Inter-reviewer reliability of literature screening for human and machine-assisted systematic reviews: A mixed methods review

MSR27

Hanegraaf P [1] , Wondimu A [2] , Mosselman JJ [1] , de Jong G [1] , Abogunrin S [3] , Queiros L [3] , Lane M [3] , Postma MJ [2,4,5,6] , Boersma C [2,4,7] , van der Schans J [2,4,5,7]

[1.] Pitts, Zeist, the Netherlands; [2.] Health-Ecore B.V., Zeist, The Netherlands; [3.] F. Hoffmann-La Roche, Basel, Switzerland; [4.] Unit of Global Health, Department of Health Sciences, University Medical Center Groningen (UMCG), University of Groningen, The Netherlands; [5.] Department of Economics, Econometrics & Finance, University of Groningen, Faculty of Economics & Business, Groningen, The Netherlands; [6.] Department of Pharmacy, Unit of PharmacoEpidemiology & PharmacoEconomics, University of Groningen, Groningen, Netherlands; [7.] Department of Management Sciences, Open University, Heerlen, The Netherlands;

TAKE AWAY

A minimal strong agreement between reviewers of machine learning assisted SLRs is recommended to ensure overall acceptance of machine learning in SLRs.

BACKGROUND

Synthesis of available evidence by means of a systematic literature review (SLR) forms the foundation to informed medical decision-making, making it one of the most important sources of evidence. [1] The rigorous character of SLRs combined with the increasing volume of evidence and the need for systematic updates to prevent evidence to become outdated [2,3] has put excessive pressure on researchers involved in evidence generation and assessment. The potential impact of outdated and incomplete health information goes beyond the field of evidence generation and is likely to result in sub-optimal treatment of patients. As a result, automating aspects of the SLR process could lead to better and up-to-date informed medical decision-making, and thus indirectly improve the health of individual patients and entire populations. Machine learning automation efforts have been proposed to reduce the workload and potentially enhance the quality of SLRs. However, the level of inter-reviewer reliability (IRR) in both human and machine learning automated SLRs is yet unclear. The need for rigorously produced, disseminated, and easily accessed evidence of machine learning validation is one of the key aspects to advance the field of machine learning in evidence synthesis. The demonstration of accuracy versus human classification is one of the first steps in the significant introduction of machine learning into evidence synthesis. [4]

AIM

The aim of this work is to assess the level of agreement of human-executed SLRs and to assist in setting objectives for machine learning algorithms and creating a benchmark for determining the level of conflicts between machine learning algorithm classification and human classification.

METHODS

This mixed methods review consists of two parts to create a comprehensive synthesis of quantitative and qualitative data of IRR for screening and data extraction in SLRs. In the first part of this study, we performed a review of systematic literature reviews of randomised controlled trials to assess the IRR reported in SLRs. The Pitts web application (www.pitts.ai) was used for both the literature screening as well as the data extraction. [5]

The Protocol of the review was registered in PROSPERO: CRD42023386706.

In the second part of this research, we surveyed authors of SLRs on their expectations of machine learning automation and human performed inter-reviewer reliability in SLRs to determine the expected IRR by authors of SLRs by means of a survey on:

- Expectations on the IRR of SLRs between two human reviewers
- Expectations on the IRR of SLRs between a human reviewer and a machine learning agent
- Participants' own experience on presenting the IRR of their SLRs

