Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human-GenAI Comparison of a COSMIN Systematic Review: A Swiss Case Study

Matthew Kerry

kerr@zhaw.ch

ZHAW Zurich University of Applied Sciences

Thanh Elsener

ZHAW Zurich University of Applied Sciences

Short Report

Keywords: Patient-reported outcome measures, COSMIN, artificial intelligence, ChatGPT, systematic review, screening, measurement properties

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11465622/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Artificial intelligence may accelerate systematic reviews of patient-reported outcome measures (PROMs), but its ability to identify eligible studies and reproduce Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) assessments remains uncertain.

Methods

ChatGPT performance was compared with human reference decisions in a systematic review of generic PROMs relevant to Switzerland. Screening performance was evaluated across 4,374 title/abstract records and 1,512 full-text reports, of which 50 were ultimately included by human reviewers. Sensitivity, specificity, precision (positive predictive value), negative predictive value, and accuracy were calculated. A secondary, post-hoc analysis incorporated 12 human-verified corrections. Agreement on COSMIN measurement-property ratings, evidence quality, and overall recommendations was evaluated for five instruments.

Results

At title/abstract screening, ChatGPT correctly advanced 76 of 1,499 human-positive records (sensitivity 5.1%; precision 97.4%; specificity 99.9%). At full-text screening, ChatGPT correctly included 17 of 50 human-included studies and incorrectly included 11, yielding sensitivity of 34.0%, precision of 60.7%, and specificity of 99.2%. Following 12 human-verified corrections, sensitivity increased to 58.0% and precision to 72.5%. Agreement on matched COSMIN measurement-property ratings ranged from 52.9% (PROMIS) to 86.8% (SF-36), whereas evidence-quality agreement ranged from 34.6% (SF-36) to 62.5% (WORQ). Overall instrument recommendations agreed for four of five instruments (80%).

Conclusions

ChatGPT demonstrated low sensitivity for identifying eligible studies, despite high specificity. COSMIN agreement varied substantially, particularly for evidence-quality assessments. Human verification remains essential for reliable AI-assisted PROM systematic reviews.

Background

Systematic reviews of patient-reported outcome measures (PROMs) require comprehensive literature identification, eligibility screening, critical appraisal, and synthesis of measurement-property evidence. The COSMIN methodology provides structured standards for evaluating instrument validity, reliability,

responsiveness, and the certainty of supporting evidence. These tasks are resource-intensive and may be candidates for support by large language models (LLMs). [1, 2]

However, the consequences of screening errors are asymmetric. Incorrect exclusions may remove relevant evidence from a review, while incorrect inclusions increase the burden of subsequent appraisal. Moreover, correctly reproducing an overall instrument recommendation does not necessarily demonstrate accurate assessment of individual measurement properties or certainty of evidence. [2, 3] Machine-learning-assisted screening methods have also been proposed to prioritize review records while retaining human oversight [8].

We examined ChatGPT's performance during a systematic review of generic PROMs relevant to Switzerland. The objectives were to quantify agreement with human screening decisions and to compare ChatGPT-generated COSMIN ratings and instrument recommendations against human assessments.

Methods

Review workflow and reference standard

The systematic review considered generic PROMs, including SF-36, EQ-5D, WHOQOL-BREF, PROMIS, and WORQ. The search strategy covered PubMed, CINAHL, Embase, and additional sources, followed by title/abstract and full-text screening. [1, 2] The published Swiss COSMIN review by Elsener, Kerry, and Biller-Andorno provides the human-review study context against which the AI-assisted workflow was compared [6].

Human reviewer decisions were designated the definitive reference standard. The title/abstract comparison comprised 4,374 records, and the full-text comparison used 1,512 reports, of which 50 were ultimately included by human reviewers. ChatGPT classifications were compared with the human decisions without retrospectively changing the reference standard. The published review is cited as the source study; the screening counts and comparison metrics reported here derive from the study-specific human–AI comparison dataset, rather than being attributed to the publication [6].

