JOURNAL OF NURSING SCHOLARSHIP

Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews

Julian Hirt, MSc, RN[1,2,*], Jasmin Meichlinger, MSc, RN[1], Petra Schumacher, BScN, MScN, RN[3], & Gerhard Mueller, Prof, Dr, MSc, RN[4]

1 Research Associate, Institute for Applied Nursing Science, Department of Health, University for Applied Sciences FHS, St. Gallen, Switzerland
2 International Graduate Academy, Institute for Health and Nursing Science, Medical Faculty, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
3 Professor, IMC University of Applied Sciences Krems, Department Health Sciences, Krems, Austria
4 Professor, Institute of Nursing Science, Department of Nursing Science and Gerontology, UMIT–Private University for Health Sciences, Medical Informatics and Technology, Tyrol, Austria

Key words
Automation, bias, evaluation studies, machine learning, randomized controlled trial, RobotReviewer

Correspondence
Julian Hirt, Institute for Applied Nursing Science, FHS St.Gallen, University of Applied Sciences, St.Gallen, Rosenbergstrasse 59, 9000 St. Gallen, Switzerland.
E-mail: julian.hirt@ost.ch

Accepted December 3, 2020

doi:10.1111/jnu.12628

Abstract

Purpose: RobotReviewer is a machine learning system for semi-automated assistance in risk of bias assessment. The tools's performance in randomized controlled trials (RCTs) in the field of nursing remains unknown. We aimed therefore to evaluate the agreement in risk of bias assessment between RobotReviewer and human reviewers.
Design: Evaluation study using a retrospective diagnostic design.
Methods: We used RobotReviewer as the index test and human reviewers' risk of bias assessment reported in Cochrane reviews as the reference test. A convenience sample of electronically available English-language full texts of RCTs included in Cochrane reviews with nurs* in the title were eligible for inclusion. In this context, we assessed random sequence generation, allocation concealment, and blinding (personnel or participants and assessors) corresponding to Cochrane risk of bias version 2011. Two independent research teams performed and double-checked data extraction and analysis. We calculated sensitivity, specificity, receiver operating characteristic (ROC) curve, the area under the ROC curve, predictive values, observed percentage of agreement, and Cohen's kappa (including confidence intervals, if applicable).
Findings: The selection process yielded 190 RCTs published between 1958 and 2016 in 23 Cochrane reviews published between 2000 and 2018. Missing assessments of risk of bias domains in Cochrane reviews or RobotReviewer yielded varying sample sizes per risk of bias domain. Sensitivity ranged from 0.44 to 0.88 and specificity from 0.48 to 0.95. Positive predictive value was highest for allocation concealment (0.79) and lowest for blinding assessors (0.25). Cohen's kappa was moderate for randomization (0.52), allocation concealment (0.60), and for blinding of personnel/patients (0.43). Blinding of outcome assessors had only slight agreement (0.04).
Conclusions: This is the first evaluation of risk of bias assessment by RobotReviewer in RCTs included in nursing-related Cochrane reviews. It yielded a moderate degree of agreement with human reviewers for randomization and allocation concealment, and an adequate sensitivity for detecting low risk of selection bias.
Clinical Relevance: Based on our results, using the RobotReviewer for risk of bias assessment in RCTs can be supportive in some risk of bias domains. However, human reviewers should supervise the semi-automated assessment process.

Systematic reviews lead to significant decisions in clinical practice (Ioannidis, 2016). Since the methodological quality of systematic reviews has not only highly scientific and practical relevance but also ethical relevance, the risk of bias assessment in systematic reviews is crucial to determine the internal validity of randomized controlled trials (RCTs). Bias is defined as systematic error of study results and is caused by incorrect research methods (Higgins & Altman, 2011). Therefore, the aim of risk of bias assessment is to critically appraise potential methodological flaws in reported research study results, which might lead to deviations from the true effect of an intervention on an outcome that would have been revealed without bias (Higgins et al., 2011). There are several types of bias in different study designs and several tools to assess risk of bias (Sterne et al., 2019). The most frequently used tool for assessing risk of bias in RCTs is the Cochrane risk of bias tool (Higgins & Green, 2008). The Cochrane Handbook methodologically guides Cochrane review authors in assessing risk of bias domains in RCTs. The fundamentals of risk of bias assessment are the judgment of systematic differences in baseline characteristics between treatment groups (selection bias), in the provision of care between treatment (performance bias), in study withdrawals between treatment groups (attrition bias), in outcome assessment between treatment groups (detection bias), and between reported and unreported findings (reporting bias; Higgins & Altman, 2011). The Cochrane risk of bias tool is widely used in both Cochrane and non-Cochrane systematic reviews of RCTs (Farrah, Young, Tunis, & Zhao, 2019; Sterne et al., 2019). For authors of Cochrane reviews, using the proposed risk of bias assessment tool is mandatory. Recently, the revised risk of bias assessment (RoB 2.0) was published with fundamental changes compared to the first and the updated version of the tool. The use of RoB 2.0 in upcoming Cochrane reviews is recommended but not mandatory (Sterne et al., 2019). Therefore, one can assume that risk of bias assessment following the Cochrane methodology proposed in 2011 is still in use and will be applied in upcoming reviews.

