COMPARING AN EXPERT-IN-THE-LOOP AI WORKFLOW WITH MANUAL SYSTEMATIC LITERATURE REVIEW: IMPLICATIONS FOR EFFICIENCY AND EVIDENCE QUALITY IN HEOR

SA48

Elizabeth Weathers, PhD, RN, RGN, FAAN,[1,2] Melissa Bathish, PhD, RN, CPNP-PC,[3] Samy Ateia, MSc,[4] Julie Thompson, PhD,[5] Thomas Hopkins, MD, MBA,[2] and Nicole Darling, PhD, EIT.[2]

1=College of Health and Agricultural Sciences, University College Dublin, Dublin, Ireland; 2=ECNE Research, LLC, St. Petersburg, FL, USA; 3=School of Nursing, University of Michigan, Ann Arbor, MI, USA; 4=Computer Science Dept, University of Regensburg, Regensburg, Germany; 5= Statistics Dept, Duke University, Raleigh, NC, USA.

BACKGROUND

Systematic literature reviews and literature-surveillance workflows are essential to evidence-based healthcare and HEOR, where timely evidence synthesis informs health technology assessment, reimbursement submissions, economic modeling, real-world evidence generation, value assessment, and clinical or policy decisions. PRISMA 2020[1] reinforces the importance of transparent, reproducible reporting through structured checklists and flow diagrams for systematic reviews.

AI-enabled tools may improve the scalability of evidence synthesis, but require human-centric design, a clear context of use, risk-based validation, documentation, performance assessment, lifecycle oversight, and transparent reporting, consistent with emerging FDA and EMA guidance and principles.[2-4]

SafeSearch Pro™ is an **expert-supervised literature-surveillance workflow** designed to standardize evidence identification, screening, classification, and extraction while retaining human validation. This evaluation examines whether **SafeSearch Pro™** can support review performance comparable to dual independent human review while reducing reviewer time and workload.

METHODS

A comparative methods evaluation was conducted using two parallel workflows applied to the same review question, search strategy, eligibility criteria, and reviewer-calibration process:

- ✓ **SafeSearch Pro™ workflow**: AI-enabled citation screening, relevance classification, data extraction, and structured human validation
- ✓ **Manual workflow**: Dual independent human screening and extraction using conventional systematic literature review procedures.

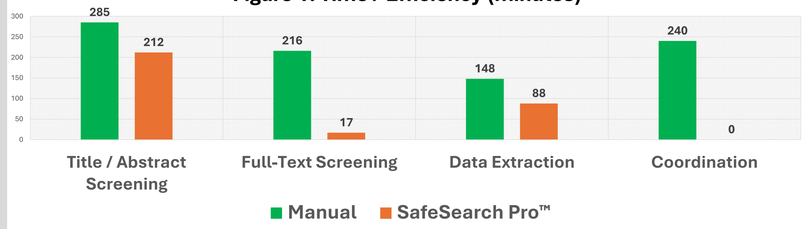
Reviewers completed a calibration exercise before screening to align interpretation of inclusion, exclusion, and extraction criteria. Outcomes were assessed across four domains: evidence yield, efficiency, reviewer workload / burden, and decision consistency.

Evidence-yield outcomes included records screened, full-texts reviewed, and final included studies. Efficiency outcomes included time spent on screening, full-text review, extraction, and coordination. Reviewer workload was assessed using post-task measures of cognitive load, perceived effort, and fatigue. Decision consistency was evaluated using intra-arm disagreement rates and agreement statistics.

Feature	Manual workflow	AI-enabled workflow
Literature sources searched	PubMed + Embase	PubMed + Semantic Scholar
Citation pool screened	753 citations	2,534 citations

RESULTS

Figure 1. Time / Efficiency (minutes)



***SafeSearch Pro™* expert-supervised workflow reduced total review time by 64%.**

Figure 2. PAAS Mental Effort
(1=very, very low mental effort; 9= very, very high mental effort)

