A mixed-methods review of the systematic reviewer's reliability using kappa statistics

Piet Hanegraaf[1], Abraham Wondimu[2], Jacob-Jan Mosselman[1], Gerrit de Jong[1], Seye Abogunrin[3], Luisa Queiros[3], Simon Van der Pol[2,4], Maarten Postma[2,4], Cornelis Boersma[2,4,5]

[1]Pitts, Zeist, the Netherlands; [2]Health-Ecore, Zeist, the Netherlands; [3]F. Hoffmann-La Roche Ltd., Basel, Switzerland; [4]Unit of Global Health, Department of Health Sciences, University Medical Center Groningen (UMCG), University of Groningen, The Netherlands; [5]Department of Management Sciences, Open University, Heerlen, The Netherlands.

INTRODUCTION

A combination of human and machine efforts has been proposed to reduce the workload of conducting systematic literature reviews (SLRs), and potentially enhance screening and data extraction quality.

As a first step to understanding the potential improvements machine approaches can bring, it is important to consider the quality of human-executed SLRs. One way to potentially determine how well the researchers involved in an SLR understood the topic being investigated is the reporting and level of the disagreements between the researchers[1]. This can be used as a proxy for the quality of the evidence summarized in such SLRs. Cohen's kappa is a way to measure inter-rater reliability (IRR)[2] and it is often used in SLRs to express the agreement levels between reviewers, taking into account not only the percentage of agreement between the reviewers, but also the chance of random agreement. The kappa score ranges between 0 and 1. A result between 0 and 0.20 means that there is no agreement between the reviewers and results of 0.90 or higher is considered as almost perfect agreement[3].

Notably, the kappa score can also potentially be used to compare the performance of machine learning approaches to those of human systematic reviewers. The systematic review of SLRs presented in this work is part of a mixed methods approach that aims to assess the quality of human-executed SLRs. This work on human performance in SLRs can assist in setting objectives for machine learning algorithms and creating a benchmark for determining the level of conflicts between machine learning algorithm performance and human performance.

METHODS

Using a systematic review approach in the PubMed database, we identified records of SLRs reporting the IRR, using the kappa statistic of the reviewers. For this part of the SLR study, we implemented a search query that limited results to only those records that contained the words kappa or Cohen's kappa.

Search strategy:

"(("Systematic Review" [Publication Type] OR "Meta-Analysis" [Publication Type]) AND ("randomized controlled trial*" OR "randomized clinical trial*" OR "controlled clinical trial*" OR "controlled trial*" OR "randomized" OR "trial") AND ("Cohen's kappa" OR "Cohen's Kappa statistic" OR "Cohen's kappa coefficient" OR "Cohen's K" OR "kappa test" OR "kappa*")) AND (y_5[Filter])"

To limit the scope of the review and to increase the comparability of the included studies, we only included systematic reviews of randomized controlled trials (RCTs). Protocol registration was required to ensure that only high-quality systematic reviews were considered (Table 1). Two reviewers independently screened the titles and abstracts of identified records. Following this, full texts of eligible abstracts were independently critically appraised to confirm eligibility. Data extraction was also independently performed by two reviewers based on pre-set extraction criteria. All records were screened and extracted using the Pitts web application (https://www.pitts.ai). The machine learning components of the Pitts web application were not employed for any steps of this systematic review.

Table 1: Study selection criteria used for the systematic review

SPIDE(R)	Inclusion Criteria
Sample	• Systematic literature reviews* of randomized controlled trials of pharmacological interventions • Two or more reviewers involved in literature screening and/or extraction • Publication in English or Dutch
Phenomenon of Interest	• Report level of agreement between reviewers
Design	• Double-blind screening • Double-blind data extraction • Reported number of in- and excluded studies • Protocol registered in PROSPERO or equivalent database
Evaluation	• Kappa score reported for at least one step of the SLR process or having data that can be used to calculate the kappa score

* Individual patient data meta-analysis, network meta-analysis, scoping review, realist reviews were not considered

RESULTS

The search was run on the 17th of October 2022. A total of 104 publications (Figure 1) were identified based on applying the search query over a 5-years retrospective time-window. From the records identified by the systematic review, most did not report a kappa statistic or another IRR.

As part of the study selection procedure, we excluded 46 records during abstract screening and another 51 manuscripts were excluded during full-text screening. For those records that did meet the inclusion criteria (n = 7), the data was not reported consistently for the different literature review steps. Kappa scores ranged from 0.62 to 0.98, with heterogeneity in reporting; some only recorded IRR for abstract or full-text screening or data extraction. All kappa scores were explicitly mentioned in the text, and we were not able to calculate any additional kappa scores if not reported in text for any additional studies.