Conducting a risk of bias assessment imposes high demands on reviewers' expertise as well as on resources such as time and costs to conduct risk of bias assessment (Marshall & Wallace, 2019). These challenges have led to an increased development of electronic applications aimed to promote automated reviewing (O'Connor et al., 2019).

RobotReviewer is a freely available online tool devised to improve efficiency in the systematic review process (Marshall, Kuiper, & Wallace, 2016). It has been developed specially to support risk of bias assessment in RCTs for the Cochrane domains random sequence generation, allocation concealment, blinding of participants and personnel, as well as blinding of outcome assessors (Gates, Vandermeer, & Hartling, 2018). For each domain, the system determines whether a trial is at low risk of bias and identifies relevant text passages that support bias judgments of human reviewers. RobotReviewer performs a semi-automated risk of bias assessment in English-language RCTs using text analysis and machine learning (Marshall et al., 2016). Semi-automation means offering assessment suggestions and correcting them, if needed. It is important to underline that the tool makes no claim to correctness. However, it is intended for preliminary or supplementary assessment validated by human reviewers (Marshall & Wallace, 2019).

A retrospective evaluation study investigating the performance of RobotReviewer included 1,180 trials on different health topics. It demonstrated a Cohen's kappa agreement between 0.10 and 0.48 for the RobotReviewer's assessment in comparison with human reviewers among the different risk of bias domains. Furthermore, the analysis yielded a sensitivity between 0.28 and 0.76, and a specificity between 0.72 and 0.90 for detecting a low risk of bias in different domains (Gates et al., 2018). Regarding these results, RobotReviewer complemented by human reviewers' assessments might contribute to an efficient systematic review process. However, the sample consisted of RCTs from different health fields and did not consider specific scientific disciplines such as nursing. To focus on nursing seems necessary since interventions are characterized by complexity, specific contexts, and practical challenges such as sufficient blinding that might influence risk of bias assessment (Polit, 2011).

The accuracy of risk of bias assessment using RobotReviewer in the field of nursing remains unclear, and it is not known whether the tool might be useful for assessing risk of bias in nursing-related RCTs. We aimed therefore to evaluate the agreement between RobotReviewer's and human reviewers' risk of bias assessment.

Methods

Study Design

To evaluate RobotReviewer's performance, we applied a retrospective diagnostic study design. We used the Standards for Reporting Diagnostic Accuracy (STARD)

to structure this paper (Bossuyt et al., 2015). We did not prospectively register our study. Since we did not use clinical or patient data, ethical approval was not required.

Sample and Inclusion Criteria

We included a convenience sample of RCTs in Cochrane reviews containing nurs* in the title field. At the time of our research (August 30, 2018), the Cochrane library was under construction and showed irregularities. Therefore, we searched MEDLINE via PubMed for Cochrane reviews using the following search strategy: (nurs*[Title]) AND "The Cochrane database of systematic reviews"[Journal]. We included electronically available full texts of English-language RCTs (also cluster and cross-over RCTs).

Index and Reference Standard

The index test corresponded to the risk of bias assessment in RCTs using RobotReviewer (Marshall et al., 2016). For this study, we used RobotReviewer between September 27 and October 9, 2018. RobotReviewer is a freely available web-based machine learning tool developed by researchers of health and information sciences. It aims to support evidence synthesis through automatic data extraction and determination of risk of bias assessment. We inserted the full-text pdf of an RCT via the interface https://robotreviewer.vortext.systems/. RobotReviewer automatically determined the following four risk of bias domains: random sequence generation (selection bias), allocation concealment (selection bias), blinding of participants and personnel (performance bias), and blinding of outcome assessors (detection bias). RobotReviewer's results are instantly presented dichotomously per each domain by using the categories low and high or unclear risk of bias. The result sheet contains references to the text passages to justify the tool's decision per each domain. A concise description of the tool's functions and requirements is given elsewhere (Gates et al., 2018; Marshall et al., 2016). To ensure time-based data extraction, we downloaded RobotReviewer's result sheets containing risk of bias assessment containing the results of risk of bias assessment. Missing data for RobotReviewer's assessment occurred when RobotReviewer did not provide an assessment for some or all risk of bias domains.

