AI Support Reduced Screening Burden in a Systematic Review with Costs and Cost-Effectiveness Outcomes (SR-CCEO) for Cost-Effectiveness Modeling

MSR61

Ewa Borowiack, Ewelina Sadowska, Artur J Nowak, Jan Brożek
Evidence Prime, Krakow, Poland

Objective

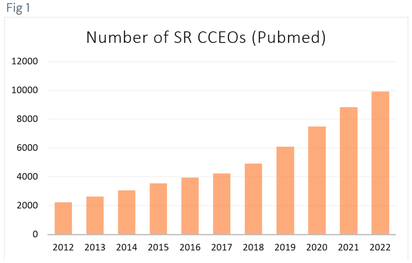
We explored whether an AI-assisted, semi-automated three-stage screening approach (AI-assisted single screening – AISS) improves efficiency in identifying the relevant studies for SRs of economic evaluations and if it is a safe alternative to the traditional two-reviewers screening approach.

Introduction

SR-CCEOs are crucial elements of the health economic studies, that identify inputs for economic models, describe strengths and limitations of health economic studies and inform decision makers where to allocate resources.

Through the last years we observe systematic growth of SR-CCEO studies. Based on our targeted search in Pubmed, over the past decade, the number of records related to cost-utility analysis has increased almost 5 times, from 2,226 in 2012 to 9,553 in 2022 [Fig 1]. This estimation does not include the many unpublished reviews undertaken as part of the Health Technology Assessment.

Fig 1



SR-CCEOs should follow the same structured approach as SRs of effectiveness. Screening is one of the most resource-intensive stages and typically requires two independent reviewers to avoid errors. However, due to resource and budget limitations, authors tend to use single screening in the selection of CCEO studies. According to the research study, only 54% of published SR-CCEOs used double screening approach. [2] As a result, there's increased risk of missings that affects the quality of SR.

ISPOR Good Practices for Critical Appraisal of SR-CCEO lists two screening approaches to accelerate the process: single reviewer screening and text mining. However, their effectiveness is still debated. [1]

Single screening conducted without support may lead to a significant loss of relevant studies (The median proportion of missed studies for experienced screener 3% (range: 0 to 21%) , for the junior screener 13% (range: 0 to 58%). [4]

Methods

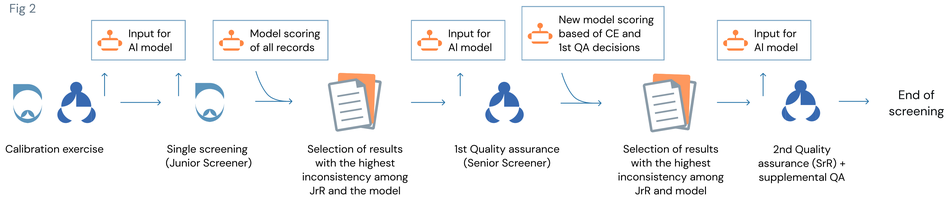
Sample

We selected one SR-CCEO of the cost-effectiveness of pharmacological management for osteoarthritis for this study. [3] In the original review, after a literature search in PubMed, EMBASE, Cochrane Library, Health Technology Assessment (HTA) database, and National Health Service Economic Evaluation database (NHS EED), authors identified 43 studies that met eligibility criteria.

Screening simulation - process

We deduplicated all records retrieved from the same databases as in the original SR, using the machine learning model. We restricted search to the same cut-off date - 4 November 2021. At the same time, screening instruction and relevant desired and undesired keywords were entered into Laser AI. Records obtained after deduplication were uploaded to Laser AI for three-stage screening [Fig 2]:
1) calibration exercise (CE) – training phase for both screeners and AI model. Calibration exercise was planned to be conducted in rounds, until the disagreement rate among screeners was relatively low.
2) main screening phase – single screening by a junior reviewer (JrR) supported by AI-features (e.g., AI-records prioritization) and non-AI features (eg., keywords, records filtering).
3) quality assurance (QA) - performed by a senior reviewer (SrR) only on those records for which the decisions of the junior screener and the AI model disagreed. AI model's record suggestions were based on the screeners decisions from the calibration exercise. After the first QA, model's suggestions for the rest of the records were re-ranked according to both calibration exercise and senior screener's decisions. Number of QA sessions was based on senior screener decision. In case of possible systematic error made by junior screener, senior screener was able to conduct supplemental rounds of QA (without model suggestions) based on simple keywords filtering.

Fig 2



Metrics

Primary endpoint

- The primary endpoint of the study was the proportion of records and studies missed compared with the reference standard – the original SR.