Figure 1: Flow Diagram for the SLR

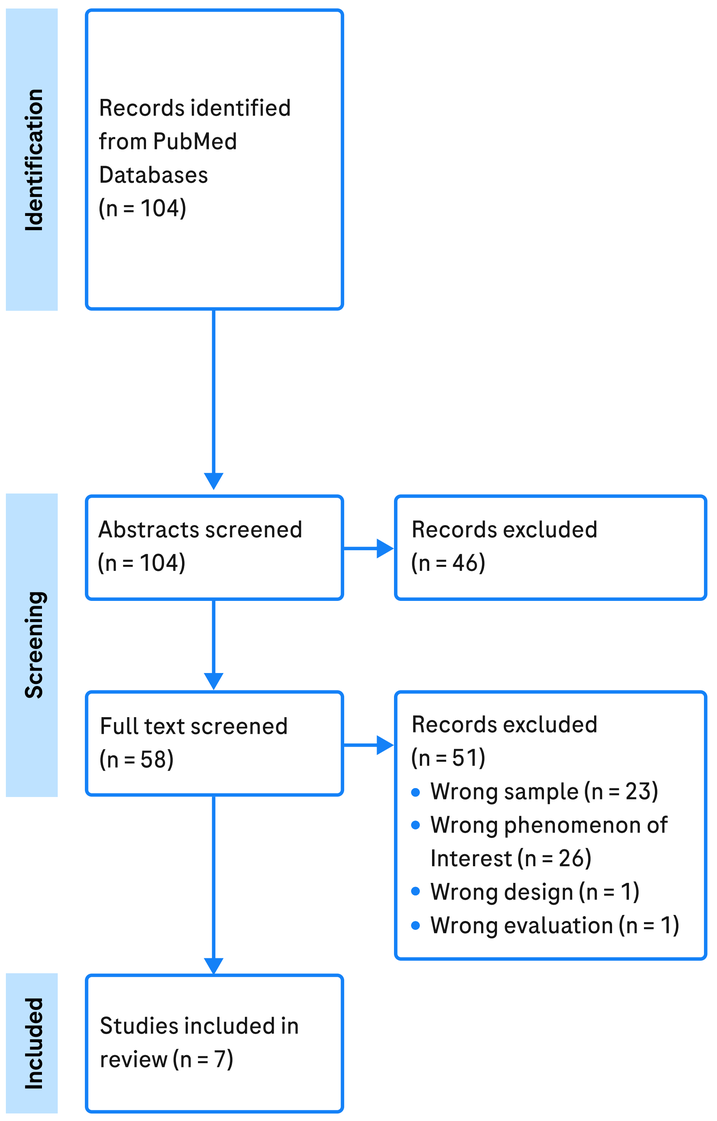

Table 2: Details on studies identified reporting the kappa score

Study	Kappa Score Reporting by SLR Step			
	Title and abstract screening	Full-text screening	Title and abstract and full text screening combined	Data extraction
Gupta et al. 2020 [4]	0.72	0.62	-	-
Cuyul-Vásquez et al. 2020 [5]	-	-	0.73	-
Barbari et al. 2020 [6]	-	0.82	-	-
Ainiwaer et al. 2021 [7]	-	-	0.84	0.98
Zhang et al. 2021 [8]	-	-	0.77	-
Allemann et al. 2017 [9]	-	-	-	0.92
Currell et al. 2019 [10]	-	-	0.89	-

DISCUSSION AND CONCLUSION

Our searches were restricted to identify records which mentioned the word "kappa", and were published in the last 5 years. Without adopting this approach, we would have had approximately 45,000 hits. However, to mitigate this, we designed the searches to attempt to identify studies with the mention of kappa reporting in any part of the article.

We found scarce reporting on kappa scores for IRR in systematic reviews of RCTs. In those studies that reported kappa statistics, variability in the kappa scores was found. Differences in researcher experience in combination with the difficulty of the screening task, time constraints on reviewing and level of preparation for screening and data extraction, likely drove variations in inter-reviewer kappa estimates. Future SLRs should report IRR kappa scores as a best scientific practice to showcase how well the researchers involved in SLRs understood the work they did.

This work is part of a mixed-methods approach, currently reporting on the first stage of the project. As a next step, we will apply a search query without 'kappa' related search terms and survey authors of systematic reviews on the perceived value of the kappa scores. These anticipated future results will build further on the quality of human-executed SLRs and the value-add of machine-assisted methods for conducting SLRs.

REFERENCES

1. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology. Sociol Methods Res. 2018 Sep 24;004912411879937.
2. Park CU, Kim HJ. Measurement of Inter-Rater Reliability in Systematic Review. Hanyang Med Rev. 2015;35(1):44.
3. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276–82.
4. Gupta S, Scheuter C, Kundu A, Bhat N, Cohen A, Facente SN. Smoking-Cessation Interventions in Appalachia: A Systematic Review and Meta-Analysis. Am J Prev Med. 2020 Feb;58(2):261–9.
5. Cuyul-Vásquez I, Berríos-Contreras L, Soto-Fuentes S, Hunter-Echeverría K, Marzuca-Nassr GN. Effects of resistance exercise training on redox homeostasis in older adults. A systematic review and meta-analysis. Exp Gerontol. 2020 Sep;138:111012.
6. Barbari V, Storari L, Ciuro A, Testa M. Effectiveness of communicative and educative strategies in chronic low back pain patients: A systematic review. Patient Educ Couns. 2020 May;103(5):908–29.
7. Ainiwaer A, Zhang S, Ainiwaer X, Ma F. Effects of Message Framing on Cancer Prevention and Detection Behaviors, Intentions, and Attitudes: Systematic Review and Meta-analysis. J Med Internet Res. 2021 Sep 16;23(9):e27634.
8. Zhang Y, Li J-J, Wang A-J, Wang B, Hu S-L, Zhang H, et al. Effects of intensive blood pressure control on mortality and cardiorenal function in chronic kidney disease patients. Ren Fail. 2021 Dec;43(1):811–20.
9. Allemann SS, Nieuwlaat R, Navarro T, Haynes B, Hersberger KE, Arnet I. Congruence between patient characteristics and interventions may partly explain medication adherence intervention effectiveness: an analysis of 190 randomized controlled trials from a Cochrane systematic review. J Clin Epidemiol. 2017 Nov;91:70–9.
10. Currell SD, Liaw A, Blackmore Grant PD, Esterman A, Nimmo A. Orthodontic mechanotherapies and their influence on external root resorption: A systematic review. Am J Orthod Dentofacial Orthop. 2019 Mar;155(3):313–29.