MSR35

Training and Validation Methods of Artificial Intelligence Tools For Automated Literature Reviews in Pharmacoepidemiology and Health Technology Assessment: A Systematic Review

Aurore Bergamasco[1], Richard Chiv[1], Yola Moride[1,2]
1. YolaRx Consultants, Montreal, Quebec, Canada
2. Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

INTRODUCTION

- Systematic reviews (SRs) are essential in pharmacoepidemiology (PE) and health technology assessment (HTA) to support evidence-based decision-making.
- Traditional SR workflows are labor-intensive and time-consuming, especially during screening and data extraction.
- The emergence of commercial AI tools such as ChatGPT (OpenAI) has sparked growing interest in artificial intelligence (AI)-driven automation of SR processes.
- Although these AI tools increasingly support SR tasks, concerns persist about their transparency, validation, and regulatory acceptability.
- While the National Institute for Health and Care Excellence (NICE) has acknowledged AI's potential in evidence generation, practical implementation remains limited.

OBJECTIVE

The objective of this systematic review was to evaluate how available AI tools or platforms used for systematic reviews in pharmacoepidemiology and health technology assessment are trained, validated, and tested.

METHODS

- A systematic search was conducted in MEDLINE and Embase (January 1, 2019 – January 2, 2025) to identify sources on training, validation, and testing of commercially available AI tools for systematic reviews.
- Data extracted included: AI tool name, intended purpose (screening, data extraction, and/or quality assessment), stage of development, evaluation methods, dataset characteristics (type, size, therapeutic area)
- To minimize publication bias, pragmatic searches of supplementary sources were also performed.
- Two independent reviewers conducted abstract screening and data extraction.
- The protocol adhered to the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 checklist and was registered in International prospective register of systematic reviews (PROSPERO) (*CRD pending*).

RESULTS

- We identified 1,540 sources on AI use for literature reviews, of which **94** (6.1%) mentioned training and/or validation.
- The most frequently reported AI tools were **proprietary internal AI models** (66.7%), followed by **ChatGPT** (OpenAI) (29.6%) (Figure 1).
- 80% of reviews using AI were systematic reviews (Figure 2).
- Most AI applications were used for **title and/or abstract screening** (**68.1%**), and **24.1%** for **data extraction** (Figure 3).
- Training methods were described in 21.3% and validation methods in 54.3% of sources.
- **Training dataset size** ranged from **50** to **11,923 abstracts**.
- **Validation dataset size** ranged from **472** to **2,369**.
- Therapeutic areas of the training/validation datasets were largely **unknown** (**42.7%**), while **16.5%** were **oncology**-related (Figure 4).
- Across all AI models, **Sensitivity:** 16.7 – 100%, **Specificity:** 34 – 100%, **Accuracy:** 10 – 100%.

Fig 1. Distribution of AI tools for literature reviews

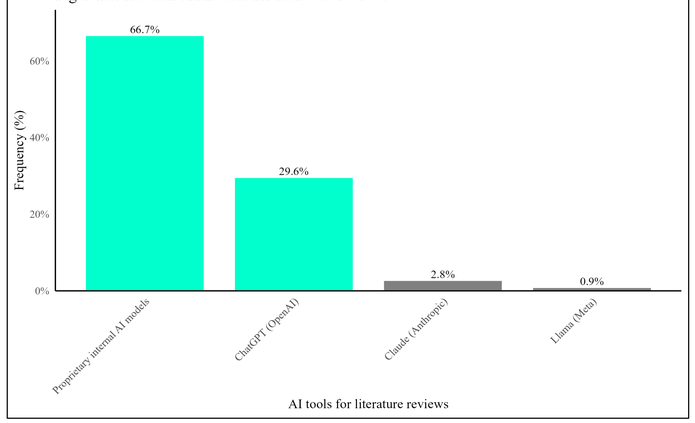


Fig.2 Distribution of types of literature reviews conducted with AI

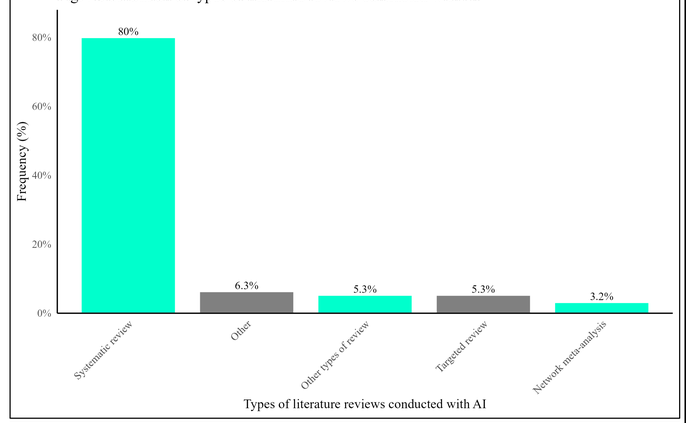
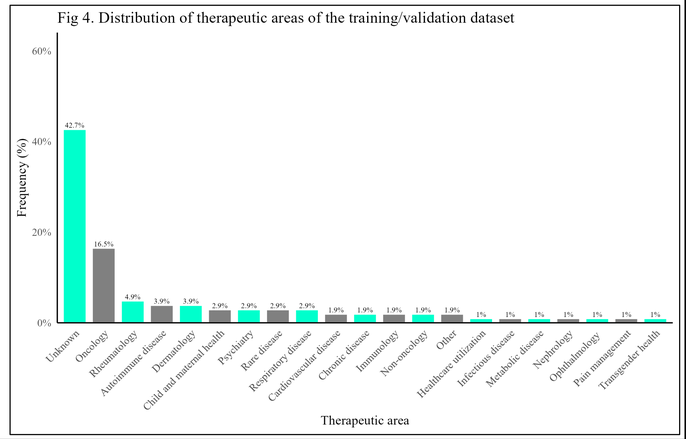


Fig. 3 Distribution of tasks conducted using AI

Task	Count	Percentage (%)
Records deduplication	2	2.1
Titles and/or abstracts screening	64	68.1
In-depth screening	5	5.3
Data extraction	12	12.8
Classification	4	4.3
Flares identification	1	1.1
Study quality assessment	4	4.3

Fig 4. Distribution of therapeutic areas of the training/validation dataset



CONCLUSION

This systematic review highlights the increasing adoption of AI tools for automating systematic review workflows in PE and HTA. While title/abstract screening remains the most common application, the reporting of training, validation, and testing methods varies significantly, raising concerns about consistency and transparency. These findings emphasize the urgent need for standardized evaluation frameworks to ensure the reliability, reproducibility, and credibility of AI tools in systematic reviews.

The study received no funding.

Presented at ISPOR 2025, Montreal, QC, Canada (Wednesday, May 14, 2025)