Secondary endpoints

- Total time spent on screening, workload, and time saved in comparison with double screening. To measure resource use in our approach we obtained data (mean time per record) directly from the tool and compared with the times recorded manually by screeners.
- Additionally we estimated the proportion of records missed in each screening phase and effectiveness of AI model in terms of QA phase (estimation based on number of relevant records (missed by junior screener) proposed by model during QA).

Stage	Stage description	Results
Deduplication	with machine learning model	Search results: 5679 records ⟶ Records excluded: 944
Pilot exercise	**Double screening** (sample of records) 1. Screening guide validation 2. Training phase for both screeners and AI model	Sample: 100 records ⟶ **Disagreement rate: 5%**
Main screening	**Single screening** (all records) supported by: 1. AI features: records prioritization 2. Non-AI features: keywords: highlights, filter-based clustering	Number of screened records: **4359** ⟶ **Number of records missed: 4** **Proportion of records missed: 10.53%**
Quality assurance	**Single screening in rounds** (sample of records) Records for which the decisions of the first screener and the AI model disagreed	Number of QA rounds: 3 Number of records screened during QA rounds: 552 ⟶ **Number of relevant records (missed by 1st reviewer) proposed by model during QA: 3** **Model Effectiveness: 75%**
All stages	**Records screened in double: 652** **Records screened by single reviewer: 3807**	**Proportion of records missed: 2.63% [Sensitivity: 97.56%]** **Proportion of studies missed: 0%**

Results

We identified 4459 records and 215 were finally included for full text screening. During calibration exercise, reviewers screened 100 records with the 5% of disagreement rate, what allowed to start the main phase of the experiment. During the single screening phase, junior screener missed 4 out of 38 relevant records (10.53%), and 3 of them were chosen by the AI model for the QA phase and included by the senior screener. QA phase was conducted in 3 rounds, with 150 and 201 records proposed by AI model and with 201 records screened during supplemental QA.
Altogether, following our approach, 37 out of 38 records were included. However, considering conducted studyfication, we found all of the 38 studies included in the original review. Overall, among 4459 records, 652 were screened in double and 3807 by single reviewer -Title and abstract screening workload was reduced by 43%, whereas the estimated time saving was 19 hours and 16 minutes (total screening time 20h 57min).

Total screening time	20h 27min
Time savings	19h 16min
Workload savings	42.69%

Conclusion

To the best of our knowledge, this project is the first user-based evaluation on SR-CCEO that combines human and AI effort in a three stage screening process.
As expected, based on published evidence [4] single screening may lead to exclusion of relevant studies.
Introduction of the quality assurance phase based on model suggestions is a promising risk management strategy in a single screening approach.
In the follow-up experiment conducted on the systematic review of cost-effectiveness studies of telehealth-delivered diet and exercise interventions [5], the proportion of records missed was 0% [Sensitivity: 100%].
Our results suggest that the AI-assisted single screening (AISS) approach might reduce human effort in SR-CCEOs. Further validation in a broader range of SR-CCEOs is ongoing.

References

1. Mandrik OL, Severens JLH, Bardach A, Ghabri S, Hamel C, Mathes T, Vale L, Wisløff T, Goldhaber-Fiebert JD. Critical Appraisal of Systematic Reviews With Costs and Cost-Effectiveness Outcomes: An ISPOR Good Practices Task Force Report. Value Health. 2021 Apr;24(4):463-472. doi: 10.1016/j.jval.2021.01.002. PMID: 33840423.
2. Luhnen M, Prediger B, Neugebauer EAM, Mathes T. Systematic reviews of health economic evaluations: A structured analysis of characteristics and methods applied. Res Synth Methods. 2019 Jun;10(2):195-206. doi: 10.1002/jrsm.1342. Epub 2019 Mar 4. PMID: 30761762.
3.Shi J, Fan K, Yan L, Fan Z, Li F, Wang G, Liu H, Liu P, Yu H, Li JJ, Wang B. Cost Effectiveness of Pharmacological Management for Osteoarthritis: A Systematic Review. Appl Health Econ Health Policy. 2022 May;20(3):351–370. doi: 10.1007/s40258-022-00717-0. Epub 2022 Feb 9. PMID: 35138600; PMCID: PMC9021110.
4. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019;19(1):132.
5. Law L, Kelly JT, Savill H, et al. Cost-effectiveness of telehealth-delivered diet and exercise interventions: A systematic review. J Telemed Telecare. 2022:1357633x211070721.

Presented at ISPOR 2023, Boston. Contact us at hello@laser.ai