As a reference standard, we used human reviewers' risk of bias assessments reported in Cochrane reviews. They followed a structured, highly standardized approach, for example, risk of bias assessment conducted by two independent persons and following assessment criteria published by Cochrane (Higgins & Green, 2011). We also extracted data of the reference test for the mentioned risk of bias domains. Since the RobotReviewer's output consists of a dichotomous risk of bias assessment using the categories low and high or unclear risk, we clustered high and unclear risk reported in Cochrane reviews to a combined risk of bias category. Missing data for Cochrane review's assessment were caused by the fact that Cochrane reviews did not provide an assessment on all the domains.

Data Extraction

J.H. and J.M. conducted data extraction. P.S. and G.M. double-checked this extraction for both risk of bias assessments (RobotReviewer and human reviewers). If blinding of participants or personnel and outcome assessors was combined in Cochrane reviews, we applied the respective judgment for both domains to match it with RobotReviewer's assessment. Since RobotReviewer distinguishes between low and high or unclear risk of bias, we categorized data emerging from Cochrane reviews according to low and high or unclear risk of bias. If there were missing data for a risk of bias domain (on the part of RobotReviewer or human reviewers), we did not consider this domain for analysis.

Analysis

We calculated RobotReviewer's performance using the following parameters and their 95% confidence intervals (95% CIs) based on cross tables for each of the four domains: sensitivity (the rate of correctly identified domains as low risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews), specificity (the rate of correctly identified domains as high or unclear risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews), positive predictive value (the rate of low risk of bias domains correctly identified by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews), negative predictive value (the rate of high or unclear risk of bias domains correctly identified by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews), Cohen's kappa (percentage of agreement between RobotReviewer's assessment and assessment by human reviewers reported in Cochrane reviews considering the possibility of agreement occurring by chance), and observed percentage of agreement (percentage of

agreement between RobotReviewer's assessment and human reviewers' assessment reported in Cochrane reviews).

Cohen's kappa was interpreted as poor (<0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), or very good (0.81–1.00) (Altman, 1991). In order to visualize the ratio between sensitivity and specificity, we generated receiver operating characteristic (ROC) curves as well as area under the ROC curves (AUC) and its 95% confidence interval (Hajian-Tilaki, 2014). We used SPSS version 24 (IBM Corp., Armonk, NY, USA) to analyze data. The SPSS file containing our data set is available upon reasonable request. We conducted no a priori sample size calculation, given the explorative character of this study.

Results

Sample

We identified 47 Cochrane reviews with nurs* in the title; 21 of them were updates or duplicates. In case of updates of Cochrane reviews, we included the latest version. Three of the Cochrane reviews did not contain any RCT. Finally, we included 23 Cochrane reviews containing 215 RCTs. We excluded 25 RCTs due to nonavailability of electronic full-text PDFs (n = 18), language (n = 4), and duplication (n = 3). We therefore included 190 RCTs in our analysis. Figure 1 illustrates the search and study selection process in detail. A list of the identified Cochrane reviews and included RCTs can be found in the supplementary material.

Study Characteristics

The Cochrane reviews were published between 2000 and 2018, half of them after 2013 and most of them in 2017. The included RCTs were published between 1958 and 2016, half of them after 2013 and most of them in 2003 and 2004. Of these RCTs, 188 (98.9%) were included in Cochrane reviews, referring their risk of bias assessment to the Cochrane Handbook methodology. RCTs were published in 98 different journals, mostly in the British Medical Journal (n = 21) and Age and Ageing (n = 10).

