SWAR 52: Artificial Intelligence as a screener and data extractor of records for a systematic review

Objective of this SWAR
To evaluate the efficacy and efficiency of artificial intelligence (AI) for study screening and data extraction in a systematic review.

Study area: Study Identification, Data extraction
Sample type: Review Authors, Editors
Estimated funding level needed: Unfunded

Background
To truly democratize science, we must ensure that the means of its production are as accessible as its conclusions. In many settings for systematic reviews and evidence syntheses, the availability of a second or third reviewer for critical tasks, such as screening studies for eligibility and data extraction, is a significant practical barrier. Given the advanced state of today's AI, the use of a validated AI tool as a substitute for a human is not only a reasonable option but is arguably a responsible course of action for a researcher facing such constraints. This outline for a Study Within a Review (SWAR) (1) presents a generic design to assess this.

An initial corpus of scholarly articles would be compiled through a systematic search of a literature database such as PubMed, Embase or Scopus. The search would be restricted to articles published in the last 10 years to ensure relevance and manageability. The search query will consist of keywords and Boolean operators. For example: ("topic A" OR "topic B") AND "specific methodology C". All retrieved titles and their corresponding abstracts would be imported into a reference management software and duplicates removed. This final set of unique records would constitute the sample for the screening intervention.

Screening
The compiled records would be screened independently and in parallel by two entities: a human reviewer and an AI tool. The eligibility criteria for the review would be as follows and the aim would be to categorise each title as "Include", "Exclude" or "Uncertain":
Inclusion: For a specific prevalences condition (e.g. caries), (a) must be a primary research study; (b) must involve human participants.
Exclusion: (a) review article, editorial or commentary; (b) not published in English; (c) fails to meet specific methodological requirements.

The lists of "Include" classifications from both the human reviewer and the AI will be compared. The primary outcome will be the measure of inter-rater reliability, calculated using Cohen's Kappa coefficient, to assess the level of agreement beyond chance. Raw percentage agreement, sensitivity and specificity of the AI, using the human reviewer's selections as the reference standard, will also be reported.

Data extraction
For all articles selected as "Include" by both the human and the AI, a standardized data extraction form will be used. Both the human reviewer and the AI will independently extract the following key data points from the full text of these articles: number of participants, any index, sex, age. The extracted data will be compared for concordance and accuracy.

Interventions and Comparators
Intervention 1: Human reviewer: A single researcher (the "reviewer") will manually screen each title and abstract against a predefined set of inclusion and exclusion criteria.
Intervention 2: AI intervention: Simultaneously, the same set of titles and abstracts will be processed by Google Gemini 2.5 PRO, with no human intervention during the AI's classification process. 10 articles will be screened per prompt.
Intervention 3: Human reviewer: The “reviewer” will extract a pre-determined set of data from the full article of each included record.

Intervention 4: AI intervention: The AI tool will be used to extract the same pre-determined set of data from the full article of each included record.

Index Type: Searching; Protocol; Dissemination

Method for Allocating to Intervention or Comparator:

Outcome Measures

Primary: Primary Outcomes

1) Agreement on study selection (metric: Cohen's Kappa (κ) coefficient). This will measure the inter-rater reliability between the human reviewer and the AI tool for the title and abstract screening phase. It quantifies the level of agreement in classifying articles as "Include" or "Exclude," while accounting for the possibility of agreement occurring by chance. A higher Kappa value indicates stronger agreement.

2) Accuracy of data extraction (metric: Data point concordance rate). This will be calculated for the subset of articles classified as “Include” by both the human and the AI. It is defined as the percentage of specified data points (e.g., study population size, primary outcome measure, study design) that are identically extracted by both the AI and the human reviewer from the full text. This is a direct measure of the AI's accuracy in performing the data extraction task.

Secondary:

1) Diagnostic accuracy of AI screening (metrics: Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV)). Treating the human reviewer's decisions as the reference standard, these metrics will evaluate the AI's performance as a screening tool. Sensitivity will measure the AI's ability to correctly identify the relevant articles. Specificity will measure the AI's ability to correctly reject irrelevant articles. PPV and NPV will provide insight into the predictive performance of the AI's classifications.

2) Efficiency of the process (metric: Time to completion in hours/minutes). This will be a direct comparison of the time required for the human reviewer to complete the screening and data extraction tasks versus the computation time required for the AI to perform the identical tasks.

3) Discrepancy analysis (metric: Qualitative categorization of disagreements). A descriptive analysis will be performed on all instances where the AI and the human reviewer disagreed on study selection or data extraction. The nature of these discrepancies will be categorized (e.g., "AI missed nuanced exclusion criterion," "Human error in data entry," "Ambiguous language in abstract") to identify any systematic patterns of error for either the AI or the human.