RESULTS

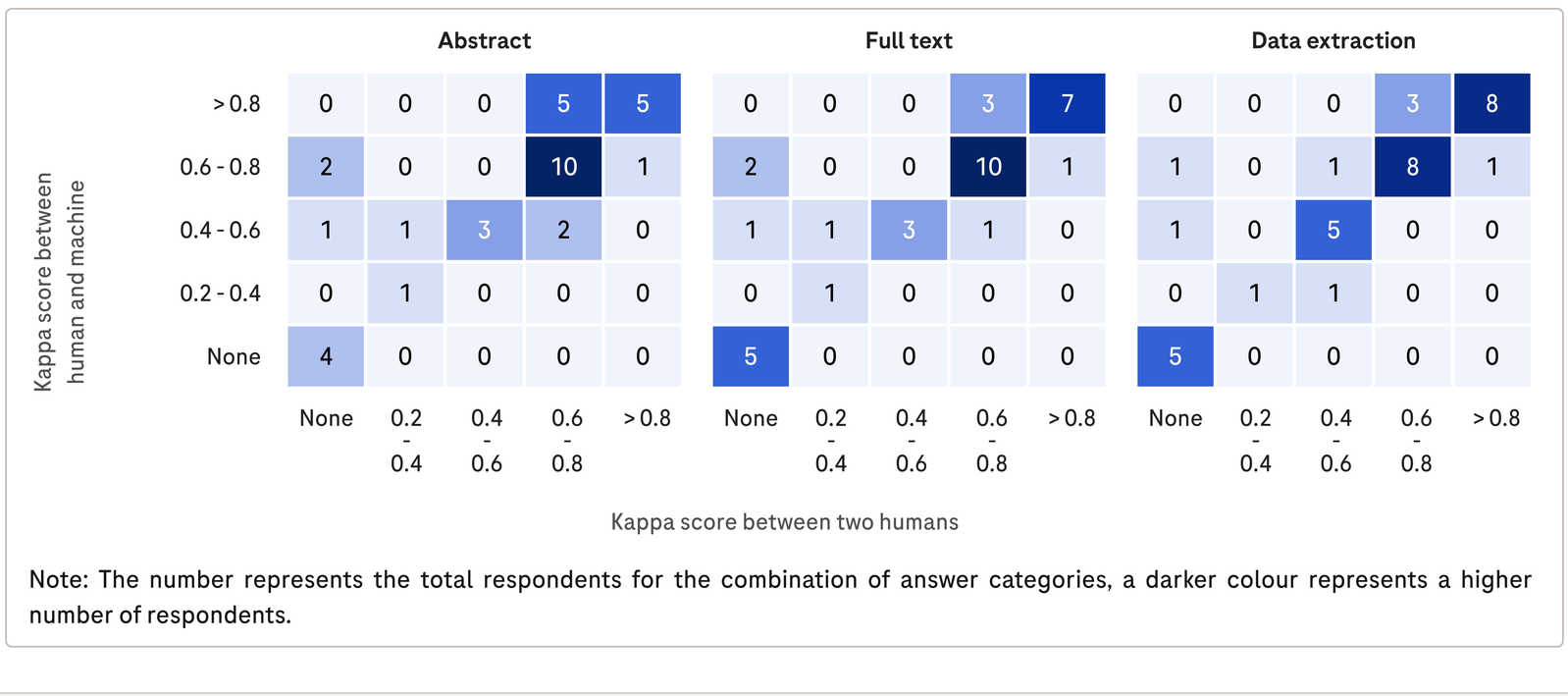
A total of 836 records were identified based on applying the search query after removal of duplicates. As part of the study selection procedure, we **excluded 423 records during abstract screening** and another **363 studies were excluded during full-text screening**. The primary reason for exclusion at full-text level was that the full text study did not report on the level of agreement between reviewers (n=307). In total, **45 articles met the eligibility criteria** and were included in this review. The Cohen's kappa scores of this study between reviewers for the abstract screening and full text screening were both 0.72.

The average Cohen's kappa score reported was:

- **0.82 (SD= 0.11, n=12)** for abstract screening
- **0.77 (SD= 0.18, n=14)** for full text screening
- **0.86 (SD=0.07, n=15)** for the whole screening process
- **0.88 (SD=0.08, n=16)** for data extraction

In total, **37 respondents completed the survey**. The survey results for the expected IRR showed **overlapping expected Cohen's kappa values compared to the SLR performed ranging between approximately 0.6-0.9**, indicating a moderate to strong agreement between reviewers. The majority of the respondents indicated that **automated literature screening systems should be above average with respect to their ability to screen (72.97%) or extract (70.27%)** accurately before they should be used for peer-reviewed and journal-published SLRs. No trend was observed between reviewer experience (i.e., scientific experience, publication experience SLRs, machine learning experience SLRs, and screening and data extraction ability) and expected IRR in both human based SLR and machine learning assisted SLR. An important common response on not reporting IRR was the fact that the reviewers included a third assessor to solve disagreement and the IRR only reflects part of the consensus building process. Misinterpretation of the IRR as a quality measure of the reviewing process was mentioned as another underlying reason to not record or report IRR in SLRs.

Figure 1: Lowest acceptable agreement expressed in Cohen's Kappa score between two reviewers (human-human or human-machine learning agent) for double-blinded literature abstract screening, full text screening, and data extraction decisions to be published in a scientific journal



Note: The number represents the total respondents for the combination of answer categories, a darker colour represents a higher number of respondents.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- First assessment of threshold of agreement between human reviewers of systematic literature reviews
- First reference for a threshold of agreement for machine learning assisted systematic literature reviews
- Underreporting of inter-reviewer reliability metrics may not accurately reflect the true agreement
- Sample size of the survey is small, undermining both the internal and external validity

CONCLUSION

Human performed SLRs likely show a moderate agreement between reviewers, while authors expect machine learning assisted SLRs to perform better. This mixed-methods review gives first guidance on the human IRR benchmark, which could be used as a minimal threshold for IRR in machine learning assisted SLRs. A minimal strong agreement between reviewers of machine learning assisted SLRs is recommended to ensure overall acceptance of machine learning in SLRs.

REFERENCES

1. Gough D, Elbourne D. Systematic research synthesis to inform policy, practice and democratic debate. Soc Policy Soc. 2002;1(3):225-236. https://doi.org/10.1017/S147474640200307X
2. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007 Aug 21;147(4):224-33. doi: 10.7326/0003-4819-147-4-200708210-00179.
3. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, Salanti G, Meerpohl J, MacLehose H, Hilton J, Tovey D, Shemilt I, Thomas J; Living Systematic Review Network. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol. 2017 Nov;91:23-30. doi: 10.1016/j.jclinepi.2017.08.010.
4. Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. Syst Rev. 2021 Jan 8;10(1):16. doi: 10.1186/s13643-020-01569-2. [8] Park CU, Kim HJ. Measurement of Inter-Rater Reliability in Systematic Review. Hanyang Med Rev. 2015;35(1):44.
5. Living systematic review software | Pitts [Internet]. [cited 2022 Nov 24];Available from: https://pitts.ai/

Would you like to contribute to our discrete choice experiment on the accuracy of AI in SLRs? Scan the QR-code for more information.