Test Results

The results (low risk, high or unclear risk, missing) of RobotReviewer's assessment (index test) and assessment by human reviewers reported in Cochrane reviews (reference test) in risk of bias domains are presented separately in Tables 1 and 2. Missing data occurred due to a lack of RobotReviewer's output (untraceable reasons) or a lack of assessment by human reviewers reported in Cochrane reviews. Table S1 illustrates assessments made by Cochrane reviewers and by RobotReviewer. Additionally, we provide justifications for RobotReviewer's assessment on Open Science Framework (OSF; see https://osf.io/hgpzw/, https://doi.org/10.17605/OSF.IO/HGPZW). Each folder in OSF

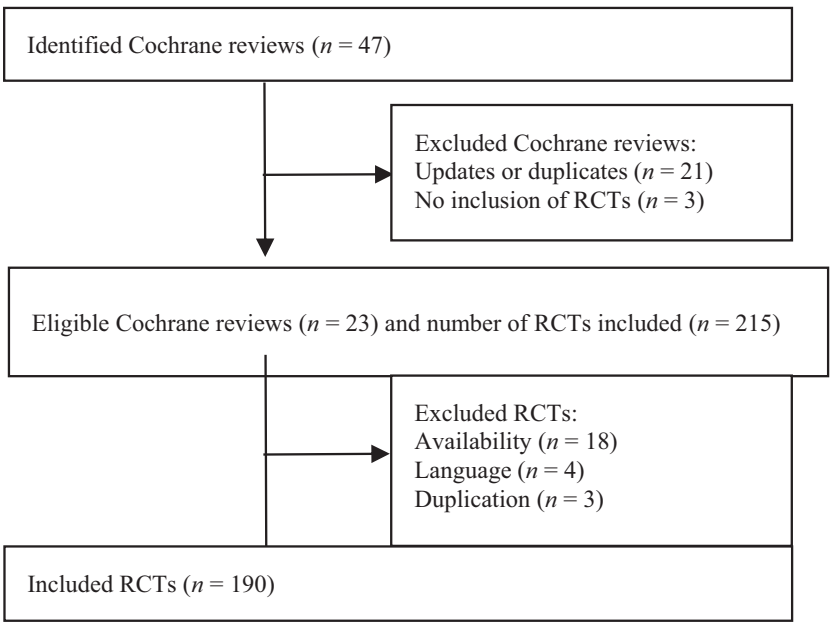

Figure 1. Search and study selection process.

Table 1. RobotReviewer's Assessment (Index Test, *N* = 190)

	Random sequence generation, *n* (%)	Allocation concealment, *n* (%)	Blinding of participants and personnel, *n* (%)	Blinding of outcome assessors, *n* (%)
Low risk	121 (64)	81 (43)	15 (8)	89 (47)
High/unclear risk	60 (32)	100 (53)	166 (87)	92 (48)
Missing	9 (5)	9 (5)	9 (5)	9 (5)

corresponds with the Cochrane review that contains the RCTs that we assessed using RobotReviewer. Table S2 is a list of Cochrane reviews with contained RCTs.

RobotReviewer's sensitivity, specifity, predictive values, Cohen's kappa, and percentage of observed agreement with the assessment by human reviewers reported in Cochrane reviews are illustrated in Table 3.

The rate of correctly identified domains as low risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews (sensitivity) was highest for random sequence generation (0.88; 95% CI 0.81, 0.95) and allocation concealment (0.77; 95% CI 0.68, 0.86). Thus, RobotReviewer might identify 8 to 10 of 10 studies concerning random sequence generation and personnel, and 7 to 9 of 10 studies concerning allocation concealment. The rate of correctly identified domains as high or unclear risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews (specificity) was highest for blinding of participants and personnel (0.95; 95% CI 0.90, 0.99) and allocation concealment (0.82; 95% CI 0.75, 0.90). Thus, RobotReviewer might identify a low risk of bias in 9 to 10 of 10 studies concerning blinding of participants and personnel, and 8 to 9 of 10 studies concerning allocation concealment. Highest rates of domains correctly identified as low risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews (positive predictive value) referred to allocation concealment (0.79; 95% CI 0.70, 0.88) and random sequence generation (0.77; 95% ci 0.69, .086). Highest rates of domains correctly identified as high or unclear risk of bias by RobotReviewer in correspondence with the assessment by human reviewers reported in Cochrane reviews (negative predictive values) concerned blinding of participants and personnel (0.91; 95% CI 0.74, 1.07), allocation concealment (0.81; 95% CI 0.72, 0.89]), and random sequence generation (0.78; 95% CI 0.69, 0.87). Cohen's kappa was poor for blinding of outcome assessors (0.04; 95% CI -1.14, 0.22) and moderate for blinding of participants and personnel (0.43; 95% CI 0.14, 0.72), random sequence generation (0.52; 95% CI 0.36, 0.68), and allocation concealment (0.60; 95% CI 0.48, 0.72). Percentage of observed agreement ranged between 50% and 87% for all four bias domains.