Analysis Plans

1. Analysis of diagnostic accuracy: The diagnostic accuracy of the AI screening tool will be evaluated using the human reviewer's classifications as the reference standard. A 2x2 contingency table will be constructed as follows:

Human: Included; Excluded

AI: Included True Positives (TP); False Positives (FP)

AI: Excluded False Negatives (FN); True Negatives (TN)

From this table, the following metrics will be calculated, along with their 95% confidence intervals using the Clopper-Pearson method:

Sensitivity: TP / (TP + FN)

Interpretation: The proportion of truly relevant articles that the AI correctly identified.

Specificity: TN / (TN + FP)

Interpretation: The proportion of truly irrelevant articles that the AI correctly rejected.

Positive Predictive Value (PPV): TP / (TP + FP)

Interpretation: The probability that an article included by the AI is truly relevant.

Negative Predictive Value (NPV): TN / (TN + FN)

Interpretation: The probability that an article excluded by the AI is truly irrelevant.

2. Analysis of process efficiency
The efficiency of the screening and data extraction process will be compared between the human reviewer and the AI tool. Total time will be recorded for both the human reviewer (in hours/minutes) and the AI (in CPU time, converted to hours/minutes). The time-to-completion data will be summarized using descriptive statistics (mean, median, standard deviation, range). To test for a statistically significant difference in efficiency, a two-sample t-test will be used if the time data are normally distributed. If the assumption of normality is violated (as assessed by the Shapiro-Wilk test), the non-parametric Mann-Whitney U test will be employed as a robust alternative. The results will be reported as the mean difference in time with a corresponding 95% confidence interval.

3. Analysis of Discrepancies
A qualitative and quantitative analysis will be performed on all disagreements between the human and AI classifications. A coding scheme will be developed to categorize the reasons for each discrepancy (e.g., misinterpretation of inclusion/exclusion criteria, processing of non-textual data like figures, ambiguity in the source text). Two independent researchers will review each instance of disagreement and assign a category from the coding scheme. Inter-coder reliability for the categorization process will be assessed using Cohen's Kappa. The frequency and percentage of each category of disagreement will be calculated and presented in a table. This analysis will provide insights into any systematic error patterns of the AI tool, which is critical for understanding its limitations and potential for future refinement.

Possible Problems in Implementing This SWAR
1. Limitations related to the reference standard
a) Single human reviewer: The study protocol designates a single human reviewer as the reference or "gold" standard. This presents the most significant limitation. The performance of the AI is judged against one individual whose classifications are subject to their own biases, fatigue, and potential for error. A more robust design would use a consensus judgment from two or more independent human reviewers to establish a more reliable ground truth.
b) Human reviewer bias: The human reviewer is susceptible to cognitive biases, such as confirmation bias or a tendency to be more lenient or strict as the review progresses (instrument drift). These inconsistencies can affect the perceived accuracy of the AI.

2. AI-specific challenges and biases
a) Algorithmic and training data bias: The AI tool's performance is fundamentally dependent on the data it was trained on. If its training corpus underrepresented certain study designs, niche terminology or research from specific geographical regions, its performance may be systematically weaker on those articles.
b) Inability to interpret nuance: The AI may struggle with tasks requiring deep contextual understanding, such as interpreting irony, sarcasm or highly nuanced language within an abstract, potentially leading to incorrect classifications.
c) Sensitivity to data format: For the data extraction phase, the AI's accuracy may be highly sensitive to the formatting of source documents (e.g., multi-column PDFs, complex tables, data presented only in figures). This could lead to a high rate of extraction failure for certain types of articles.
d) The "Black Box" problem: Depending on the specific AI model used, its decision-making process may not be fully transparent. While the discrepancy analysis aims to mitigate this, it can still be challenging to diagnose the precise reason for certain AI errors.

3. Limitations to external validity (generalizability)
a) Specificity of the AI tool: The findings of this study will be specific to the particular AI model and version used. The findings, whether positive or negative, may not be generalizable to other AI screening tools, which may use different architectures and training data.
b) Specificity of the research domain: The performance of the AI would be tested in a single, specific subject area. The tool's effectiveness could differ significantly in another field with distinct

jargon, evidence standards, and publication conventions. Therefore, claims about the utility of the AI for systematic reviews in general must be made with caution.
c) Language restriction: If the literature search is restricted to a single language (e.g., English), this will introduce a known selection bias and limit the generalizability of the findings to a global body of literature.

References

1. Devane D, Burke NN, Treweek S, Clarke M, Thomas J, Booth A, et al. Study within a review (SWAR). Journal of Evidence-Based Medicine 2022;15(4):328-32.

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Luiz Fellipe Nakamai
Contact email address: nakamai@alumni.usp.br
Date of idea: 13/06/2025
Revisions made by: Luiz Fellipe Nakamai
Date of revisions: 13/06/2025