Evaluating the Performance and Optimal Use Strategy of an Artificial Intelligence-Assisted Tool for Oncology Literature Review

SA3

Ahmed Mostafa Ahmed Kamel,[1]* Joshua T. Harvey,[2] Wenxi Tang,[3] Xiaoliang Wang,[3] Gregory Maglinte,[3] Lin Zhan[3]

[1]Auburn University, Auburn, AL, USA; [2]University of Massachusetts Chan Medical School, Worcester, MA, USA; [3]BeOne Medicines, Ltd, San Carlos, CA, USA

CONCLUSIONS

- AI-assisted literature review tools were highly time-saving, particularly for large reviews with more than 2000 records to screen
- Performance declined when the review question was broad and methodologically complex, such as including all HEOR evidence related to a compound
- Important concepts still require human reinforcement, including missing tags such as cost-effectiveness and manual checking of extracted endpoints
- The best strategy is a hybrid approach: AI for speed and prioritization, with human oversight for search refinement, missing tags, full-text verification, and final data validation

INTRODUCTION

- Systematic literature reviews (SLRs) are central to health economics and outcomes research (HEOR), but they are time- and labor-intensive. Nested Knowledge (NK) provides artificial intelligence (AI)-assisted search, screening, and extraction tools that may accelerate review workflows
- The objective of this study was to evaluate the performance and practical use of NK for oncology HEOR literature review by comparing an AI-only workflow with a manual review workflow, focusing on search yield, screening performance, time savings, and extraction validity
- Two separate oncology literature review workstreams were conducted. The solid tumor workstream focused on HEOR studies of programmed death-ligand 1 (PD-1) inhibitors in advanced or metastatic non-small cell lung cancer (NSCLC) and was used to compare a fully manual PubMed review with an AI-only NK workflow. The hematology workstream focused on clinical outcomes for ibrutinib in blood cancers and was evaluated using an AI-assisted workflow with dual human verification and comparison against Centers for Medicare & Medicaid Services (CMS) Maximum Fair Price (MFP) citations

METHODS

Manual Review

- A systematic PubMed search identified HEOR studies on PD-1 inhibitors in advanced or metastatic NSCLC
- Study types included real-world evidence, cost-effectiveness analysis/budget impact, patient-reported outcomes, SLRs, and indirect treatment comparisons
- Manual title/abstract and full-text screening were followed by data synthesis on indication, publication year, treatment line, and study type
- Inclusion criteria included: HEOR-focused studies in NSCLC involving PD-1 inhibitors
- Exclusion criteria included: non-HEOR studies; non-NSCLC populations; studies not involving PD-1 inhibitors; preclinical/animal studies; biomarker-only, pharmacokinetic/pharmacodynamic-only, or early-phase studies without HEOR outcomes; editorials, protocols, case reports, conference abstracts without full text; non-English full texts

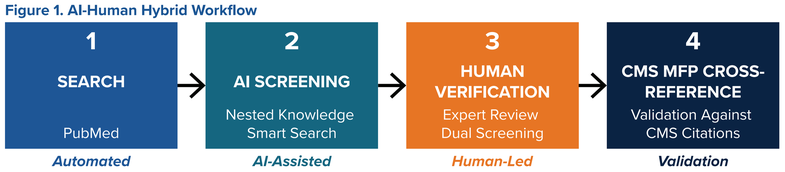
AI Workflow (NK)

- NK Smart Search generated the search strategy and retrieved PubMed records
- Screening models were trained iteratively: 50 manually screened papers with >10 inclusions, then bulk exclusions for irrelevant cancer types, followed by 100 manually screened articles (Model 2) and 200 manually screened articles (Model 1)
- Default NK tags plus additional custom tags were used for extraction; extracted data were exported to Excel for review
- Metrics compared with manual review were time consumed, recall based on final-study overlap, false negatives, false positives, and content validity of AI extraction

AI With Human Oversight

- A targeted literature review was also conducted for clinical outcomes of ibrutinib in blood cancers using PubMed (**Figure 1**)
- NK Smart Search was used to generate the search strategy and retrieve records, followed by AI-assisted screening
- The AI-human hybrid workflow employed automated screening with manual verification **by dual independent reviewers**
- Findings were cross-referenced against the 52 clinical outcome citations in the CMS MFP document to assess concordance

Figure 1. AI-Human Hybrid Workflow



Abbreviations: AI: artificial intelligence; CMS: Centers for Medicare & Medicaid Services; MFP: Maximum Fair Price.

