AI-assisted Qualitative Data Extraction and Evidence Mapping in an Umbrella Review

Sheena Singh,[1] Tirza Boyle,[2] Linda Kalilani[3]
[1]Cytel, Inc., London, UK; [2]GSK, Philadelphia, PA, USA; [3]GSK, Durham, NC, USA

Background

- Evidence synthesis approaches such as targeted literature reviews, scoping reviews, and umbrella reviews are methodologically rigorous but typically lower risk than health technology assessments or regulatory submissions, making them strong candidates for artificial intelligence (AI)-assisted workflows.
- Such applications can also help build confidence for broader adoption in more complex evidence synthesis settings..
- When combined with structured human quality control (QC), AI-enabled processes have the potential to enhance efficiency across evidence identification, extraction, and visualization without compromising accuracy or reproducibility.

Objectives

- The aim of this study was to evaluate the feasibility and performance of an AI-assisted workflow for qualitative data extraction and evidence mapping within an umbrella review of safety outcomes in solid tumors.

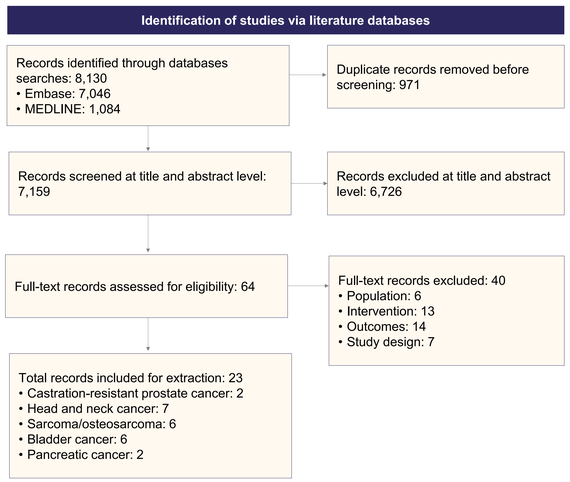
Methods

- We conducted an umbrella review assessing safety outcomes in solid tumors using an AI-assisted workflow within Nested Knowledge.
- Custom extraction tags were configured across key domains including study design, population, interventions, inclusion/exclusion criteria, and safety outcomes with each tag defining the response format (free-text, numeric, or binary yes/no) and a natural language description to guide AI extraction.
- Performance was evaluated across extraction rate, accuracy, precision, and time savings versus a fully manual process.

Results

- In total, 23 studies were identified for inclusion in the umbrella review (Figure 1).

Figure 1. PRIOR flow diagram



Results (cont.)

Qualitative Study-Characteristic Extraction

- AI-assisted extraction demonstrated strong overall performance for qualitative study-characteristic data points.
- Of the 230 expected data points across 23 included studies, 210 were successfully extracted (extraction rate: 91.3%), with an overall accuracy of 90.0% when missing extractions were treated as incorrect. (Table 1).

Table 1. Extraction rate and accuracy of AI-assisted extraction for study-characteristic data points across 23 included studies

Study characteristic	Expected (n)	Extracted by AI n (%)	Missing n (%)	Correct data, (n)	Accuracy (%)	Precision (%)
Study (SLR) objectives	23	18 (78.0%)	5 (22.0%)	18	78.0%	100.0%
Country/region	23	14 (61.0%)	9 (39.0%)	11	48.0%	78.6%
SLR search period	23	23 (100%)	0 (0%)	23	100%	100.0%
Number of included studies	23	22 (96.0%)	1 (4.0%)	21	95.0%	95.5%
Study design	23	22 (96.0%)	1 (4.0%)	22	96.0%	100.0%
Brief population description	23	21 (91.0%)	2 (9.0%)	21	91.0%	100.0%
Intervention description	23	22 (96.0%)	1 (4.0%)	22	96.0%	100.0%
Inclusion criteria	23	22 (96.0%)	1 (4.0%)	22	96.0%	100.0%
Exclusion criteria	23	23 (100%)	0 (0%)	23	100%	100.0%
Analysis description	23	23 (100%)	0 (0%)	23	100%	100.0%
Overall	230	210 (91.3%)	20 (8.7%)	206	90.0%	98.1%

Note: Formatting differences (e.g., wording, structure, or order) were not considered extraction errors if the semantic content matched the human reference.
Abbreviations: AI, artificial intelligence; SLR, systematic literature review

- Precision was defined as the proportion of extracted data points that were correct and found to be 98.1%, indicating that when the AI software did extract a data point, it was mostly accurate.
- Low accuracy was therefore primarily attributable to missing extractions (n=20, 8.7%) rather than incorrect content.
- Performance was consistent across most study characteristics, with nine of 10 fields achieving precision of 95.5% to 100.0%.
- The exception was country/region, with the lowest extraction rate (61.0%) and precision (78.6%), likely reflecting variability in how geographic information was reported across the included systematic literature reviews.

