ELEVATE GEN-AI FRAMEWORK-BASED INTERIM EVALUATION OF THE MADEAI PLATFORM FOR LITERATURE SCREENING OF A CLASS II MEDICAL DEVICE

ISPOR 2026

MadeAi

SA52

AUTHOR(S)
Viji Queen, PharmD[2] | **Revanth M,** B.E[2] | **George Alisha,** B.E.[2] | **Swathirajan C R,** Ph.D[2] | **Angeline Dhas,** BS[1] | **Vince Salerno,** PharmD, RPh[1]
[1]MadeAi, Cambridge, MA, USA, [2]MadeAi, Nagercoil, India.

Background

Systematic literature reviews (SLRs) are essential for regulatory evidence generation, where rigorous methods, traceability, and timely delivery are critical. However, conventional title and abstract screening is labor-intensive, time-consuming, and prone to reviewer variability, particularly for large evidence datasets. Generative artificial intelligence (GenAI) offers the potential to improve screening efficiency and consistency within evidence synthesis workflows. However, adoption in regulated settings requires validation beyond traditional accuracy metrics alone. The ELEVATE-GenAI framework provides a structured, multidomain approach to assess GenAI systems across key areas such as accuracy, comprehensiveness, factuality, reproducibility, and operational readiness. Applying this framework to SLR platforms may support reliable adoption of GenAI in regulated evidence synthesis.

Objectives

- Conduct an interim evaluation of MadeAi, a GenAI-enabled literature screening platform
- Apply the ELEVATE-GenAI framework across performance, comprehensiveness, and factuality domains
- Assess screening-stage behavior using a Class II medical device golden dataset

Dataset: 2,302 Records → Human Adjudication: Golden Standard → GenAI Screening (MadeAi): AI screening results → Domain-Based Evaluation: Accuracy, Comprehensiveness, Factuality

Methodology

Dataset:

2,302 records | Class II medical device | Curated and adjudicated by subject matter experts | Title and abstract screening level

Evaluation Framework:

Screening outputs were assessed using three ELEVATE-GenAI domains:

Accuracy | Comprehensiveness | Factuality

Human-adjudicated decisions served as the golden reference standard.

DOMAIN	EVALUATION FOCUS	METRIC(S) USED
Accuracy	Inclusion and exclusion classification	Recall, Precision, F1
Comprehensiveness	Retention of relevant evidence	Overall Article Relevance
Factuality	Alignment of exclusion rationale	Concordance With Human Primary Reasons

Results

ACCURACY

INCLUSION DECISIONS
- **RECALL: 83%**
- **PRECISION: 80.7%**
- **F1 SCORE: 81.9%**

Interpretation: The platform demonstrated strong sensitivity for identifying relevant studies at the title and abstract stage. The balance between recall and precision suggests stable classification behavior without extreme over-inclusion. This indicates reliable detection of potentially eligible evidence while maintaining reasonable decision specificity.

EXCLUSION DECISIONS
- **RECALL: 96%**
- **PRECISION: 96%**
- **F1 SCORE: 95.9%**

Interpretation: Exclusion performance showed highly consistent identification of non-relevant studies. The higher exclusion recall compared to inclusion recall reflects conservative screening behavior, prioritizing retention of potentially relevant evidence over premature exclusion. In regulated evidence synthesis, such conservative behavior reduces the risk of missing relevant studies at early screening stages.

COMPREHENSIVENESS

OVERALL ARTICLE RELEVANCE: 93%

Interpretation: Ninety-three percent of articles identified as relevant were aligned with the human-adjudicated golden dataset, indicating strong coverage of relevant evidence. This supports screening-stage comprehensiveness and suggests that evidence capture was not substantially compromised by automation.

FACTUALITY

CONCORDANCE BETWEEN AI EXCLUSION DECISIONS AND HUMAN PRIMARY EXCLUSION RATIONALES: 84%

Interpretation: Eighty-four percent alignment between AI-generated exclusion reasoning and human primary exclusion categories demonstrates substantial agreement in decision logic. This domain-level assessment extends beyond classification accuracy to evaluate reasoning alignment, which is critical for transparency and auditability in regulated workflows.

Domain-Level Summary

Across the accuracy, comprehensiveness, and factuality domains, the evaluation indicates consistently strong system behavior. Classification accuracy was high and stable across inclusion and exclusion decisions, evidence coverage remained substantial at the screening stage, and exclusion reasoning demonstrated meaningful alignment with adjudicated human standards. Collectively, these findings support the value of structured, multi-domain validation for GenAI-assisted literature screening within regulated HEOR contexts.

Conclusion

This interim evaluation demonstrates that applying the ELEVATE-GenAI framework enables structured, domain-based assessment of GenAI-assisted literature screening within a regulated HEOR context. Beyond classification accuracy, inclusion of comprehensiveness and reasoning concordance provides a more interpretable and transparent characterization of system behavior. Domain-level reporting supports standardized evaluation of GenAI tools intended for regulated evidence synthesis workflows.