Avaliação Crítica de Revisões Sistemáticas: Investigadores *versus* Inteligência Artificial

67

Carlota Duarte de Mendonça[1,2,3], Inês Cardoso[1], Bruno Rosa[1,2,3], Susana Dias[1,2], Ruben Pereira[1,2], António Mata[1,2,3]

carlotamendonca@edu.ulisboa.pt

[1] Faculdade de Medicina Dentária (FMDUL), Universidade de Lisboa, Lisboa, Portugal;
[2] Oral Biology and Biochemistry Research Group at LIBPhys-FCT UIDB/04559/2020 da FMDUL, Lisboa, Portugal;
[3] Center for Evidence-Based Dental Medicine da FMDUL, Lisboa, Portugal.

INTRODUÇÃO E OBJETIVO

A avaliação crítica de revisões sistemáticas é essencial para a Medicina Dentária Baseada na Evidência. Contudo, a aplicação de instrumentos como o *Risk Of Bias in Systematic Reviews* (ROBIS)[1] exige tempo, formação metodológica e julgamento especializado, tornando o processo de avaliação crítica moroso e dependente da interpretação humana[2-9]. Os modelos de linguagem de grande escala (LLMs), devido ao seu fácil acesso e à capacidade de processar grandes volumes e diferentes formatos de informação científica, poderão apoiar esta fase da elaboração de sínteses de evidência[10,11]. No entanto, a sua fiabilidade na aplicação do ROBIS[1] e a concordância com investigadores calibrados permanecem insuficientemente esclarecidas.

O objetivo deste estudo foi avaliar a concordância entre investigadores calibrados e um modelo de linguagem de grande escala (LLM) na avaliação crítica de revisões sistemáticas em Medicina Dentária, utilizando o instrumento ROBIS.

MÉTODOS

Estudo *in silico*, transversal

30 revisões sistemáticas, 60 avaliações independentes
Intervenções em Medicina Dentária indexadas na PubMed e selecionadas aleatoriamente

Aplicação do Instrumento ROBIS

Investigadores calibrados (2)
Avaliação independente

Discussão e consenso
Terceiro investigador (em caso de desacordo)

Gold standard
Classificação final obtida pelo consenso

LLM
ChatGPT Pro 5.4

Avaliação Independente com recurso a PROMPT estruturado (via *Prompt Cowboy*)
2 avaliações por revisão em sessões diferentes para avaliar a reprodutibilidade do LLM (Teste-reteste)

Análise Estatística
kappa ponderado (wk); percentagem de concordância; teste de homogeneidade marginal
Nível de significância de α=0,05

RESULTADOS

PERFIL DE RESPOSTAS "YES"
60,9%
Gold standard
40,9%
LLM

TESTE DE HOMOGENEIDADE MARGINAL
12 em 29
% concordância (Pa)
45,4%
p<0,001
p<0,05, diferenças sistemáticas entre as classificações do ChatGPT e do *gold standard*

REPRODUTIBILIDADE ChatGPT Pro 5.4
Kappa ponderado (wk)
Teste-reteste
0,81
IC 95%: 0,78-0,83
p<0,001
Concordância Excelente

CONCORDÂNCIA ChatGPT Pro 5.4 vs. *Gold Standard*
Kappa ponderado (wk)
0,43
IC 95%: 0,39-0,47
Concordância Moderada

CLASSIFICAÇÃO FINAL DO RISCO DE VIÉS
Kappa ponderado (wk)
0,08
IC 95%: −0,08–0,25
Concordância Ligeira

DISCUSSÃO

- O modelo LLM não reproduz de forma consistente o raciocínio dos investigadores ao longo dos diferentes domínios do ROBIS.
- A reprodutibilidade interna do LLM, contrasta com níveis moderados de concordância com o *gold standard*. Estas divergências resultam de diferenças na interpretação dos critérios metodológicos, e não de inconsistência interna dos modelos.
- No entanto, estas divergências na forma como o modelo interpreta e aplica critérios metodológicos sugerem que, na sua configuração atual, não deverão substituir a avaliação humana, constituindo antes uma ferramenta de apoio complementar em tarefas de triagem inicial sob supervisão humana.
- O tempo reduzido de aplicação da ferramenta ROBIS pelo LLM reforça o seu potencial contributo para a eficiência do processo, ainda que condicionado pelas limitações de fiabilidade identificadas.

CONCLUSÃO

Na avaliação metodológica de revisões sistemáticas através do instrumento ROBIS, o LLM demonstrou elevada reprodutibilidade nas classificações atribuídas, porém apenas uma concordância moderada com investigadores calibrados.

REFERÊNCIAS

1. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225-34. 2. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2. 3. Ismail AI, Bader JD, Affairs ADACoS, Division of S, Journal of the American Dental A. Evidence-based dentistry in clinical practice. J Am Dent Assoc. 2004;135(1):78-83. 4. Straus SE, Glasziou P, Richardson WS, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. 5th Edition ed2018. 5. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance G. Evidence based medicine: a movement in crisis? BMJ. 2014;348:g3725. 6. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Collaboration; 2011. 7. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. 8. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6. 9. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163. 11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6).