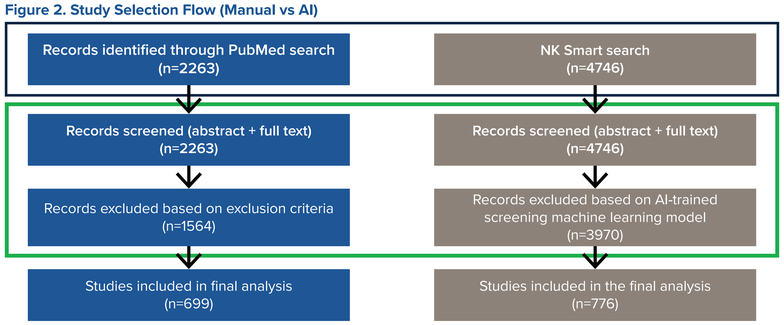
RESULTS

Manual review	2263 screened/699 included
AI review	4746 screened/776 included
Overlap	361 studies (51.6%)
Time	90 hours vs 6 hours

Search and Screening

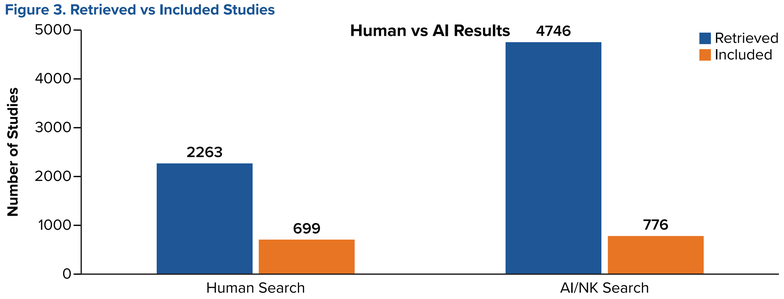
- Manual review screened 2263 records and included 699 studies; 1564 records were excluded based on predefined criteria (**Figures 2** and **3**)
- AI/NK Smart Search retrieved 4746 records, of which 776 were included after machine learning-based screening and 3970 were excluded (**Figures 2** and **3**)

Figure 2. Study Selection Flow (Manual vs AI)



Comparison of records identified, screened, and included using manual PubMed search and AI-assisted NK workflow.
Abbreviations: AI: artificial intelligence; NK: Nested Knowledge

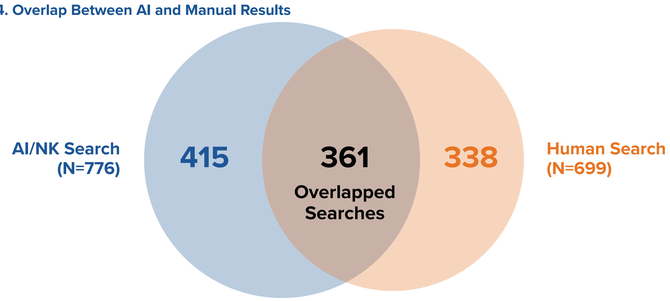
Figure 3. Retrieved vs Included Studies



AI retrieved more studies than manual search, with similar numbers included after screening.
Abbreviations: AI: artificial intelligence; NK: Nested Knowledge.