Figure S1 illustrates the ROC curves according to the four risk of bias domains. The AUC ranged between 0.75 for random sequence generation (95% CI 0.66, 0.84), 0.80 for allocation concealment (95% CI 0.73, 0.87), 0.69 for blinding of patients and personnel (95% CI 0.52, 0.78), and 0.53 for blinding of outcome assessors (95% CI 0.40, 0.66).

Discussion

Semi-automated tools supporting risk of bias assessment are part of international scientific discussions about why and how to automate steps in systematic reviews (Marshall & Wallace, 2019; O'Connor et al., 2019). This is due to the increasing number of studies potentially included in systematic reviews and high demands on methodological accuracy. To contribute to the discussion, we intended to evaluate the agreement in risk of bias assessment between RobotReviewer and human reviewers in Cochrane reviews. This topic has not yet been specifically addressed in the field of nursing science (Gates et al., 2018; Marshall et al., 2016). We assessed a convenience sample of 190 RCTs published between 1958 and 2016. Since we used consistent methods for calculating sensitivity, specificity, and Cohen's kappa, our results can be compared with a

Table 2. Assessment of Human Reviewers Reported in Cochrane Reviews (Reference Test, *N* = 190)

	Random sequence generation, *n* (%)	Allocation concealment, *n* (%)	Blinding of participants and personnel, *n* (%)	Blinding of outcome assessors, *n* (%)
Low risk	78 (41)	85 (45)	16 (8)	27 (14)
High/unclear risk	59 (31)	104 (55)	99 (52)	94 (50)
Missing	53 (28)	1 (5)	75 (40)	69 (36)

Table 3. RobotReviewer's Performance

	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	Cohen's kappa (95% CI)	Percentage of observed agreement
Random sequence generation (*n* = 129)	0.88 (0.81, 0.95)	0.62 (0.48, 0.75)	0.77 (0.69, 0.86)	0.78 (0.69, 0.87)	0.52 (0.36, 0.68)	78
Allocation concealment (*n* = 180)	0.77 (0.68, 0.86)	0.82 (0.75, 0.90)	0.79 (0.70, 0.88)	0.81 (0.72, 0.89)	0.60 (0.48, 0.72)	80
Blinding of participants and personnel (*n* = 109)	0.44 (0.19, 0.68)	0.95 (0.90, 0.99)	0.58 (0.30, 0.86)	0.91 (0.74, 1.07)	0.43 (0.14, 0.72)	87
Blinding of outcome assessors (*n* = 115)	0.58 (0.39, 0.77)	0.48 (0.38, 0.59)	0.25 (0.14, 0.35)	0.80 (0.70, 0.90)	0.04 (−1.14, 0.22)	50

Note. Sensitivity = the rate of correctly identified domains as low risk of bias by RobotReviewer in agreement with human reviewers; CI = confidence interval; specificity = the rate of correctly identified domains as high/unclear risk of bias by RobotReviewer in agreement with human reviewers; positive predictive value = the rate of low risk of bias domains correctly identified by RobotReviewer in agreement with human reviewers; negative predictive value = the rate of high/unclear risk of bias domains correctly identified by RobotReviewer in agreement with human reviewers; Cohen's kappa = percentage of agreement between RobotReviewer's and human reviewer's assessment considering the possibility of agreement occurring by chance; percentage of observed agreement = percentage of agreement between RobotReviewer's and human reviewers' assessment.

previous evaluation of the RobotReviewer (Gates et al., 2018).

