Large Language Models for Health Economics and Health Technology Assessment: A Targeted Review

Dolin O,[1] Lim V,[1] Chalmers K,[1] Hepworth T,[1] Gonçalves-Bradley DC,[1] Langford B,[1] Rinciog C[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 economics (HE) and health technology assessment (HTA) has increased in recent years.
- Recently published NICE guidance on the use of artificial intelligence (AI) in evidence generation[1] reflects this interest as HTA bodies anticipate increased adoption of these tools.
- Despite this, guidelines are generally limited, as use of LLMs is not yet common practice. Currently, the feasibility of integrating LLMs into existing workflows, as well as the extent of their use, is not well understood.

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 applied in HE modelling and HTA submissions.
- 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 the use of LLMs for HE modelling and to understand the extent to which such methods are used in HTA.
- Embase was searched from November 30, 2022, to May 20, 2024, and supplemented by searches of congresses (ISPOR, HTAi, Cochrane Colloquium), HTA guidance (NICE, SMC, EUnetHTA, IQWiG, HAS, CDA-AMC, PBAC) and ISPOR good practice guidelines.
- 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 screening, full-text review, and data extraction were performed by a single reviewer; 20% of records were quality checked by a second reviewer.

Abbreviations: AI, artificial intelligence; CDA-AMC, Canadian Agency for Drugs and Technologies in Health; GPT, generative pre-trained transformer; HAS, French National Authority for Health; HE, health economics; HTA, health technology assessment; HTAi, health technology assessment international; IQWiG, Institute for Quality and Efficiency in Healthcare; LLMs, large language models; MS, Microsoft; NICE, National Institute for Healthcare Excellence; PBAC, Pharmaceutical Benefits Advisory Committee; SMC, Scottish Medicines Consortium; VBA, Visual Basic.

Results

Figure 1. Screening process

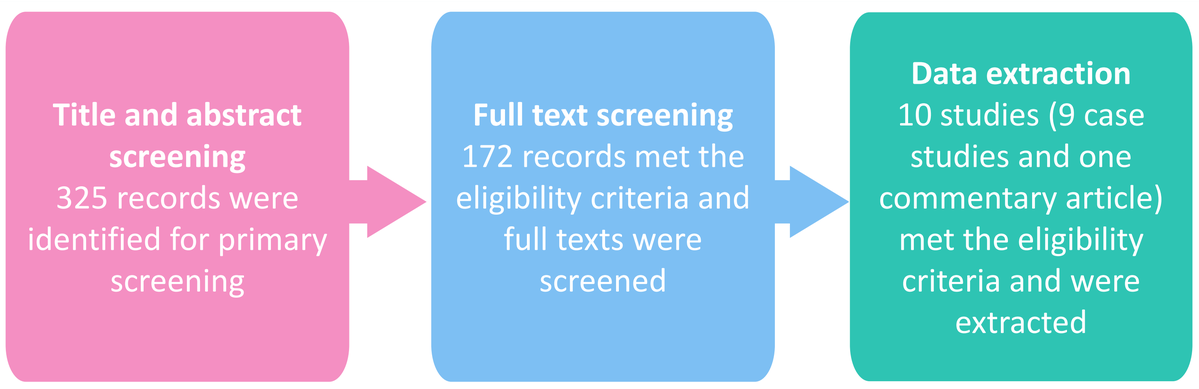
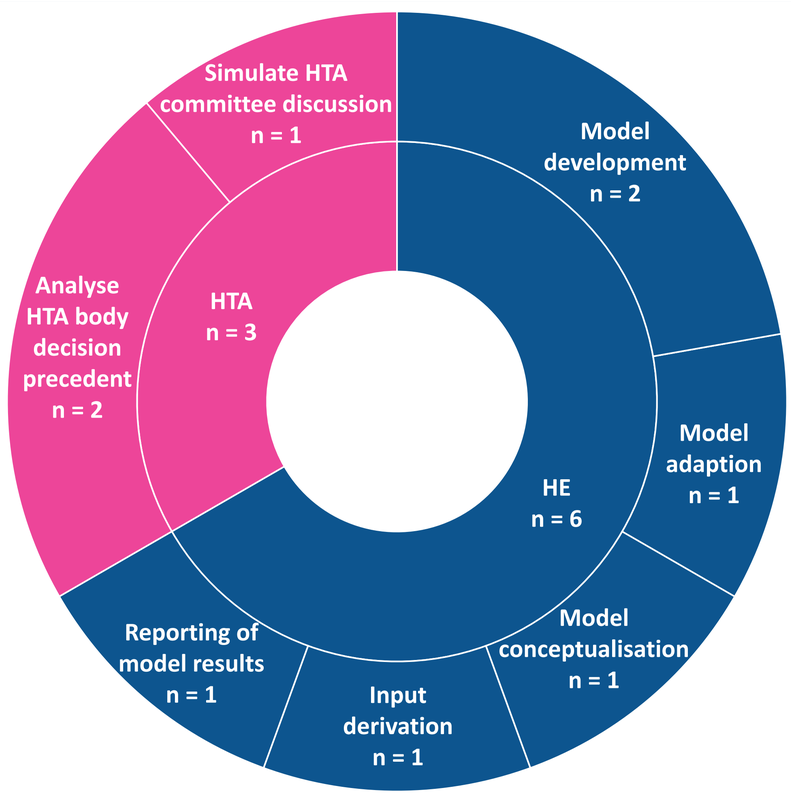


Figure 2. Tasks examined by case studies



Abbreviations: HE, health economics; HTA, health technology assessment.

