Large Language Models for Outcomes Research: A Targeted Review

Dolin O,[1] Lim V,[1] Hepworth T,[1] Gonçalves-Bradley DC,[1] Langford B[1]

[1]Symmetron Limited, London, England • Poster inquiries: odolin@symmetron.net • www.symmetron.net • Presented at ISPOR EU 2024 Barcelona Annual Meeting

Introduction

- Interest in using large language models (LLMs) for health outcomes research (e.g. systematic literature review [SLR], real-world evidence [RWE] generation) has increased in recent years.
- A recently published NICE position statement on the use of AI in evidence generation[1] reflects this interest, as HTA bodies anticipate an increase in the adoption of these tools.
- Despite this, guidelines are generally limited, as use of LLMs is not yet common practice. A key challenge is determining which specific tasks LLMs may be suited for in health outcomes research and understanding the level of supervision required to ensure reliable outputs.

Objectives:

- Assess how LLMs – including both open-source and proprietary LLMs designed to represent (e.g. BERT), process and/or generate (e.g. GPT-4) text – are currently used for health outcomes research.
- Identify limitations and concerns regarding LLM use.
- Highlight key areas for future research to ensure responsible and effective use of LLMs.

Methods

- A targeted review was conducted to identify case studies and guidance on LLM usage in health outcomes research (including qualitative and quantitative evidence synthesis and real-world data analysis).
- Embase was searched from November 30, 2022, to May 20, 2024. Supplemental searches of congresses (ISPOR, HTAi, Cochrane Colloquium), Health Technology Assessment (HTA) guidance (NICE, SMC, EUnetHTA, IQWiG, HAS, CADTH, PBAC) and ISPOR good practice guidelines were performed.
- A time limit was imposed on searches to focus on studies published after the widespread popularisation of LLMs (and subsequent interest in HE applications) following the release of ChatGPT in late November 2022.
- Title and abstract (T&A) screening, full-text review, and data extraction were performed by a single reviewer; 20% of records were quality checked by a second reviewer.

Abbreviations: CADTH, Canadian Agency for Drugs and Technologies in Health; EHRs, electronic health records; GPT, generative pre-training transformer; HAS, French National Authority for Health; HO, health outcomes; HTA, health technology assessment; LLMs, large language models; NICE, National Institute for Healthcare Excellence; NMA, network meta-analysis; PBAC, Pharmaceutical Benefits Advisory Committee SMC, Scottish Medicines Consortium; T&A, title and abstract

Results

Tasks examined by case studies

- Included studies were case studies **(**64/69; **Figure 1),** reviews (2/69), editorials (2/69), and an ISPOR good practice report (1/69).
- Research tasks examined by case studies are listed in **Figure 2**; data extraction and T&A screening were most frequently examined.

Figure 1. Screening process

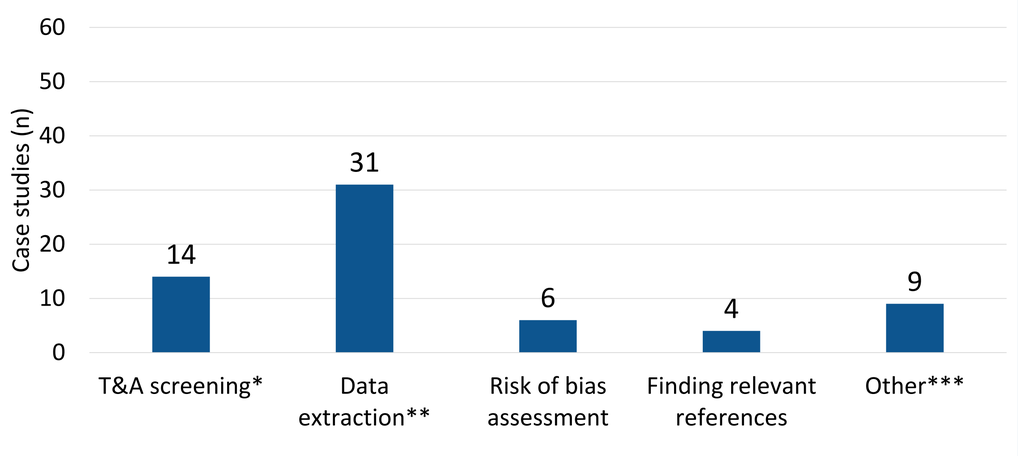
Title and abstract screening
325 records were identified for primary screening