Specificity was highest for blinding of participants and personnel as well as for allocation concealment. Similar results and highest values for specificity were reported by Gates et al. (2018) for blinding of participants and personnel as well as allocation concealment. Additionally, blinding of outcome assessors and random sequence generation yielded values above 0.70. However, this was different in our study, with random sequence generation and blinding of outcome assessors reaching values below 0.70. The correct identification of high or unclear risk of bias categories (specificity) needs further clarification on whether it should be judged as high or unclear risk of bias. Therefore, RobotReviewer's sensitivity might be of higher value for reviewers to answer the question of reliability concerning the tool's capacity for semi-automated risk of bias assessment (Gates et al., 2018). Sensitivity was highest for random sequence generation and allocation concealment. The study by Gates et al. yielded similar results concerning sensitivity for random sequence generation and allocation concealment. Sensitivity for blinding of participants and personnel as well as outcome assessors showed lower values. This is also the case in the results of Gates et al., even if blinding of outcome assessors yielded higher values in our study. Blinding of participants and personnel is often not possible in nursing interventions. Additionally, its judgment might be guided by a high degree of assessors' subjectivity. Therefore, we did not expect high accuracy of RobotReviewer's results for these two domains (Marshall & Wallace, 2019).

In our study, Cohen's kappa results were similar to the findings of Gates et al. (2018), moderate for randomization and blinding of participants and personnel, and slight agreement for blinding of outcome assessors. However, Cohen's kappa value for allocation concealment reached higher agreement in our study (0.60; 95% CI 0.48, 0.72) compared with the results of Gates et al. (0.45; 95% CI 0.4, 0.51).

Concerning the precision of our results, we found differences in comparison with the findings of Gates et al. (2018). The range of confidence intervals for sensitivity, specificity, and Cohen's kappa was much smaller in Gates et al. Klicken oder tippen Sie hier, um Text einzugeben than in our study. However, the reason for these remarkable differences and the lower precision of our results might be caused by the much larger sample size of 1,180 RCTs in Gates et al. Klicken oder tippen Sie hier, um Text einzugeben compared with our sample size and a maximum number of 180 RCTs. More precise results would have required a predefined and larger sample size (Daly, 2000, p. 143). Therefore, future studies could use our results to calculate a required sample size allowing robust results to ensure generalizability (Hajian-Tilaki, 2014).

Our results explicate the complexity of semi-automated risk of bias assessment in several ways. First, there are domains that might be less reliable with regard to a specific judgment of RobotReviewer. Therefore, reviewers should be aware whether to judge RobotReviewer's performance based on a specific result provided by the tool (positive and negative predictive values) or according to human reviewers' assessment in Cochrane reviews and the tools' capacity to judge correctly (sensitivity or specificity; Trevethan, 2017).

Second, some domains might be more suitable than others for using RobotReviewer as a supportive tool

in risk of bias assessment concerning nursing-related topics. For example, it might be useful to assess the risk of selection bias via RobotReviewer since the performance is quite accurate compared with blinding domains as indicated by ROC curves and corresponding AUC. Furthermore, random sequence generation and allocation concealment need to be assessed on an overall study level and not specifically for one or several outcomes. This might only be useful if reviewers need to assess several studies. The decision for or against using RobotReviewer might be up to reviewers. For example, reviewers might judge the tool's accuracy as sufficient or helpful for its supportive use when facing a high number of studies for risk of bias assessment or facing studies that are reported in a different structure than in scientific journals such as grey literature publications. Furthermore, a recent study indicates that using RobotReviewer might be less time consuming than manual assessment by human reviewers (Soboczenski et al., 2019).

Third, although the tool has not been developed to replace assessment by human reviewers (Marshall et al., 2016), research shows that RobotReviewer's reliability is comparable to that between two independent human reviewers (Hartling et al., 2013). However, such assumptions should be regarded with caution. Confirmation in further studies is required since it was not our aim to evaluate RobotReviewer's performance as a fully automated tool (Gates et al., 2018).

We used the risk of bias assessment in Cochrane reviews as a reference test in our study. The RCTs included in our study were part of Cochrane reviews published between 2000 and 2018. Therefore, they were subject to methodological recommendations existing at that time. However, half of the included Cochrane reviews were published after 2013. Here one could assume that methodological recommendations of the Cochrane Manual from 2011 have been applied. However, our analysis of human reviewers' assessments in Cochrane reviews yielded missing data among the four risk of bias domains, between 1% for allocation concealment and up to 40% for blinding of participants and personnel. Our missing data might be caused by nonmandatory use of all risk of bias domains in the past (Sterne et al., 2019) or deviations from the Cochrane methodology (Barcot, Boric, Dosenovic, et al., 2019; Barcot, Boric, Poklepovic Pericic, et al., 2019; Propadalo et al., 2018). Furthermore, research shows disagreements in risk of bias assessment by human reviewers. Bertizzolo, Bossuyt, Atal, Ravaud, and Dechartres (2019) stated low agreement for RCTs included in more than one Cochrane review, ranging from approximately 20% to 40% depending on risk of bias domain. This might indicate a lack of validity of our chosen reference standard. Corresponding to the statement by Puljak (2018), more precise information about the grading criteria provided in the included Cochrane reviews would be beneficial for interpreting our results. However, no data are available for this given our retrospective study design. Therefore, prospective diagnostic data would be needed.

