Old Data, New Tricks: Accuracy and Efficiency Gains for AI-driven SLR Updates

Kelly Bell,[1] Ramsha Khan,[2] Sara Lucas,[3] Caitlyn Solem[4]
[1]GSK, Collegeville, PA, USA, [2]Cytel, Toronto, ON, Canada, [3]Cytel, London, UK, [4]GSK US, Bethesda, MD, USA

Background

- Artificial intelligence (AI) is becoming increasingly integrated into systematic literature review (SLR) workflows.
- Using AI, such as machine-learning (ML)-driven models, to screen titles and abstracts offers the potential to reduce manual burden and accelerate turnaround of reviews for health technology assessment (HTA) submissions.
- However, hesitancy among HTA agencies means that uptake of AI use remains limited.
- The 2024 position statement by the National Institute for Health and Care Excellence acknowledges AI's potential to enhance evidence generation but emphasizes that its use should augment, not replace, human involvement.[1]
- There is a need for robust evidence demonstrating that AI-assisted approaches can deliver reliable, high-quality outputs comparable with traditional methods.

Objectives

- This study aimed to 1) compare the accuracy of screening of titles and abstracts with ML-based systems under two training scenarios when screening a three-month SLR update; and 2) evaluate the efficiency gains of AI-assisted screening of titles and abstracts against human screening.

Methods

- ML models available on the Nested Knowledge® platform[2] were trained using two alternative scenarios.

1. The ML model was trained using the dataset from an original HTA-standard clinical SLR dataset in extensive-stage (ES) small-cell lung cancer (SCLC), previously screened by two experienced human reviewers (n = 5,850 records).
2. The ML model was trained on the minimum recommended number of human-screened citations for Nested Knowledge (n = 50 records, of which ≥10 are advance/include).[3]

- The trained ML models were used to screen the titles/abstracts of records identified during a three-month update conducted for the SLR (n = 244 records).
- The model scanned the abstracts and assigned an advancement probability score to each, ranging from 0 (irrelevant) to 1 (highly relevant).
- The probability scores were compared with the human screening decisions of the original SLR.
- To measure ML-based efficiency gains, the time taken for models to be trained and screen titles and abstracts in three HTA-standard SLRs (clinical, health-related quality of life [HRQoL], and economic) was compared with human dual-screening times.

Results

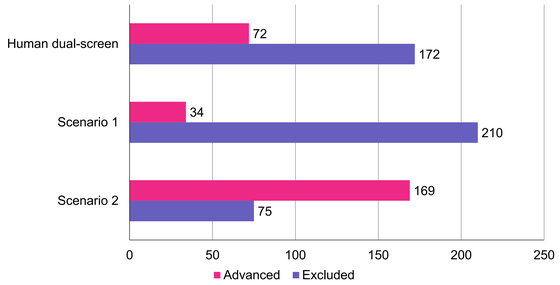
Human dual-screen performance

- Of 244 search hits in the clinical SLR update, 72 were advanced at title and abstract screening in the human dual-screen review.
- Of these, six were included at full-text screening.

Scenario 1 performance

- The ML model trained on 5,850 records advanced 34 of 244 records at title/abstract screening. (Figures 1 and 4)

Figure 1. Title/abstract advanced and excluded studies



- The model established a probability score threshold of 0.52.
- Of the six studies ultimately included at full-text screening by human dual review, the ML model correctly identified five (with a probability score >0.52).
  - The missing study had a low probability score (<0.52); ES SCLC was not explicitly stated in the title/abstract.
- The model was strong at identifying negatives (high specificity). (Table 1)
- Precision was high, so predictions were usually correct.
- Recall was low; many positives were missed at title/abstract screening, which led to a relevant study being missed.
- There was a risk of missing borderline studies (Figure 2).

