Validating the AI Assisted Abstract Screening Feature of Nested Knowledge Platform

145449

[1]Anne-Mary Lewis-Mikhael, MD, PhD, [2]Xuan Wang MD, MSC

[1]ICON plc, Ontario, Canada; [2]ICON plc, Stockholm, Sweden

INTRODUCTION

- Systemic literature reviews (SLRs) play an important role in Health Technology Assessment (HTAs). However, with the huge amounts of publications, the process is very time-consuming, specifically for the screening process.
- Accordingly, the use of artificial intelligence (AI) platforms such as Nested Knowledge (NK) has become more popular due to potential efficiencies while conducting literature reviews.
- NK has multiple functionalities that help in conducting literature reviews with high efficiency, including the AI assisted screening.

AI assisted Screening

Robot Screener replaces one human reviewer with an AI reviewer in nests with a Dual Screening mode

It does require training (50 adjudicated screening decisions and 10 advancements or inclusions) prior to being switched on, but continually trains itself thereafter.

Then, a human adjudicator reviews the preliminary screenings and makes the final decision.

OBJECTIVE

- We aimed to assess the accuracy of one of NK AI's key features, AI assisted screening.

METHODS

- We compared all cross validations when utilizing the AI-assisted abstracts screening. These include different measures for accuracy, AUC (area under the curve), and recall.
- A key metric evaluated is "Recall", which reflects the tool's ability to identify relevant studies. Precision, on the other hand, measures the proportion of studies flagged by the tool that are relevant. These metrics provide a comprehensive picture of Robot Screener's effectiveness in both flagging studies that are likely to be includable or of interest and thus preventing 'missed' studies (Recall) and in excluding studies that are likely outside of a review's purview to save the effort of the adjudicator in excluding these non-relevant studies (Precision).
- NK recommends 50 expert-screened abstracts as a minimum training requirement for the AI model.
- Given that our literature review had different population subgroups and several outcomes to be addressed, we compared cross validation-based accuracy after screening 225 and 450 abstracts, as well as after screening all abstracts that were to be screened (1,550 abstracts).

RESULTS

- After screening 225 and 450 abstracts, the accuracy increased minimally from 0.71 to 0.72.
- After screening 1550 abstracts, the accuracy reached 0.89.
- The AUC reflects how well the model discerns between included and excluded records.
- It increased very slightly from 0.83 to 0.84 after screening 225 and 450 abstracts. After completing the 1550 screenings, it changed slightly to 0.89 (0.80+ indicates a high AUC).
- The recall was the only measured aspect that might have changed after screening all 1550 abstracts, indicating that the model will less frequently lean towards exclusion of relevant records.
- The recall was 0.89, 0.91 and 0.76 after screening the 225, 450 and 1550 abstracts, respectively.
- NK recommends minimum acceptable recall of >70%, which was met for the three different thresholds of abstracts screened.

Predictions | Cross Validation

Measure	Score
AUC	0.84
Recall	0.91
Precision	0.48
F1	0.63
Accuracy	0.72

Figure 1, Measures and scores obtained after screening 450 abstracts.

CONCLUSION

- The AI screening feature offered by NK can be effectively utilized after training the model with a sufficient number of abstracts.
- Our exercise proved the great AUC, recall, and accuracy of the AI assisted abstract screening offered as an important method for providing screening assistance
- These results confirm previous findings of two recent studies that have highlighted the effectiveness of the performance of robot screening in NK in diverse types of reviews.
- High recall that has even significantly outperformed human reviewers have been found, where the recall rate was 97.1% versus 94.4% respectively. In the other study, robot Screener's Recall rate (0.79 ± 0.18) was comparable to human reviewers (0.80 ± 0.20).

CONCLUSION cont'd

- High recall is considered crucial as it ensures that no relevant studies are missed in the initial screening phase. This ensures comprehensiveness which is critical in Health Economics and Outcomes Research (HEOR).

CONTACT INFORMATION

AnneMary.LewisMikhaelSaad@iconplc.com;

Xuan.Wang@iconplc.com