Recently, the updated Cochrane Handbook for Systematic Reviews of Interventions was published (Higgins & Thomas, 2019). The revised tool for risk of bias assessment shows some fundamental changes. For example, reviewers must answer one or more signaling questions leading to low risk of bias, some concerns, or high risk of bias following a question-based algorithm. The assessment within each domain will be guided to an overall risk of bias judgement. Details are provided elsewhere (Sterne et al., 2019). The guidance on RoB 2.0 cannot yet be addressed by using RobotReviewer. Since (a) the risk of bias assessment methodology described in the Cochrane Handbook is most frequently used to assess the internal validity of a study and (b) the use of RoB 2.0 is not yet mandatory for Cochrane review authors, we can assume that the risk of bias assessment described in the 2011 version of the Cochrane Handbook might still be in use and will be used in future Cochrane and non-Cochrane systematic reviews (Farrah et al., 2019; Sterne et al., 2019). Furthermore, we should wait to see if and how many other reviews use the revised version RoB 2.0. Therefore, this study might support reviewers faced with the decision to use or not to use RobotReviewer as a semi-automated tool for risk of bias assessment.

To the best of our knowledge, this is the first evaluation of RobotReviewer in the context of a specific health topic such as nursing. The strengths of our approach are the transparent and reproducible inclusion of RCTs and independent double-checked data extraction of RobotReviewer's and human reviewers' risk of bias assessment. The explorative character of our study using a convenience sample of RCTs included in nursing-related Cochrane reviews might be a limitation concerning the generalizability of our results. Furthermore, the validity of our study is limited due to missing information in Cochrane reviews on human reviewers' level of experience in assessing risk of bias. Further research should therefore concentrate on prospective diagnostic studies considering the level of experience and professional background of human reviewers (Puljak, 2018).

Missing values are also limiting our study results. Missing risk of bias assessments by RobotReviewer were also reported in a previous study. This was unforeseeable and should be taken into account when using the tool (Gates et al., 2018). Missing values in Cochrane

reviews might have occurred due to inconsistent use of risk of bias domains in the past (Barcot, Boric, Dosenovic, et al., 2019; Barcot, Boric, Poklepovic Pericic, et al., 2019; Propadalo et al., 2018; Sterne et al., 2019). Given the explorative character of our study, we did not impute missing values.

Conclusions

This is the first evaluation of RobotReviewer's assessment of risk of bias in RCTs included in nursing-related Cochrane reviews. It yielded moderate agreement with human reviewers' assessment reported in Cochrane reviews for randomization and allocation concealment as well as an adequate sensitivity for detecting low risk of selection bias. Our study confirms the results of a previous study indicating that semi-automated assessment might be an accurate option to supplement human risk of bias assessment. However, higher precision and generalizability of the results requires an a priori sample size evaluated in a prospective study design. To avoid a lack of validity in the reference test, more precise information about grading criteria for human assessment should be ensured. Facing our results, some domains might be suitable for using RobotReviewer as a supportive tool for risk of bias assessment concerning nursing-related topics. However, human reviewers should supervise the assessment.

Acknowledgment

Open Access funding enabled and organized by ProjektDEAL.

Clinical Resources

- Cochrane Handbook for Systematic Reviews of Interventions (all versions). https://training.cochrane.org/cochrane-handbook-systematic-reviews-interventions
- RobotReviewer tool. https://www.robotreviewer.net/

References

Altman, D. G. (1991). *Practical statistics for medical research*. Boca Raton, FL: Chapman and Hall.