Figure 2. Scenario 1 model screening profile

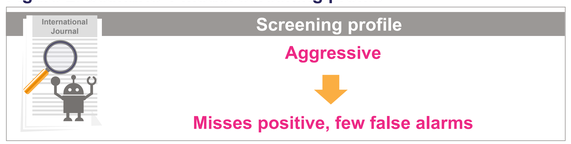


Table 1. Performance of both scenarios against human dual screen

Metric	Metric definition	Scenario 1	Scenario 2
Accuracy	How often is the model correct?	85.7%	52.0%
Precision	How many predicted positives are truly positive?	88.2%	36.7%
Sensitivity	How many actual positives are correctly identified?	49.1%	86.1%
Specificity	How many actual negatives are correctly identified?	97.8%	37.8%
F1 score	Harmonic mean of precision and sensitivity.*	0.632	0.515

* Used to deal with imbalanced datasets where accuracy may be misleading; for example, during screening of a database search with a high proportion excludes.

Scenario 2 performance

- The ML model trained on 50 records advanced 169 of 244 clinical SLR update records at title and abstract screening (Figures 1 and 4)
- The model established a probability score threshold of 0.10.
- Of the six studies ultimately included at full-text screening by human dual review, the ML model correctly identified all six (with a probability score >0.10).
- The model captured most positives (high sensitivity). (Table 1)
- Precision and specificity were low due to a high number of false positives.
- Despite noise, a high proportion of positives was captured (Figure 3).

Figure 3. Scenario 2 model screening profile

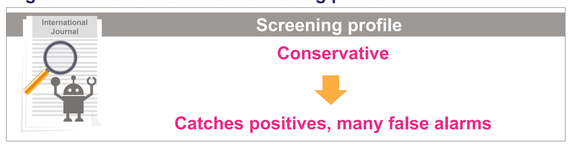
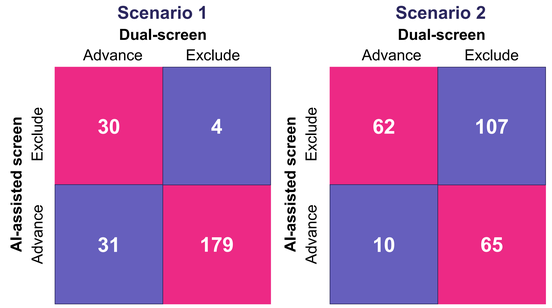


Figure 4. Title/abstract screening confusion matrices



Performance comparison

Strict vs permissive
Scenario 1 used a high threshold to aggressively filter; scenario 2 used a low threshold to advance almost all records.

Clean vs comprehensive
Scenario 1 was precise but missed positives; scenario 2 captured positives but created heavy noise.

Workload trade-off
Scenario 1 minimized human effort with risk; scenario 2 maximized safety with extra screening burden.

Efficiency gains

- Across three SLR topics, substantial time savings were observed with the use of ML-assisted screening compared with human dual screen. (Table 2)
- Using a fully automated approach, screening time was reduced by 98% to 100% in all three reviews.
- When adopting a hybrid approach of one ML model screener and one human screener, screening time was reduced by 37% in the clinical review and 50% in both the HRQoL and economic reviews.
- Time savings translated into reductions in reviewer hours, freeing capacity for downstream tasks such as full-text review and data extraction.
- Fully automated screening maximized efficiency; however, a hybrid approach may be more appropriate to ensure that human oversight is maintained.

Table 2. Efficiency of human dual screen and ML model screen

SLR topic	Search hits (n)	Dual-screen (hours)*	Hybrid-screen (hours)*	AI-screen (hours)
Clinical	2,873	68	34	1
HRQoL	4,247	100	50	0.6
Economic	5,288	124	62	0.6

* Assuming a human screening rate of 85 records per hour
Abbreviations: AI, artificial intelligence; HRQoL, health-related quality of life; SLR, systematic literature review

Conclusions

- ML model-assisted title and abstract screening can substantially reduce workload and accelerate SLR execution and update turnaround times.
- Models can feasibly replace one reviewer with high sensitivity, but require a human for final discretion on included studies.
- The balance between sensitivity, specificity, and human oversight will depend on the acceptable risks and intended review purpose.

References

1. National Institute for Health and Care Excellence. 2024. https://www.nice.org.uk/corporate/ecd11
2. Nested Knowledge. 2026. https://about.nested-knowledge.com
3. Nested Knowledge. 2025. https://about.nested-knowledge.com/docs/using-and-interpreting-the-screening-model/

Disclosures and acknowledgements
GSK study # 300530.

The authors would like to thank Andrew Easton and Colleen Dumont of Cytel, Inc. for poster development.