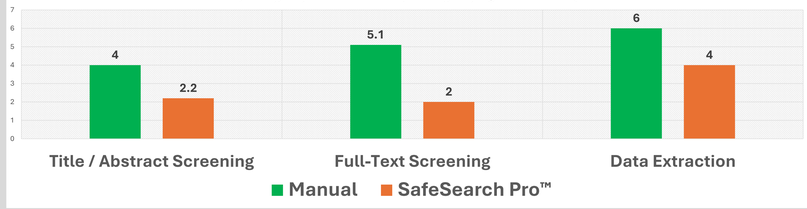
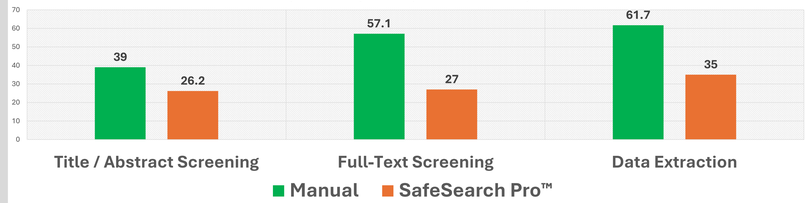


Figure 3. RSME Perceived Effort
(0=absolutely no effort; 150= extreme effort)



***SafeSearch Pro™* expert-supervised workflow reduced perceived workload / burden across all review phases, with PAAS and RSME scores consistently lower than manual review.**

Table 1. Evidence-yield, efficiency, and reviewer workload outcomes

Outcome / Metric	Manual review	AI-enabled workflow	Interpretation
Records screened	753	2,534	**+236% more records screened**
Included studies	16	16	**Same final included study count**
Operational time	10.8 h	5.3 h	**51% reduction**
Total time incl. coordination	14.8 h	5.3 h	**64% reduction**
Title / Abstract throughput	159 records / h	717 records / h	**4.5x faster**
Full-text throughput	15 papers / h	92 papers / h	**6.1x faster**
Inter-arm discordance	31.5%	1.1%	**28x fewer proportional disagreements**
Outcome / Metric	**Estimate**	**95% CI**	
Sensitivity (PPV adjusted)	92.3%	66.7%–98.6%	**Comparable evidence capture**
F1 Score (adjusted)	82.8%	-	**Based on sensitivity and adjusted PPV**

Note: Final adjudication of manual-review decisions is ongoing. Therefore, classification metrics are preliminary because manual decisions have not yet been fully adjudicated to establish a verified gold-standard reference set.

CONCLUSIONS

In this comparative evaluation, an expert-in-the-loop AI workflow screened a larger citation pool than manual review while producing the same final included-study count.

SafeSearch Pro™ reduced operational review time by 51%, reduced total time including coordination by 64%, lowered reviewer-reported burden across review phases, and showed fewer proportional intra-arm disagreements.

SafeSearch Pro™ transformed a labor-intensive HEOR review workflow into a faster, lower-burden, more consistent expert-validation process, screening 3.4× more citations and achieving the same included-study yield in less time.

These findings support AI-assisted, expert-supervised literature surveillance as a scalable approach for HEOR evidence generation.

KEY TAKEAWAYS

1. **AI expanded evidence reach**: The AI workflow screened 2,534 citations versus 753 manually, a 236% larger citation pool.
2. **Included-study yield was preserved**: Both workflows identified a final sample of 16 included studies.
3. **Review time fell substantially**: AI reduced operational time by 51% and total time including coordination by 64%.
4. **Reviewer burden was lower across phases**: PAAS and RSME scores favored the AI workflow during title/abstract screening, full-text screening, and extraction.
5. **Decision consistency improved**: Intra-arm disagreement was 1.1% with AI versus 31.5% manually.
6. **Comparable database coverage:** Both workflows demonstrated comparable biomedical database coverage across PubMed/Embase in the manual arm and PubMed/Semantic Scholar in the AI arm.

REFERENCES

[1] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
[2] European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Final version adopted by CHMP 9 September 2024 and by CVMP 11 September 2024..
[3] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. FDA Digital Health Center of Excellence. Updated periodically.
[4] U.S. Food and Drug Administration; European Medicines Agency. Guiding Principles of Good AI Practice in Drug Development. January 2026. Content current as of January 14, 2026.

SCAN ME

SafeSearch Pro™ aligns with emerging Good AI Practice principles by using a clearly defined context of use, expert human oversight, structured validation, traceable documentation, fit-for-purpose performance assessment, and transparent reporting to support responsible evidence generation.

Interested in hearing more? Scan the QR code to email the lead author.