- A total of 361 studies overlapped between the manual and AI final sets, corresponding to 51.6% overlap among included studies (**Figure 4**)
- The AI workflow yielded 415 false positives (irrelevant studies included by AI) and 338 false negatives (relevant studies missed by AI) (**Figure 4**)
- Total review time decreased from about 3 weeks/90 hours for manual review to about 6 hours using the AI-only workflow

Figure 4. Overlap Between AI and Manual Results



Venn diagram showing overlap (n=361), false positives (n=415), and false negatives (n=338) between workflows.
Abbreviations: AI: artificial intelligence; NK: Nested Knowledge.

Content Validity of AI Extraction

- A subset of the included papers (N=28) were evaluated for extraction of funding source, study type, sample size, overall survival (OS), and progression-free survival (PFS)
- Study type was correctly identified as observational in only eight papers (28.6%); the remaining papers were misclassified as randomized clinical trials, reviews, preclinical studies, or other categories
- Sample size was missed in 14 papers (50%) and correctly captured in the remainder
- Model performance improved with additional training data (200 manually screened articles vs 100) (**Table 1**)

Table 1. Performance of AI Screening Models Across Training Iterations

	History	Records (inc)	Recall	Precision	F1	Accuracy	AUC
Model 1 ➡	**2025-08-05**	2252 (315)	0.98	0.78	0.87	0.96	0.99
Model 2 ➡	**2025-07-28**	1968 (34)	0.83	0.46	0.58	0.98	0.98
	2025-07-28	1960 (27)	0.28	0.61	0.38	0.99	0.91
	2025-07-18	62 (29)	0.97	0.62	0.75	0.69	0.85

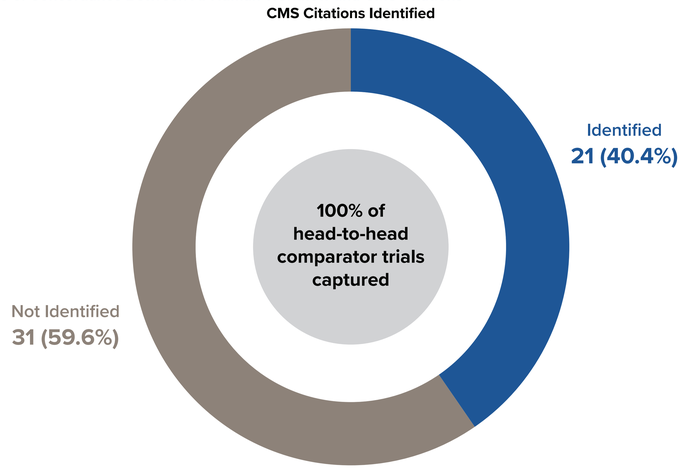
Abbreviation: AUC, area under the curve; inc, included.

- OS was extracted in four of the 20 papers that reported OS (20% accuracy)
- PFS was extracted in two of the 20 papers that reported PFS, and one PFS result was incorrectly classified under OS

AI With Human Oversight

- Of 52 clinical studies cited by the CMS MFP document, the AI-human hybrid review identified 21 (40.4%) (**Figure 5**)
- All key head-to-head trials were captured
- Unmatched citations were predominantly foundational single-agent trials predating the Bruton tyrosine kinase inhibitor era (**Figure 5**)

Figure 6. Concordance Between AI-Human Workflow and CMS Citations



Abbreviation: CMS: Centers for Medicare & Medicaid Services.

Take-Home Message

AI-assisted review can dramatically improve efficiency in oncology HEOR, but search optimization, study selection quality, and data extraction accuracy still depend on structured human oversight. A combined AI + human workflow is the most reliable strategy for timely and decision-relevant evidence generation.

DISCLOSURES

WT, XW, GM, LZ: employees and equity holders at BeOne Medicines, Ltd.

ACKNOWLEDGMENTS

This study was funded by BeOne Medicines, Ltd. Medical writing was supported by BeOne Medicines, Ltd

CORRESPONDENCE: Ahmed Mostaha Ahmed Kamel, amk0147@auburn.edu

Presented at the 2026 ISPOR Annual Conference; May 17-20, 2026; Philadelphia, PA, USA