Accuratezza IA Generativa: Valutazione Rischio di Bias e Data Extraction

Luca Carrer, Niccolò Maschi, Tiziano Innocenti, Stefano Salvioli

INTRODUZIONE E OBBIETTIVI

Le revisioni sistematiche costituiscono un punto fondamentale della pratica basata sull'evidenza, ma sono spesso caratterizzate da elevati costi temporali. I recenti sviluppi in ambito di intelligenza artificiale, in particolare dei modelli linguistici di grandi dimensioni come ChatGPT-4o, offrono nuove opportunità per supportare e automatizzare parzialmente alcuni processi. Il presente studio mira a valutare le prestazioni di ChatGPT-4o nella valutazione del Rischio di Bias (RoB) tramite strumento RoB 2.0 e nell'estrazione dati da trial controllati randomizzati (RCT) sull'esercizio fisico per il low back pain cronico (CLBP).

MATERIALI E METODI

Questo studio trasversale ha incluso 150 RCT precedentemente analizzati da revisori umani. ChatGPT-4o è stato testato in due compiti distinti: (1) valutazione del RoB, effettuata tramite un singolo prompt strutturato e, successivamente, confrontata con le valutazioni degli esperti nei cinque domini e nel giudizio complessivo; (2) estrazione dati su 34 variabili predefinite, utilizzando sia un prompt semplificato sia uno dettagliato. Le valutazioni dei revisori umani sono state assunte come gold standard. L'analisi statistica ha incluso il coefficiente k di Cohen e l'accuratezza complessiva, oltre a sensibilità, specificità, valore predittivo positivo (PPV), valore predittivo negativo (NPV) e F1-score.

RISULTATI

L'accordo tra ChatGPT-4o e i revisori umani nella valutazione del RoB è risultato basso (kappa di Cohen = 0,14), con una tendenza alla sottostima del rischio di bias. La sensibilità si è attestata al 40%, mentre specificità e PPV sono risultati superiori per le classificazioni a basso rischio. Per quanto riguarda l'estrazione dati, ChatGPT-4o ha dimostrato prestazioni solide, con un'accuratezza superiore all'84% e un F1-score oltre il 90% per entrambe le modalità di prompt. Le allucinazioni sono state rare (<0,02%). Il tempo medio per completare i compiti è passato da 30-50 minuti (revisori umani) a meno di 5 minuti con ChatGPT-4o.

CONCLUSIONI

ChatGPT-4o ha mostrato un'affidabilità limitata nella valutazione del RoB, probabilmente a causa della complessità interpretativa di tale compito. Tuttavia, ha fornito solidi risultati nell'estrazione dati, specialmente quando guidato da prompt accuratamente progettati. L'elevata precisione, la rapidità di esecuzione e la bassa incidenza di allucinazioni evidenziano il potenziale del modello come revisore secondario nei flussi di lavoro delle revisioni sistematiche. Rimane tuttavia prematuro il suo impegno in compiti valutativi che richiedono un giudizio critico. Con ulteriori sviluppi e un'adeguata supervisione umana, i modelli linguistici di grandi dimensioni come ChatGPT-4o potrebbero contribuire in modo significativo all'ottimizzazione dei processi di revisione sistematica.

BIBLIOGRAFIA

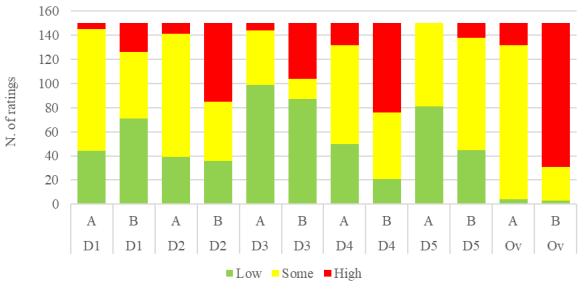
1. Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):E2412687. doi:10.1001/jamanetworkopen.2024.12687
2. Hayden JA, Ellis J, Ogilvie R, Malmivaara A, van Tulder MW. Exercise therapy for chronic low back pain. Cochrane Database of Systematic Reviews. 2021;2021(9). doi:10.1002/14651858.CD009790.pub2
3. Innocenti T, Hayden JA, Salvioli S, et al. Bias in the measurement of the outcome is associated with effect sizes in randomized clinical trials on exercise therapy for chronic low back pain: a meta-epidemiological study. J Clin Epidemiol. 2023;162:145-155. doi:10.1016/j.jclinepi.2023.09.001
4. Motzfeldt Jensen M, Brix Danielsen M, Riis J, et al. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. PLoS One. 2025;20(1):e0313401. doi:10.1371/journal.pone.0313401

Rob 2 domain	Agreement	Cohen's Kappa (95% CI)
D1	56%	0,28 (0,12-0,43)
D2	28%	-0,04 (-0,19-0,11)
D3	43%	0,01(-0,15-0,17)
D4	39%	0,12(-0,03-0,27)
D5	56%	0,20 (0,04-0,35)
Overall	29%	0,04 (-0,11-0,19)

Rob 2 domain	Sensibility (95% CI)	Specificity (95% CI)
D1	0,56 (0,42-0,69)	0,78 (0,67-0,88)
D2	0,28 (0,12-0,43)	0,64 (0,51-0,76)
D3	0,39 (0,24-0,53)	0,70 (0,58-0,81)
D4	0,37 (0,22-0,51)	0,68 (0,56-0,79)
D5	0,54 (0,40-0,67)	0,77 (0,66-0,87)
Overall	0,25 (0,09-0,40)	0,26 (0,10-0,41)

Rob 2 domain	PPV (95% CI)	NPV (95% CI)
D1	0,56 (0,42-0,69)	0,78 (0,67-0,88)
D2	0,28 (0,12-0,43)	0,64 (0,51-0,76)
D3	0,39 (0,24-0,53)	0,70 (0,58-0,81)
D4	0,37 (0,22-0,51)	0,68 (0,56-0,79)
D5	0,54 (0,40-0,67)	0,77 (0,66-0,87)
Overall	0,25 (0,09-0,40)	0,26 (0,10-0,41)

Metric	Short Prompt (95% CI)	Long Prompt (95% CI)
Overall Accuracy (%)	84,1	84,02
Error rate (%)	15,90 (-0,82-30,82)	15,98 (-0,82-30,82)
Choen's K	0,43 (0,44-0,42)	0,43 (0,44-0,42)
Sensitivity	0,83 (0,82-0,84)	0,83 (0,82-0,84)
Specificity	0,88 (0,85-0,91)	0,88 (0,85-0,91)
PPV	0,98 (0,982-0,989)	0,98 (0,97-0,98)
NPV	0,35 (0,33-0,38)	0,41 (0,38-0,42)
F1-score (%)	90,5 (83,6-97,3)	90,2 (83,2-97,1)



Comparison of the categorization of risk of bias assessments in the five domains and overall between ChatGPT and Authors: A: ChatGPT; B: Authors

Results for each category of data extraction