Full text screening
172 records met the eligibility criteria and full texts were screened

Data extraction
69 studies (34 journal articles, 27 conference abstracts, 3 letters to the editor, 2 editorials, 2 reviews, and 1 ISPOR good practices report) met the eligibility criteria and were extracted*

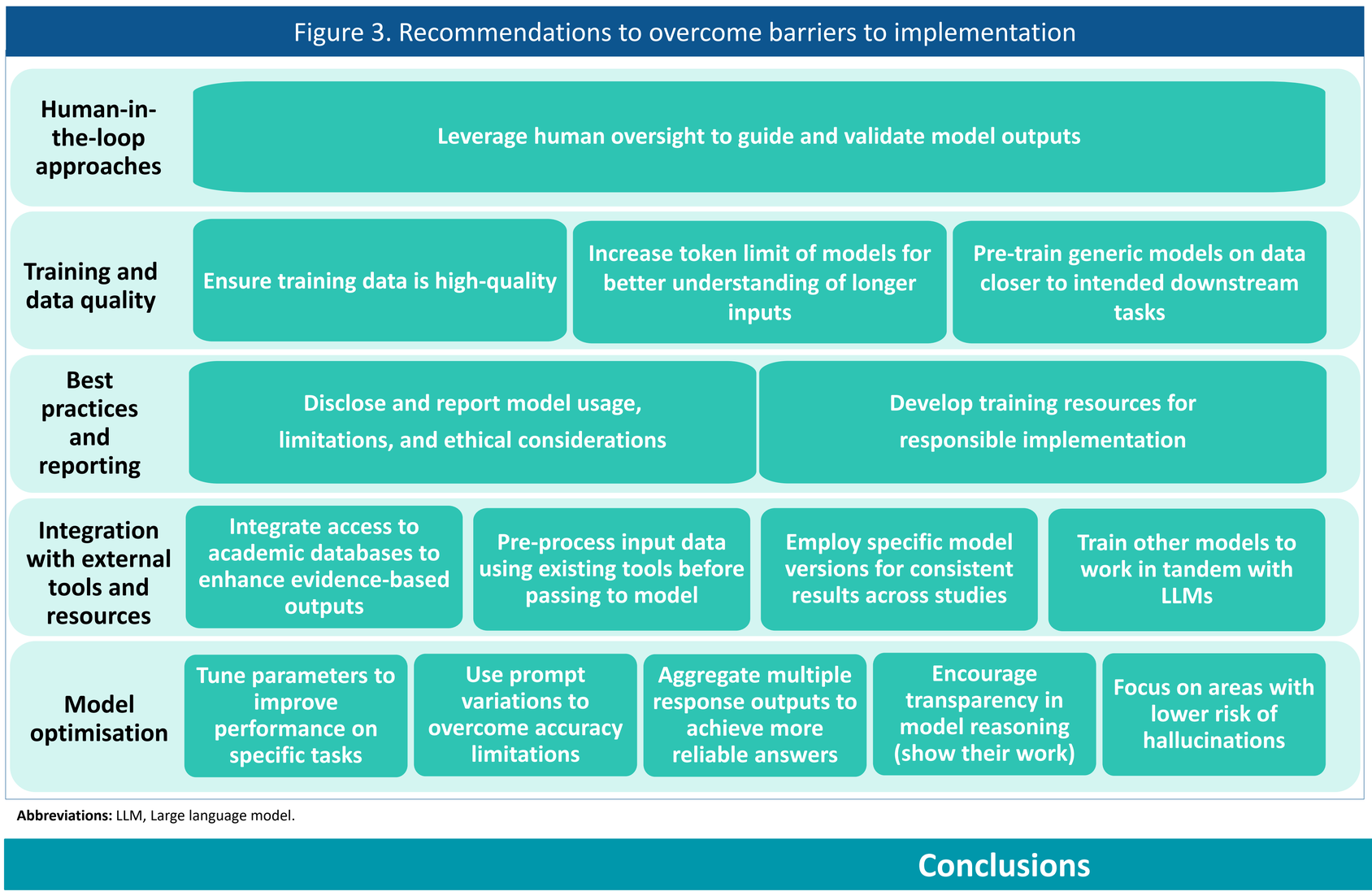
*Four studies were behind a paywall. For these, only abstracts were extracted. All other studies were freely available.