Barcot, O., Boric, M., Dosenovic, S., Pericic, T. P., Cavar, M., & Puljak, L. (2019). Risk of bias assessments for blinding of participants and personnel in Cochrane reviews were frequently inadequate. *Journal of Clinical Epidemiology*, *113*, 104–113. https://doi.org/10.1016/j.jclinepi.2019.05.012

Barcot, O., Boric, M., Poklepovic Pericic, T., Cavar, M., Dosenovic, S., Vuka, I., & Puljak, L. (2019). Risk of bias judgments for random sequence generation in Cochrane systematic reviews were frequently not in line with Cochrane Handbook. *BMC Medical Research Methodology*, *19*, 170. https://doi.org/10.1186/s12874-019-0804-y

Bertizzolo, L., Bossuyt, P., Atal, I., Ravaud, P., & Dechartres, A. (2019). Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: A research on research study using cross-sectional design. *BMJ Open*, *9*, e028382. https://doi.org/10.1136/bmjopen-2018-028382

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P., Irwig, L., … Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *British Medical Journal*, *351*, h5527. https://doi.org/10.1136/bmj.h5527

Daly, L. E. (2000). Confidence intervals and sample sizes. In D. G. Altman, D. Machin, T. N. Bryant, & M. J. Gardner (Eds.), *Statistics with confidence* (2nd ed., pp. 139–152). Bristol: BMJ Books.

Farrah, K., Young, K., Tunis, M. C., & Zhao, L. (2019). Risk of bias tools in systematic reviews of health interventions: An analysis of PROSPERO-registered protocols. *Systematic Reviews*, *8*, 280. https://doi.org/10.1186/s13643-019-1172-8

Gates, A., Vandermeer, B., & Hartling, L. (2018). Technology-assisted risk of bias assessment in systematic reviews: A prospective cross-sectional evaluation of the RobotReviewer machine learning tool. *Journal of Clinical Epidemiology*, *96*, 54–62. https://doi.org/10.1016/j.jclinepi.2017.12.015

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of Biomedical Informatics*, *48*, 193–204. https://doi.org/10.1016/j.jbi.2014.02.013

Hartling, L., Hamm, M. P., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., … Dryden, D. M. (2013). Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology*, *66*(9), 973–981. https://doi.org/10.1016/j.jclinepi.2012.07.005

Higgins, J. P. T., & Altman, D. G. (2011). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions: Cochrane book series. Version 5.1.0* (pp. 187–241). Chichester, UK: Wiley-Blackwell.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Juni, P., Moher, D., Oxman, A. D., … Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions. Version 5.0.0*. Retrieved from https://training.cochrane.org/handbook/archive/v5.0.0/

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions: Cochrane book series. Version 5.1.0*. Chichester, UK: Wiley-Blackwell. Retrieved from http://www.cochrane.org/handbook

Higgins, J. P. T., & Thomas, J. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions. Version 6* (2nd ed.). Hoboken, NJ: Wiley Online Library. Retrieved from https://training.cochrane.org/handbook

Ioannidis, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). Robotreviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, *23*, 193–201. https://doi.org/10.1093/jamia/ocv044

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*, 163. https://doi.org/10.1186/s13643-019-1074-9

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., Thayer, K. A., Shemilt, I., Thomas, J., … Wolfe, M. S. (2019). Still moving toward automation of the systematic review process: A summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *8*, 57. https://doi.org/10.1186/s13643-019-0975-y

Polit, D. F. (2011). Blinding during the analysis of research data. *International Journal of Nursing Studies*, *48*(5), 636–641. https://doi.org/10.1016/j.ijnurstu.2011.02.010

Propadalo, I., Tranfic, M., Vuka, I., Barcot, O., Pericic, T. P., & Puljak, L. (2018). In Cochrane reviews risk of bias assessments for allocation concealment were frequently not in line with Cochrane's Handbook guidance. *Journal of Clinical Epidemiology*, *106*, 10–17. https://doi.org/10.1016/j.jclinepi.2018.10.002

Puljak, L. (2018). Technology-assisted risk of bias assessment in systematic reviews requires precise definitions of risk of bias. *Journal of Clinical Epidemiology*, *99*, 168–169. https://doi.org/10.1016/j.jclinepi.2018.03.002

Soboczenski, F., Trikalinos, T. A., Kuiper, J., Bias, R. G., Wallace, B. C., & Marshall, I. J. (2019). Machine learning to help researchers evaluate biases in clinical trials: A prospective, randomized user study. *BMC Medical Informatics and Decision Making*, *19*, 96. https://doi.org/10.1186/s12911-019-0814-z

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *British Medical Journal*, *2*, l4898. https://doi.org/10.1136/bmj.l4898

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health*, *5*, 307. https://doi.org/10.3389/fpubh.2017.00307

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Figure S1. ROC-curves.

Table S1. Identified Cochrane Reviews and Included Studies.

Table S2. Full Assessments Made by Cochrane Reviewers and by RobotReviewer.