RWD158

Performance of ChatGPT-4o and Claude 3.5 Sonnet in Title and Abstract Screening for a Systematic Review in Obstetrics

S. Insuk [1], K. Boonpattharatthiti [2,3], C. Booncharoen [1], P. Chaipitak [1], M. Rashid [4], S.K. Veettil [5], N.M. Lai [6], N. Chaiyakunapruk [4,7] and T. Dhippayom [2,4*]

1.Faculty of Pharmaceutical Sciences, Naresuan University, Phitsanulok, Thailand, 2.The Research Unit of Evidence Synthesis (TRUES), Faculty of Pharmaceutical Sciences, Naresuan University, Phitsanulok, Thailand, 3.Faculty of Pharmaceutical Sciences, Burapha University, Chon buri, Thailand, 4.Department of Pharmacotherapy, College of Pharmacy, University of Utah, Salt Lake City, Utah, USA, 5.Department of Pharmacy Practice, School of Pharmacy, IMU University, Kuala Lumpur, Malaysia, 6.School of Medicine, Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Malaysia, 7. IDEAS Centre, Veterans Affairs Salt Lake City Healthcare System, Salt Lake City, Utah, USA

*Corresponding author: Teerapon Dhippayom, PharmD, PhD. E-mail: teerapond@nu.ac.th

INTRODUCTION

- Systematic reviews (SRs) are essential to evidence-based medicine.
- Screening large volumes of titles and abstracts is time-consuming and error-prone.
- Human reviewers can miss up to 10% of relevant studies.
- The growing scale of modern literature makes efficiency increasingly important.
- Generative AI models (e.g., ChatGPT, Claude) are being tested for abstract screening.
- Limited evidence exists on their comparative performance in this task.

OBJECTIVE

This study aims to evaluate the performance of ChatGPT-4o and Claude 3.5 Sonnet against junior researchers in the title and abstract screening stage of a systematic review in obstetrics, using an experienced researcher's decisions as the reference standard.

METHODS

Data Sources: We searched PubMed, EMBASE, Cochrane CENTRAL, and EBSCO Open Dissertations (inception to February 2024) as part of a systematic review on smoking cessation interventions during pregnancy.

Eligibility Criteria: Studies were eligible if they were randomized controlled trials (RCTs) conducted among pregnant women who smoked, investigated pharmacological or electronic cigarette interventions, and reported biochemically confirmed cessation outcomes.

Screening Procedure

1. Reference standard: One experienced reviewer (≥5 years in SRs).
2. Human comparators: Two junior researchers without prior SR experience, each screening half of the records.
3. AI comparators: ChatGPT-4o (OpenAI) and Claude 3.5 Sonnet (Anthropic), accessed via paid web interfaces.
4. A structured, instruction-based prompt (developed iteratively with systematic review experts) guided AI decisions. Records were submitted one at a time for classification as "include" or "exclude," with reasons provided.

Performance Metrics: Accuracy, sensitivity, precision (PPV), F1-score, and negative predictive value (NPV) were calculated against the experienced reviewer's decisions as the reference standard.

Acccess the Title and Abstract Screening Prompt here >>>

<<<<< Access the full-text article here

Insuk, S., Boonpattharatthiti, K., Booncharoen, C. et al. How Well Do ChatGPT and Claude Perform in Study Selection for Systematic Review in Obstetrics. J Med Syst 49, 110 (2025). https://doi.org/10.1007/s10916-025-02246-4

