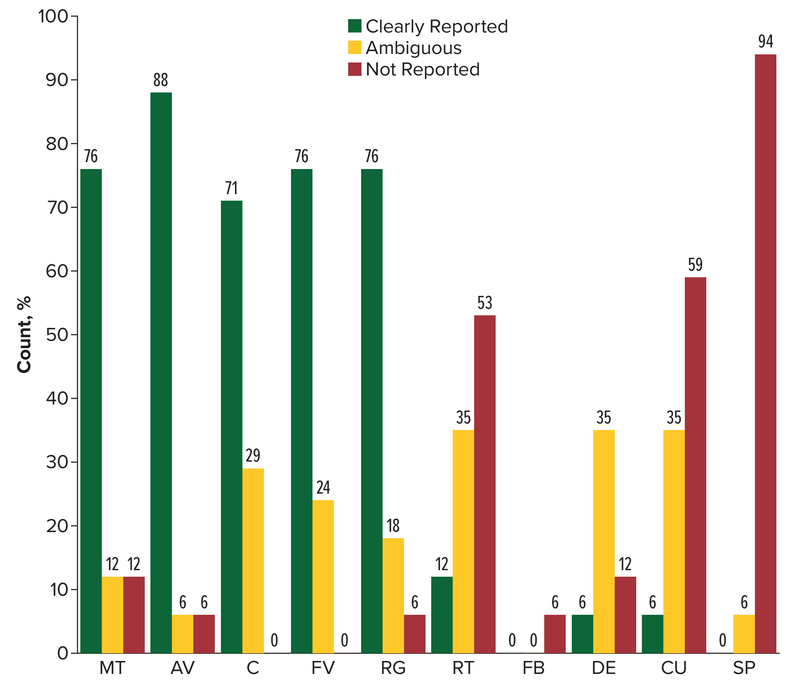
Figure 4. Heat Map of ELEVATE-GenAI Scores

Study No.	MT	Av	C	FV	RG	RT	FB	DE	CU	SP	Overall score, %
1	CR	CR	CR	CR	CR	A	NA	NA		NR	83
2	CR	CR	CR	CR	CR	A	NA	NA	NR	NR	79
3	CR	CR	A	CR	CR	CR	NA	NA	NR	NR	79
4	CR	CR	CR	A	CR	CR	NA	NA	NR	NR	79
5	CR	CR	CR	CR	CR	A	NA	NA	NR	NR	79
6	CR	CR	CR	CR	CR	A	NA	A	NR	NR	78
7	CR	CR	CR	CR	CR	NR	NA	A	NR	A	78
8	CR	CR	A	CR	CR	NR	NA	NA	A	NR	75
9	CR	CR	CR	A	CR	NR	NA	NA	A	NR	75
10	CR	CR	CR	CR	A	NR	NA	A	A	NR	74
11	CR	CR	CR	CR	CR	NR	NA	A	NR	NR	74
12	CR	CR	A	A	CR	A	NA	CR	NR	NR	74
13	NR	CR	CR	CR	A	NR	NA	A	CR	NR	70
14	A	CR	A	CR	CR	A	NA	NR	A	NR	70
15	CR		CR	CR	CR	NR	NR	A	NR	NR	67
16	NR	CR	CR	A	A	NR	NA	NR	A	NR	59
17	A	NR	A	CR	NR	NR	NA	NA	NR	NR	50

Overall Score Shading ≥80% 60-79% <60%

Abbreviations: A, ambiguous; AV, accuracy & validation; C, comprehensiveness; CR, clearly reported; CU, calibration & uncertainty; DE, deployment & efficiency; FB, fairness & bias; FV, factuality verification; MT, model & data transparency; NA, not applicable; NR, not reported; RG, reproducibility & generalizability; RT, robustness testing; SP, security & privacy.

Figure 5. Proportion of ELEVATE-GenAI Domains Clearly Reported



Abbreviations: AV, accuracy & validation; C, comprehensiveness; CU, calibration & uncertainty; DE, deployment & efficiency; FB, fairness & bias; FV, factuality verification; MT, model & data transparency; NA, not applicable; RG, reproducibility & generalizability; RT, robustness testing; SP, security & privacy.

Table 3. What Does Good Reporting Look Like?

Domain	Poor Reporting Example	Good Reporting Should Include
Security & privacy	"ChatGPT was used"	Model, data handling, and governance clearly described
Calibration & uncertainty	Accuracy only reported	Confidence intervals and uncertainty management reported
Deployment & efficiency	"Workflow was faster"	Time savings and implementation requirements quantified
Robustness testing	Single performance estimate	Performance tested across multiple scenarios
Fairness & bias	No bias assessment	Bias evaluated across populations, studies, or settings

Abbreviation: ChatGPT, Chat Generative Pre-trained Transformer.

Table 4. Why Do These Reporting Gaps Limit HTA Readiness?

Reporting Gap	Potential HTA Concern
Security & privacy	HTA workflows may involve sensitive or confidential information. Poor reporting of data governance and privacy safeguards may limit trust and adoption.
Calibration & uncertainty	HTA decisions require an understanding of uncertainty. Without reporting model confidence or uncertainty management, the reliability of AI-generated outputs is difficult to assess.
Deployment & efficiency	Efficiency claims are difficult to evaluate without implementation details. Transparent reporting helps decision-makers assess real-world feasibility and value.
Robustness testing	Evidence generation methods should produce consistent results across different settings and inputs. Limited robustness testing reduces confidence in reproducibility and generalizability.
Fairness & bias	Unassessed bias may lead to systematic errors or incomplete evidence identification. This creates uncertainty regarding the completeness and validity of the evidence base.

Abbreviation: HTA, health technology assessment.

REFERENCES

1. Fleurence RL, Dawoud D, Bian J, et al. ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report. *Value Health.* 2025;28(11):1611-1625. doi:10.1016/j.jval.2025.06.018.

DISCLOSURES

SS, AE: employees of Cytel, Ltd, which received consulting fees from BeOne Medicines, Ltd.
WT, LZ: employees and equity holders in BeOne Medicines, Ltd.

ACKNOWLEDGMENTS

This study was sponsored by BeOne Medicines, Ltd. Editorial assistance was provided by Elizabeth Hermans, PhD, and supported by BeOne Medicines.

CORRESPONDENCE: Wenxi Tang, wenxi.tang@beonemed.com

Presented at the ISPOR EU 2026 Annual Meeting; November 8-11, 2026; Vienna, Austria