Binary Safety Outcome Extraction

- Performance was comparatively lower for binary safety outcome data points, with nuanced findings when distinguishing between accuracy and precision (Table 2).
- Of the 207 expected binary data points, 150 were successfully extracted (extraction rate: 72.5%), yielding an overall accuracy of 62.3% when missing extractions were treated as incorrect.
- However, precision restricted to successfully extracted data points was substantially higher at 86.0%, indicating strong agreement with the human reference when extraction occurred. Missing extractions (n=57, 27.5%) were therefore the primary driver of reduced overall accuracy, rather than incorrect content.
- Performance varied considerably across safety outcome categories: hematological toxicities achieved the highest extraction rate (96.0%) and precision (95.5%), while infusion-related reactions and infections showed the lowest extraction rates (56.0% and 64.0%, respectively).

Table 2. Extraction rate and accuracy of AI-assisted extraction for binary safety outcomes across 23 included studies

Binary outcome	Expected (n)	Extracted by AI n (%)	Missing n (%)	Correct data, (n)	Accuracy (%)	Precision (%)
Hematological toxicities	23	22 (96.0%)	1 (4.0%)	21	91.0%	95.5%
Infections	23	15 (64.0%)	8 (36.0%)	12	52.0%	80.0%
Gastrointestinal disorders	23	20 (88.0%)	3 (12.0%)	18	78.0%	90.0%
Hepatotoxicity	23	15 (64.0%)	8 (36.0%)	11	48.0%	73.3%
Skin reactions	23	20 (88.0%)	3 (12.0%)	17	74.0%	85.0%
Ocular reactions	23	14 (60.0%)	9 (40.0%)	14	61.0%	100.0%
Infusion-related reactions	23	13 (56.0%)	10 (44.0%)	12	52.0%	92.3%
General disorders	23	14 (60.0%)	9 (40.0%)	11	48.0%	78.6%
Renal toxicities	23	17 (76.0%)	6 (24.0%)	13	57.0%	76.5%
Overall	207	150 (72.5%)	57 (27.5%)	129	62.3%	86.0%

Note: Expected data points were defined as one binary safety outcome per included study (n=23).
Abbreviation: AI, artificial intelligence

Sources of Error and Workflow Considerations

- Human QC identified a recurring error whereby AI extracted binary safety outcomes from introduction or discussion sections rather than the results section, reflecting an inherent limitation in contextual judgement relative to human reviewers.
- Tag configuration was a further key determinant of performance: suboptimal or inconsistent tag definitions contributed to variability in extraction completeness and accuracy, and standardization of tag design across the project team is critical to minimize systematic bias and ensure reproducibility.

Workflow Efficiency and Time Savings

- The AI-assisted workflow (Figure 2) delivered substantial efficiency gains at the extraction and evidence mapping stage. When accounting for the full upfront investment including tag hierarchy development, iterative prompt testing, and reconfiguration within Nested Knowledge alongside AI-assisted extraction across all 23 studies, total time for extraction and evidence mapping was reduced by approximately 80% compared with a fully manual process.
- This upfront tag configuration investment is a one-time cost that is expected to diminish as a proportion of overall effort as tag libraries are refined and reused across future projects.
- When 100% human QC coverage is additionally accounted for, the net time saving compared with a fully manual process was approximately 31%. This highlights that while AI-assisted extraction offers meaningful efficiency gains particularly for repetitive, structured data extraction tasks realistic assessment of workflow implementation must account for QC burden alongside upfront configuration costs.
- Overall time savings are therefore expected to increase with experience, as tag libraries mature, QC effort reduces with improved tag precision, and reviewer proficiency in AI-assisted workflows develops.

Figure 2. AI-assisted workflow

Conclusions

- AI-assisted extraction reduced evidence mapping time by approximately 80% versus manual processes, with a 31% net saving after 100% human QC.
- Accuracy was high for qualitative study characteristics (90.0%; precision 98.1%), with errors driven by omission rather than incorrect content.
- Performance was lower for binary safety outcomes (accuracy 62.3%; precision 86.0%), primarily due to missing extractions and extraction from non-results sections, which is addressable through improved tag configuration.
- These findings support AI-assisted extraction as a viable approach for lower-risk evidence synthesis, with efficiency gains expected to increase as tag libraries mature and reviewer proficiency develops.

References

1. Gates M, Gates A, Pieper D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. *BMJ*. 2022;378:e070849. doi:10.1136/bmj-2022-070849; 2. Cochrane Training HJT, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Accessed April 2026, https://training.cochrane.org/handbook.

Disclosures and acknowledgements

TB is an employee of GSK and holds GSK stocks and shares. LK is an employee of GSK and holds GSK stocks and shares. SS is an employee of Cytel, Inc. The authors thank Ritu Shah, Cytel, Inc for her contributions to screening and QC of this review, and Colleen Dumont, Cytel, Inc. for editorial and creative support in the development of this poster.