TITLE: How is the agreement between machine and humans? Use of RobotReviewer to evaluate the Risk of bias of randomized trials.

Susan Armijo-Olivo,[1,2] Rodger Craig,[1] Sandy Campbell[1,3]

[1] Institute of Health Economics (IHE) Edmonton AB T5J 3N4 Canada
[2] Faculty of Rehabilitation Medicine, Department of Physical Therapy, University of Alberta, Edmonton, Canada T6G 2G4
[3] John W. Scott Health Sciences Library, University of Alberta, Edmonton, Canada T6G 2R7

ABSTRACT

Background: Evidence from new technologies and treatments is growing, along with demands for evidence to inform policy decisions. Thus, it is anticipated that the need for knowledge synthesis products (i.e. Health Technology Assessments (HTAs) and systematic reviews (SRs)) will increase. Increased demands will create challenges to complete assessments in a timely manner. New technologies such as RobotReviewer, a semi-autonomous risk of bias (RoB) assessment tool, seek to decrease the time and resource burden to complete HTAs/SR. However, current evidence to validate the existing software for use in the HTA/SR process is limited.

Objectives: To test the accuracy and agreement between RobotReviewer and RoB assessments generated by consensus among human reviewers.

Methods: A random sample of randomized controlled trials (RCTs) were used. Consensus assessments between the two reviewers were compared with the RoB ratings generated by RobotReviewer. Agreement between RobotReviewer, and human reviewers was assessed using weighted kappa (κ). The accuracy of RobotReviewer was assessed by calculating the sensitivity and specificity.

Results: In total, 372 trials were included in this study. Inter-rater reliability on individual domains of the RoB tool ranged from κ= -0.01 [95% CI: -0.03, 0.001; no agreement) for overall RoB, to κ= 0.62 (95%CI: 0.534, 0.697; good agreement) for random sequence generation. The agreement was fair for allocation concealment(κ= 0.41 (95%CI: 0.31, 0.51), slight for blinding of outcome assessment (κ= 0.23 (95%CI 0.13, 0.34), and poor for blinding of participants and personnel κ=0.06 (95%CI: 0.002, 0.1). Over 70% of irrelevant quotes to make the RoB judgments were found for blinding of participants and personnel (72.6%) and blinding of outcome assessment (70.4%).

Conclusions: This is the first study in providing a thorough analysis of the usability of RobotReviewer. Agreement between RobotReviewer and human reviewers ranged from no agreement to good agreement. However, RobotReviewer selected a high percentage of irrelevant quotes in making RoB assessments. Use of Robotreviewer in isolation as a first or second reviewer is not recommended at this point.

Patient or health consumer involvement: It is hoped that the results help knowledge synthesis teams whether to use such a tool to speed up the process of knowledge synthesis.