Screening performance

We constructed 2 × 2 confusion matrices and calculated sensitivity, specificity, precision (PPV), NPV, and overall accuracy. The primary full-text analysis counted erroneous ChatGPT identification or exclusion decisions as errors. In a secondary, post-hoc analysis, 12 human-verified errors were accepted as corrections, with all remaining classifications unchanged. Duplicate study identifiers were reconciled using the confirmed study-level overlap table.

COSMIN assessment

Human and ChatGPT assessments were compared for content validity, structural validity, internal consistency, reliability, measurement error, construct validity, and responsiveness, where applicable. The analyses distinguished measurement-property ratings from methodological quality or certainty-of-evidence ratings. [1]

Agreement was calculated for matched study–property pairs. Additional analyses counted unmatched ratings as disagreements. Linear weighted Cohen's kappa was used where estimable. Overall COSMIN evidence syntheses and instrument recommendation categories were compared separately. [1]

Results

Search and screening performance

The review comprised 4,374 title/abstract records. The human reviewers advanced 1,512 reports to full-text review, with 50 ultimately included. In the supplied abstract-stage confusion matrix, ChatGPT correctly advanced 76 records, incorrectly advanced two, and failed to advance 1,423 records selected by human reviewers.

At title/abstract screening, sensitivity was 5.1%, specificity 99.9%, and PPV 97.4%. At full-text screening, ChatGPT correctly included 17 studies and incorrectly included 11, while missing 33 of the 50 human-included studies (Table 1).

Table 1
Screening confusion matrices

Title/abstract screening	**Human include**	**Human exclude**	**Total**
ChatGPT include	76	2	78
ChatGPT exclude	1,423	2,873	4,296
Total	1,499	2,875	4,374
Full-text screening			
ChatGPT include	17	11	28
ChatGPT exclude	33	1,451	1,484
Total	50	1,462	1,512

Note: The abstract-stage confusion matrix identifies 1,499 human-positive records, whereas the review workflow reports 1,512 reports advanced to full text. This discrepancy requires reconciliation before submission.

Following acceptance of 12 human-verified corrections, the number of true positives increased from 17 to 29, with sensitivity increasing from 34.0% to 58.0% and PPV from 60.7% to 72.5% (Table 2). Specificity remained 99.2%.

Table 2
Screening performance by stage and correction status

Metric	Title/abstract	Full text: original	Full text: corrected
Sensitivity	5.07%	34.00%	58.00%
Specificity	99.93%	99.25%	99.25%
Precision (PPV)	97.44%	60.71%	72.50%
NPV	66.88%	97.78%	98.57%
Accuracy	67.42%	97.09%	97.88%

Note: The corrected analysis reclassifies 12 previously missed human-included studies as true positives. It is a post-hoc, human-assisted analysis rather than independent AI performance.

The relatively high overall accuracy should be interpreted in the context of the low inclusion prevalence (50/1,512; 3.3%). Excluding every report would itself yield 96.7% accuracy, despite identifying no eligible studies. [2]

COSMIN agreement

Agreement on individual COSMIN measurement-property ratings varied substantially between instruments (Table 3). Matched-row agreement ranged from 52.9% for PROMIS to 86.8% for SF-36. Agreement on evidence-quality assessments was generally lower, ranging from 34.6% for SF-36 to 62.5% for WORQ.

Table 3
Agreement between human and ChatGPT COSMIN assessments

Instrument	Measurement-property agreement	Evidence-quality agreement	Human recommendation	ChatGPT recommendation
WORQ	75.0% (6/8)	62.5% (5/8)	B	B
PROMIS	52.9% (9/17)	58.8% (10/17)	B	B
WHOQOL-BREF	71.4% (5/7)	42.9% (3/7)	B	C
EQ-5D	72.2% (13/18)	44.4% (8/18)	C	C
SF-36	86.8% (33/38)	34.6% (9/26)	A	A

Note: Percentages are based on matched study–property rows, with denominators differing by instrument and assessment type. Recommendation categories are those recorded in the review workbook.

