Can artificial intelligence separate the wheat from the chaff in systematic reviews of health economic papers?

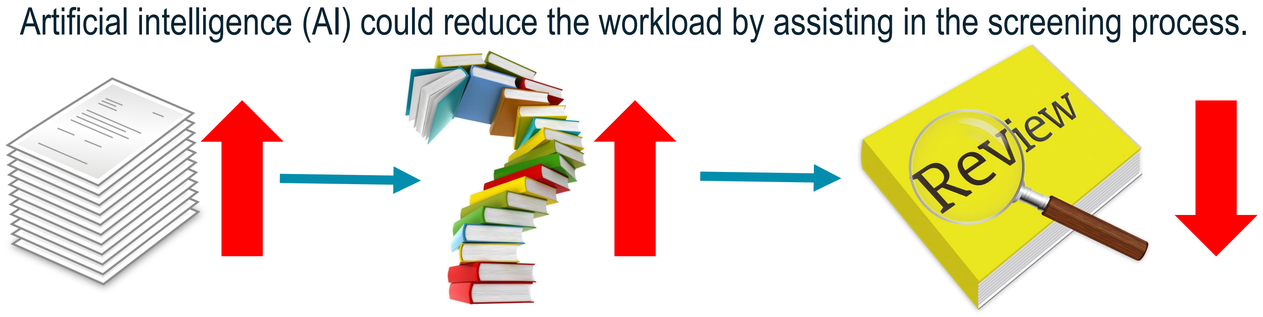
M.J. Oude Wolcherink, X.L.G.V. Pouwels, S.H.B. van Dijk, C.J.M. Doggen, H. Koffijberg

University of Twente, Health Technology and Services Research, Techmed Centre, Enschede, the Netherlands

Background

The number of published health economic evaluations has risen sharply leading to increased workload for performing systematic reviews

Artificial intelligence (AI) could reduce the workload by assisting in the screening process.



Scientific output — Workload systematic review — Feasibility to perform new systematic reviews

Aim

To assess the accuracy and efficiency of using artificial intelligence during the screening of systematic reviews within health economics.

Methods

A sample of 4,994 articles focussing on health economic evaluations of early detection strategies for cardiovascular disease was manually screened. Figure 1 indicates which articles were: relevant for full text screening, data extraction, or irrrelevant.

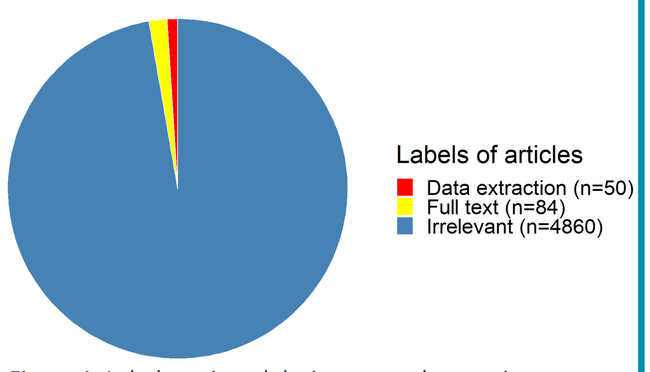


Figure 1: Labels assigned during manual screening

Primary outcomes: Accuracy (percentage of relevant articles found) & Efficiency (proportion needed to screen and time saved during screening).

The open-source ASReview[a] used machine learning to rank articles in order of the likeliness of being relevant. Initially, the ranking is based on prior knowledge, i.e. one relevant and one irrelevant article. Subsequently, the ranking is iteratively adapted by using the input of the reviewer.

Screening with AI was retrospectively assessed through 1,000 simulations in which the prior knowledge was varied. All articles included in the full text screening were labelled as 'relevant'.

Screening was continued until the stopping rules below were satisfied:

1. Screening a predetermined proportion (5%, 7.5%, 10%, 15%, and 20%)
2. Screen until a series of consecutive irrelevant articles (50, 100, 150, and 200)

Retrospectively, the number needed to screen to find all relevant articles was also assessed.

van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence 2021 3:2*. 2021;3(2):125-133. doi:10.1038/s42256-020-00287-7

Results

Table 1: Primary outcomes of screening with AI per stopping rule. Accuracy was reported as the mean percentage found, and efficiency as the proportion needed to screen (PNS) and time saved in hours.