Figure 2. Tasks examined by case studies (n = 64)



Abbreviations: T&A, Title and abstract.
*No case studies assessed full-text screening. **Data extraction of electronic health records or clinical trial reports. ***This included tasks such as generation of search strategies, drafting protocols, and generating NMA code.

Figure 3. Recommendations to overcome barriers to implementation



Abbreviations: LLM, Large language model.

Key characteristics of identified research

- Most case studies assessed LLMs' ability to replicate pre-existing findings (e.g. to correctly screen/extract studies already screened/extracted by human reviewers). Only six case studies performed new research using LLMs (where any LLM validation was a secondary focus); no studies mentioned informing HTA submissions.
- Most case studies (55/64) quantitatively assessed how well LLMs performed a task of interest. Fewer case studies (20/64) examined how long it took to complete a task of interest using LLMs (e.g. compared to traditional workflows). Only 3/64 case studies examined the cost of using LLMs to complete a task of interest.
- Most case studies (43/64) did not make any code or data publicly available.
- GPT-based models (40/64) were the most used across the reviewed studies, followed by BERT (20/64).

Recommendations for future research

- 47/69 studies recommended future research:
  - 24 suggested improving prompt designs
  - 21 suggested examining the performance of other LLMs
  - 10 suggested testing new LLM training datasets
  - 10 suggested examining the generalisability of findings across domains/contexts

Recommendations for implementation

- 26/69 studies discussed whether LLMs should be immediately implemented in health outcomes research:
  - 2 suggested avoiding immediate implementation
  - 8 suggested implementation without supervision
  - 16 suggested implementation with supervision

Barriers to implementation

- 29/69 studies identified barriers to implementing LLMs for health outcomes research.
- The following barriers were identified:
  - inaccurate responses (23/29)
  - issues with training data (6/29)
  - limited input length/type (5/29)
  - concerns surrounding cost (2/29)
  - Other barriers (4/29) included the time demands required to check model responses, ethical concerns (e.g., bias, potential for malicious use), and the requirement for human accountability (e.g. from regulatory bodies).
- 24 studies outlined measures to overcome barriers **(Figure 3).**

Conclusions

Conclusions and implications

- Included studies focussed on validating LLMs by replicating existing findings, instead of generating novel health outcomes research. The feasibility of LLM use is often illustrated through limited examples that lack comprehensive demonstrations of reliability, time savings, or cost-effectiveness. Such demonstrations are crucial to broader adoption of LLMs in health outcomes research.
- The predominant barrier to implementation identified by studies was response inaccuracy when performing tasks.
- Many of the suggestions for overcoming barriers to implementation, including those surrounding model optimisation, integration with external tools and resources, and training and data quality, seek to improve the capabilities of LLMs. Over time, this should improve LLM response accuracy and performance.
- However, even as the overall performance of LLMs improves, there still may be an unacceptable level of variability or uncertainty in performance for tasks.
- Human-in-the-loop approaches maintain the quality of research outputs while taking advantage of efficiency gains from LLMs. Most included studies recommended this approach. Despite this, there is little guidance on how to implement human-in-the-loop methods.

Limitations

- Review searches were performed in May 2024. Research in this area is rapidly evolving.
- HTA submissions and conferences focussing on electronic health records or real-world evidence were not reviewed for reports on the use of LLMs.
- Implementation and testing of LLMs within businesses looking to maintain a competitive edge (e.g. pharmaceutical companies, contract research organisations) may not be published.
- The usage of LLMs in the literature is not necessarily reflective of usage in practice.

Key messages:

- Research into LLM use for health outcomes research has focussed primarily on assessing the feasibility of various use cases.
- Methods to overcome limited accuracy or reliability of LLMs have not been examined. Future research into manually validating automated actions could mitigate variations in LLM performance and improve efficiency whilst maintaining research quality.

References: [1]NICE. Use of AI in evidence generation: NICE position statement, 2024.
Declaration of funding: *This project has been funded in full by Symmetron Limited.*