SWAR 31: Use of AI for extraction of outcome data

Objective of this SWAR
To assess the reliability of AI (compared to human reviewers) for extracting outcome data for use in systematic reviews.

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

Background
Accurate data extraction is a crucial component of a systematic review. Data extraction can be highly time consuming and accuracy can vary based upon the experience of review authors. This might lead to erroneous conclusions in the review. Newly developed software is claimed to be able to extract relevant data from reports of randomised trials directly by employing artificial intelligence (AI). However, these AI-based methods are yet to be validated.

Interventions and Comparators
Intervention 1: AI based data extraction. AI software will be tasked with extracting primary and secondary outcome data from identified randomised trials. As there are various AI data extraction software available, we will approach other NIHR evidence synthesis teams and encourage them to similarly adopt AI data extraction software. This will enable comparisons of data extraction accuracy between competing software.
Intervention 2: Human based data extraction. Two human reviewers (within each review) will extract primary and secondary outcome data from identified randomised trials and achieve consensus agreement on accuracy.

Index Type: Data extraction

Method for Allocating to Intervention or Comparator:
Both methods of data extraction will be used for each trial report.

Outcome Measures
Primary: Inter-observer level of agreement in extracted outcome data between AI and human reviewers.
Secondary: We will also evaluate time-taken to complete data extraction for Human vs AI-based data extraction. Human reviewers will record time manually when performing data extraction. Additionally, where there are differences in data extracted for AI vs human reviewers, we will record descriptive details of the type of information extracted. Where possible, we will quantitatively examine differences based on type of data extracted (e.g. Reading results from tables vs figures vs text) to determine if quality of AI-based data extraction varies on this basis.

Analysis Plans
Cohen's Kappa will be used to establish level of agreement between extracted data performed by AI versus human reviewers. We will analyse differences in time-taken to complete data extraction using t-tests for normally distributed data or Mann-Whitney U for non-parametric data. If possible, we will perform a network meta-analysis comparing performance metrics of competing AI-software.

Possible Problems in Implementing This SWAR
Our evaluation of the reliability of using AI for data extraction is dependent on human reviewers performing data extraction with a very high (or perfect) degree of accuracy. Failure of the latter is likely to compromise our conclusions on how well AI performs when undertaking data extraction.

References

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Martin Taylor-Rowan
Contact email address: martin.taylor-rowan@glasgow.ac.uk
Date of idea: 18/06/2024
Revisions made by: Martin Taylor-Rowan
Date of revisions: 16/07/2024