For content-validity assessments, observed agreement on property ratings was 20.0% for WORQ, 66.7% for PROMIS, 57.1% for WHOQOL-BREF, and 33.3% for SF-36. Agreement on methodological-quality ratings was 0.0%, 33.3%, 42.9%, and 50.0%, respectively. The EQ-5D content-validity comparison did not contain corresponding matched ratings.

When unmatched measurement-property rows were counted as disagreements, agreement decreased to 50.0% for WORQ, 40.9% for PROMIS, 50.0% for WHOQOL-BREF, 46.4% for EQ-5D, and 42.7% for SF-36.

Overall instrument recommendations

The human assessments recommended SF-36 in category A, WORQ and PROMIS in category B, WHOQOL-BREF in category B, and EQ-5D in category C. ChatGPT assigned the same recommendations to SF-36, WORQ, PROMIS, and EQ-5D, but rated WHOQOL-BREF as category C. Thus, four of five overall recommendations agreed (80%).

This relatively high recommendation-level agreement contrasted with substantially lower agreement in the underlying measurement-property and evidence-quality assessments.

Discussion

This case study identifies two distinct limitations of ChatGPT-assisted PROM systematic reviews. First, screening sensitivity was low, particularly at the title/abstract stage. Although precision and specificity were comparatively high, these metrics cannot compensate for the omission of large numbers of human-included records. In evidence synthesis, such false negatives may systematically distort the available evidence base. [3–6]

Second, COSMIN agreement depended strongly on the level of assessment. ChatGPT frequently reproduced the same overall recommendation while differing from human reviewers on individual measurement-property ratings and certainty-of-evidence judgments. This suggests that apparent agreement at the instrument-recommendation level may conceal substantive differences in the supporting appraisal. [1]

The post-hoc correction analysis demonstrated improved screening sensitivity but should not be interpreted as independent model performance. The improvement depended on human identification and correction of ChatGPT errors. [4,5] Reports of generative-AI hallucinations further support verification of AI-generated eligibility rationales and extracted evidence [7].

Strengths include the use of a complete human reference standard and comparisons across screening, study-level appraisal, and instrument-level recommendations. Limitations include the single-review setting, a limited number of instrument categories, and the potential dependence of results on the model configuration, prompts, and supplied documents. The full-text comparison also assumes that all 1,512 reports were available for AI assessment. [3–6]

Conclusions

ChatGPT showed limited sensitivity in systematic-review screening and variable agreement with human COSMIN appraisals. Agreement was particularly inconsistent for evidence-quality assessments, even when final instrument recommendations matched. Generative AI may assist with selected review tasks but should not independently determine study eligibility or COSMIN evidence synthesis without human verification.

Declarations

Author Contribution

MK conceived of study. TE conducted initial scoring analyses. MK reviewed methodology and conducted initial writeup. TE reviewed write-up.

References

1. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27:1147–1157. doi:10.1007/s11136-018-1798-3.
2. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.

3. Khraisha Q, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15:616–626. doi:10.1002/jrsm.1715.
4. Purewal A, Fautsch K, Klasova J, Hussain N, D'Souza RS. Human versus artificial intelligence: evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. Regional Anesthesia & Pain Medicine. 2026 Apr;51(4):437 – 42.
5. Martins AM, Juan LF, Oliveira A, Martins JP, Santos M. Can a chatbot be used in the full-text screening in a systematic review?. International Journal of Medical Informatics. 2026 Apr 2:106422.6. Elsener T, Kerry M, Biller-Andorno N. A COSMIN systematic review of generic patient-reported outcome measures in Switzerland. Qual Life Res. 2025;34(7):1869–1884.
6. Elsener, T., Kerry, M., & Biller-Andorno, N. (2025). A COSMIN systematic review of generic patient-reported outcome measures in Switzerland. *Quality of Life Research*, *34*(7), 1869–1884.
7. Adel A, Alani N. Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT. AI Soc. 2025;40:6799–6812. doi:10.1007/s00146-025-02406-7.
8. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3:125–133. doi:10.1038/s42256-020-00287-7.