From Data to Insights: Validating Artificial Intelligence (AI)-Generated Writing in Evidence Synthesis

MSR164

Allie Cichewicz, MSc
[1]Independent Consultant, Boston, MA, USA

Introduction

- Generative AI is increasingly integrated into systematic review software, progressing from structured tasks like screening and data extraction toward generating written summaries.
- Automated written synthesis remains a key unmet capability, and this shift to narrative output introduces new risks around accuracy, completeness, and misrepresentation of source findings that require formal validation methods.
- The RAISE (Responsible AI-Supported Evidence synthesis) framework provides guidance on transparent reporting of AI use in evidence synthesis, but practical criteria for evaluating the quality of AI-generated text against source literature are still lacking.

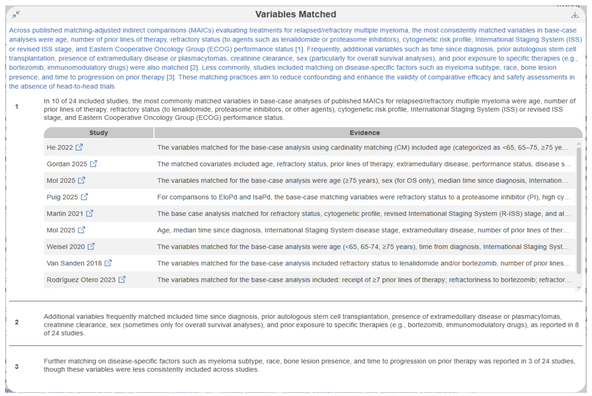
Objective

To evaluate the quality and reliability of Smart Insights, an AI-generated narrative synthesis tool in Nested Knowledge, using a structured framework for assessing scientific writing produced by AI.

Methods

- Data were extracted for three oncology reviews using Adaptive Smart Tags in Nested Knowledge: clinical effectiveness from real-world evidence (RWE; 24 studies); clinical efficacy and safety from randomized trials (RCTs; 21 studies); comparative effectiveness from matching-adjusted indirect comparisons (MAICs; 24 studies).
- For all tagged data (e.g., study characteristics, patient characteristics, outcomes, conclusions), AI-generated summaries and supporting claims were produced via Smart Insights, with each summary and claim linked back to the citation and evidence for traceability (**Figure 1**).
- Insights for each tag were evaluated across six domains (**Table 1**) with score for each domain ranging from 1 (poor) to 5 (excellent). Scores for each insight were averaged across studies for each domain to obtain a total average per domain.
- To assess consistency, Smart Insights were generated twice on the same datasets; outputs were compared qualitatively.

Figure 1. Example Smart Insight with Supporting Claims and Direct Evidence from Underlying Studies



Methods (cont'd)

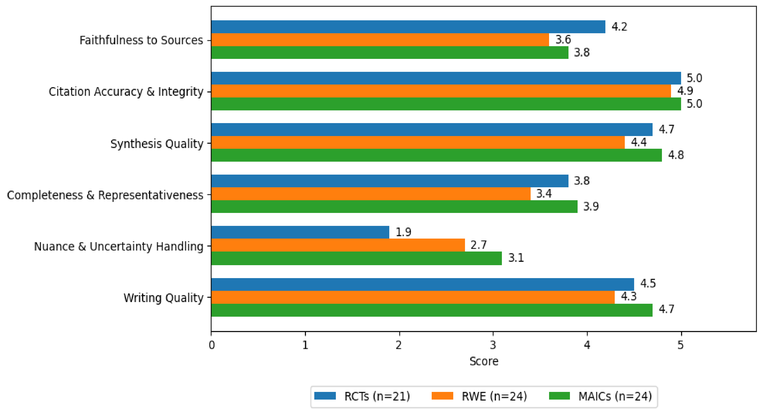
Table 1. Domains for Evaluating AI-Generated Synthesis

Domain	Description
Faithfulness to Sources	Factual claims are accurate compared with the original cited sources without fabricated or contradictory material.
Citation Accuracy & Integrity	All citations are real and match the supporting claims without fabricated or misattributed references.
Synthesis Quality	Themes, comparisons, gaps, etc. are synthesized across included studies without paper-by-paper summaries.
Completeness & Representativeness	All major themes, key findings, and seminal papers are included without critical omissions or skewing results.
Nuance & Uncertainty Handling	Explicitly expresses uncertainty in results and limitations of the findings without overconfidence or false consensus.
Writing Quality	Provides clear, coherent scientific writing.

Results

- Overall performance of Smart Insights was moderate to high across study designs. Syntheses of published MAIC findings generally score highest across most domains, followed by RCTs, with RWE slightly lower on average.
- Total scores were 3.9/5 for RWE, 4.0 for RCTs, and 4.2 for MAICs.
- As shown in **Figure 2**, consistently strong scores were observed for Citation Accuracy & Integrity (4.9 to 5.0), Synthesis Quality (4.4 to 4.8), and Writing Quality (4.3 to 4.7).
- Insights performance was generally poor on Nuance & Uncertainty Handing, ranging from 1.9 to 3.1.
- Repeated Insights generations produced minimal, non-meaningful differences, though greater textual variation was observed for text-driven fields (e.g., study limitations or author conclusions).

Figure 2. Mean Scores Across Evaluation Domains by Study Design



Conclusions & Limitations

- AI-generated written synthesis within Nested Knowledge demonstrates moderate to high performance, with strongest results in citation accuracy and synthesis quality, supporting its use as an efficiency-enhancing tool in evidence synthesis with appropriate safeguards.
- Data-driven statements appeared more accurate than qualitative, text-based summaries, though not formally evaluated; outputs did not consistently identify evidence gaps or limited support, and weaker handling of nuance and uncertainty suggests it may not be suitable for fully robust systematic reviews without structured validation and human oversight.
- Variation in synthesis by format or data type was not assessed (e.g., text vs tables, qualitative vs quantitative), limiting interpretation and representing an area for future research.
- No validated framework exists for evaluating AI-generated synthesis, requiring custom criteria and limiting comparability; standardization would improve reproducibility and benchmarking.

Disclosures: Author is a contractor of Nested Knowledge, and this study was conducted using the Nested Knowledge platform.

Presented at ISPOR 2026 - Philadelphia