Assessing Recall in Abstract Screening: Artificial Intelligence vs. Human Reviewers

MSR91

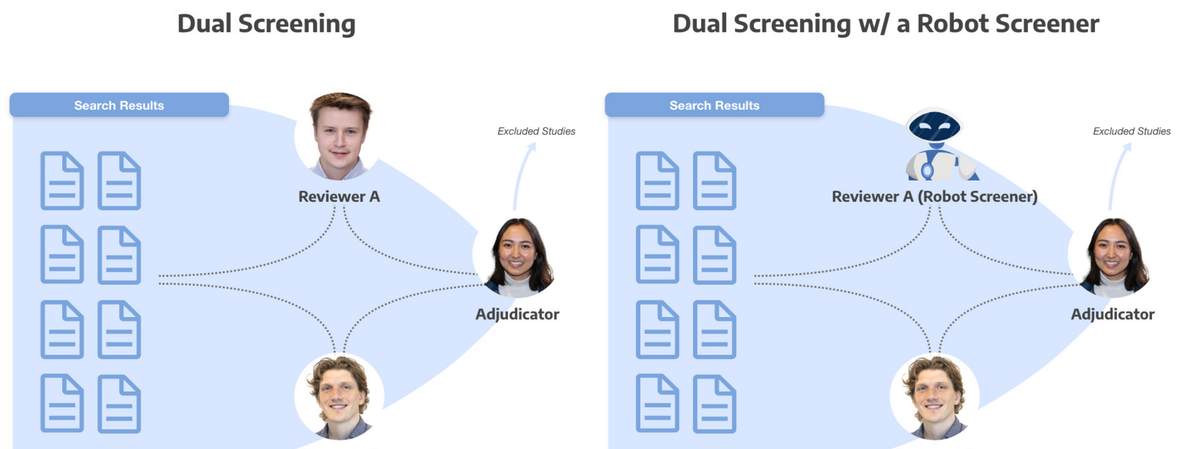
Jade Thurnham, Kevin Kallmes, Karl Holub;
Nested Knowledge, St. Paul, MN, USA.

Background

Screening for relevant records is a necessary but time-intensive task in the systematic literature review (SLR) process. To reduce human labor input, Artificial Intelligence (AI) has been proposed as a partial or total replacement for human screeners, but concerns exist about the accuracy of AI screening. The most important concern is whether AI tools have lower recall, meaning they would miss more relevant records than human reviewers, leading to incomplete evidence in SLRs. Here, we assess the performance of Nested Knowledge Robot Screener, an AI for inclusion/advancement prediction, by comparing the recall and precision of human reviewers against Robot Screener in SLRs that employed this AI.

Methods

Clinical, economic, and mental health SLRs that employed Robot Screener with at least 50 abstract-level human screening decisions in the AutoLit software were included. Human and Robot Screener abstract-level advancement decisions were compared against final, adjudicated advancement decisions to determine recall.



Results

Nineteen SLRs with 8,927 final advanced records were assessed. Human reviewers correctly advanced 8,097/8,580 records, with recall of 94.4% and precision of 86.4%. Robot Screener correctly advanced 5,791/5,965 records, with recall of 97.1% and precision of 47.3%. In a two-sided chi-squared analysis, Robot Screener's recall was significantly higher than human ($p<0.001$) and precision was significantly lower ($p<.001$).

Review Type	Reviewer Decisions
Clinical Review informing evidence repository	471
Clinical burden review	3,882
Clinical review	2,279
Mental health review	59
Clinical review informing virtual patient creation	1,684
Clinical review informing virtual patient creation	425
Clinical review (adverse events)	2,367
Economic review	2,645
Economic review	5,493
Clinical review	8,490
Mental health review	4,087
Clinical review	2,605
Clinical review	5,624
Clinical review	1,861
Clinical review	1,123
Clinical review	10,274
Clinical review	1,454
Clinical review	507
Clinical review	415

Conclusions

Robot Screener had higher recall and lower precision when compared with human abstract screeners. These findings suggest that Robot Screening may be appropriate as an assistive tool to save time in the SLR screening process without sacrificing comprehensiveness. Limitations include the fact that the selection of SLRs analyzed may not be generalizable and different numbers of records screened by humans vs. AI. Further research is necessary to assess the potential time savings of the integration of AI screening tools and the precision/recall tradeoff.

{NESTED} KNOWLEDGE®