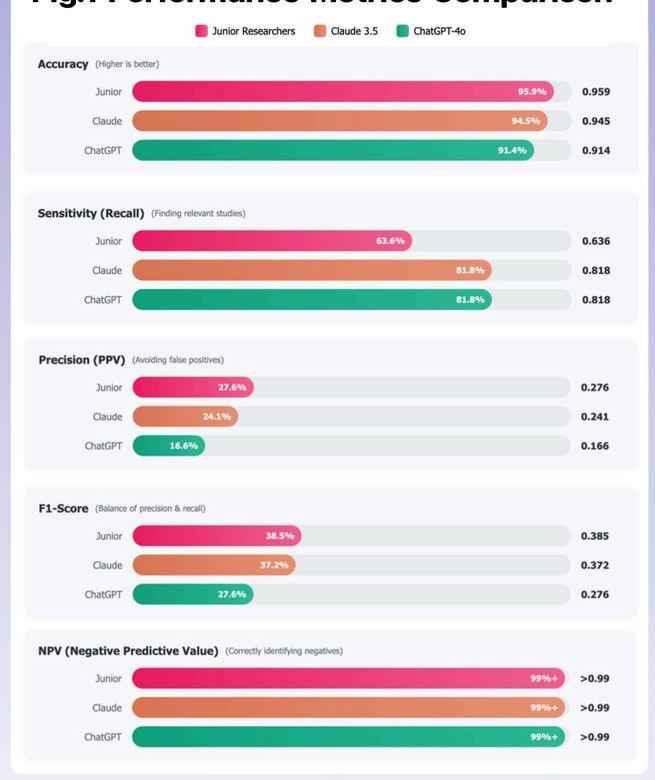
RESULTS

A literature search yielded 1,648 unique titles/abstracts for a systematic review on smoking cessation interventions during pregnancy. Out of 1,648 records: Experienced reviewer: excluded 1,615. Junior researchers: excluded 1,572. ChatGPT: excluded 1,485. and Claude: excluded 1,536.

Performance metrics *(Fig.1)*

- Junior researchers achieved the highest accuracy (0.9593) and F1-score (0.3853).
- Claude demonstrated slightly better performance than ChatGPT, with accuracy of 0.9448 versus 0.9138, and an F1-score of 0.3724 versus 0.2755, respectively.
- Both AI models showed identical recall (0.8182), higher than junior researchers (0.6364).
- All screeners exhibited high NPV: Claude (0.9961), ChatGPT (0.9959), and junior researchers (0.9924).

Fig.1 Performance Metrics Comparison

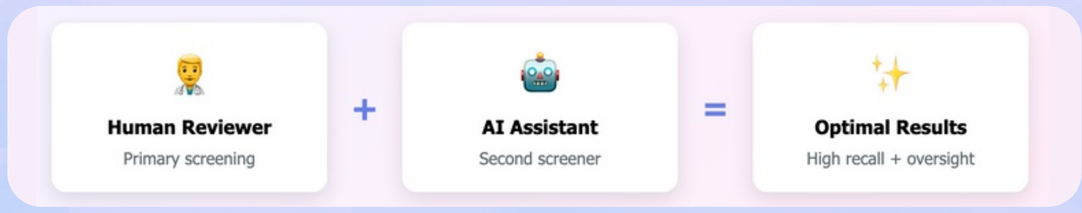


Key Takeaways

- AI models excel at sensitivity: Both ChatGPT and Claude caught 82% of relevant studies (vs 64% for junior reviewers)
- Human reviewers maintain higher precision: Less likely to include irrelevant studies (28% vs 17-24% for AI)
- Near-perfect NPV (>99%): All screeners rarely miss truly relevant studies when excluding
- Hybrid approach recommended: Combine human judgment with AI's high recall for optimal systematic review screening
- Claude slightly outperforms ChatGPT: Better accuracy (94.5% vs 91.4%) and F1-score (0.372 vs 0.276)

A hybrid model—human reviewer plus AI second screener—appears optimal. This ensures high recall while maintaining oversight, consistent with best practices in systematic reviews. *(Fig.2)*

Fig.2 Recommended Hybrid Workflow for Optimal Screening Result



Conclusions

While junior human researchers had the highest overall accuracy, generative AI models, particularly Claude, performed comparably in title/abstract screening for this review, showing high recall and NPV.

This suggests AI holds potential as a supportive tool for this stage, though human oversight remains necessary.