Stopping rule	Full text			Data extraction		
All relevant articles	Accuracy (%)	PNS (%)	Time saved (hours)	Accuracy (%)	PNS (%)	Time saved (hours)
-	100	92	6.4	100	12.8	57.8
Proportion of total	Accuracy (%)	PNS (%)	Time saved (hours)	Accuracy (%)	PNS (%)	Time saved (hours)
5%	52.9	5	59.3	92	5	59.3
7.5%	67.7	7.5	57.7	98.9	7.5	57.7
10%	73.6	10	56.2	99.9	10	56.2
15%	80.2	15	53.1	100	15	53.1
20%	83.7	20	49.9	100	20	49.9
Consecutive irrelevant articles	Accuracy (%)	PNS (%)	Time saved (hours)	Accuracy (%)	PNS (%)	Time saved (hours)
50	69.4	10.7	55.7	92.9	10.7	55.7
100	79.5	17.6	51.5	96.7	17.6	51.5
150	81.8	20	49.9	98.4	20	49.9
200	83.9	23.5	47.8	100	23.5	47.8

The mean outcomes for screening with the AI-tool until the stopping rules were satisfied are shown in Table 1 for both pre-labelled datasets. Assuming a reviewer takes on average 45 seconds per abstract[b], the time saved was calculated. Accuracy during the screening with the AI-tool is shown in Figure 2.

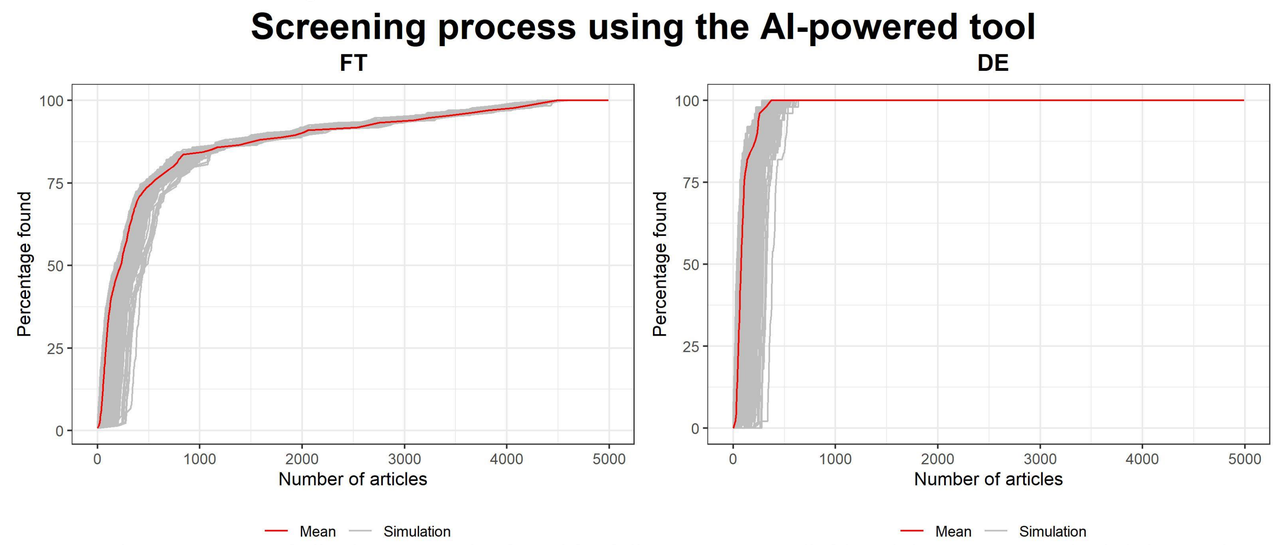


Figure 2: The proportion of all relevant articles found for full text screening (left) and data extraction (right) during the manual screening process. Each grey lines represents a simulation and the red line represents the mean of all simulations.

[b] Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*. 2010;11(1):1-11. doi:10.1186/1471-2105-11-55/FIGURES/6

Discussion & Conclusion

With a focus on articles included in full text screening and articles included for data extraction, major time savings could be obtained by using AI at no or very limited cost to accuracy.

Manual screening is used as reference, but is typically not flawless either.

The differences in finding all articles included for full text screening and articles included for data extraction are striking. This may be caused by lenient inclusions for full text screening.

Whereas these results are promising, they should be further validated in subsequent systematic reviews focussing on health economic evaluations. As it is yet unknown how other AI-tools perform within this field, future comparison between different tools is desired.

Martijn Oude Wolcherink, MSc
m.j.oudewolcherink@utwente.nl
+31534893230

The project was funded by the Dutch Heart Foundation through 3 grants: EARLY-SYNERGY (CVON2015-17), CONCRETE (CVON2017-14), and RED-CVD (CVON2017-11)

UNIVERSITY OF TWENTE. | TECHMED CENTRE