HE and HTA tasks examined

- Case studies (9/10, see **Figure 1**) primarily focussed on demonstrating the feasibility of using LLMs for different research tasks (see **Figure 2**; **Table 1**).

Methodology and results of included studies

- Most case studies reported minimal methods and results. The code and data used to generate results were only made available by one study.
- The performance, time/workload savings, and cost-effectiveness of LLM use were rarely evaluated quantitatively (see **Table 1**) by case studies. When assessed (3/9), accuracy was high, but testing was limited (e.g. focussed on one or two economic models).
- At the time of the search, no HTA guidance for LLM usage in economic modelling was identified.

Table 1. Overview of included case studies

Task	Code or data available?	Quantitative assessment of . . . performance?	time savings?	cost?
Adapt a global technical report for cost utility model to a country-specific setting (HE)[2]	X	✓	X	X
Adapt an HTA-ready Excel model from the setting of one country to another (HE)[3]	X	✓	✓	✓
Rapidly prototype a decision analytical model (HE)[4]	X	X	X	X
Program a cost-effectiveness model (MS Excel/VBA) (HE)[5]	X	X	X	X
Program partitioned survival models in R (HE)[6.i]	✓	✓	✓	X
Develop a conceptual cost-effectiveness analysis model (HE)[7]	X	X	X	X
Find relevant regulatory precedent from HTA database (HTA)[8.ii]	X	X	X	X
Replicate an HTA committee discussion (HTA)[9]	X	X	X	X
Classify terms associated with decision outcomes as positive, neutral or negative (HTA)[10]	X	X	X	X

[i] *The only journal article; all other case studies were conference abstracts.* [ii] *The only case study that did not use a GPT model.*

Abbreviations: HE, health economics; HTA, health technology assessment.

Barriers and recommendations for use

- Barriers to implementation identified by studies included response accuracy (5/10), data security (3/10), HTA body acceptance (1/10) and the ease with which code produced by an LLM can be understood (1/10).
- Five studies made recommendations about immediate use. Four studies recommended implementation with supervision (e.g. human validation of LLM outputs). One study, focused on VBA model code development, recommended using LLMs to help inexperienced programmers but avoiding implementation for senior developers.

Recommendations for future research

- Future research to advance the use of LLMs for HTA decision making and HE were suggested by six studies. They recommended the following:
  - Focus on expanding LLMs' roles in automating different sections of technical reports, in addition to automating reference generation and updating tables directly from country-specific Excel© data.
  - Enhance LLM simulations of HTA committee meetings by incorporating conditional responses from committee members based on prior discussions and identifying complex relationships between terms (e.g. positive, neutral, or negative).
  - Examine issues related to code explainability, ownership, and licensing, to ensure that the integration of LLMs in routine health economic analysis and decision-making is ethical and transparent.
  - Enhance LLM accuracy through feedback loops, prompt optimisation, and testing across various models. Explore how LLM prompts can be generalised and adapted across different decision problems and model types.
  - Evaluate LLMs for their ability to handle diverse model types, including decision trees, Markov models, and individual patient simulations, across a broader range of disease areas and scenarios.

Conclusions

Summary and implications

- The use of LLMs for HE/HTA is an emerging topic. Published evidence is minimal, difficult to replicate (code and data were not made available) and primarily exploratory (with no published evidence of LLMs being used for de novo modelling). Several potential use cases remain unexamined, including deterministic sensitivity analyses, model validation, or model adaptation (beyond updating model inputs).
- Response accuracy was the most commonly identified barrier to LLM implementation. Research recommendations made by studies to overcome this issue focussed on improving accuracy by using models more effectively (e.g. via prompt optimisation).
- In addition to this approach, research examining efficient implementation of human-in-the-loop approaches should be explored by future work. These approaches could be especially useful when LLM response accuracy is uncertain, or too variable to be left unsupervised. They would also help researchers comply with current NICE guidance,[1] which recommends a human-in-the-loop approach to maintain quality and trust in findings.

Limitations

- Review searches were performed in May 2024. Research in this area is rapidly evolving.
- 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:

- There is limited published evidence examining the use of LLMs for HE workflows.
- More high-quality research is needed to demonstrate the effectiveness of these tools, to help researchers feel more confident in research produced using LLMs.
- Future research should explore the feasibility of integrating LLMs into common HE tasks and develop effective workflows for validating model responses.

Declaration of funding: *This project has been funded in full by Symmetron Limited.*

References: [1]NICE. Use of AI in evidence generation: NICE position statement, 2024. [2]Rawlinson W TS, Reason T, Malcolm B, Gimblett A, Klijn S. Automating Economic Modeling: Potential of Generative AI for Updating Modeling Reports. 2024-05, ISPOR 2024. Volume 27. Atlanta, GA, USA: Value in Health, 2024. [3]Rawlinson W KS, Teitsson S, Malcolm B, Gimblett A, Reason T. Automating Economic Modelling: Potential of Generative AI for Updating Excel-Based Cost Effectiveness Models. 2024-05, ISPOR 2024. Volume 27. Atlanta, GA, USA: Value in Health, 2024. [4]Meissner F, Garza C. Rapid construction of an explanatory decision analytical model of treating severe depression during pregnancy with SSRI psychopharmacological therapy by the use of ChatGPT. Journal of Psychosomatic Research 2023;Conference: 10th annual scientific conference of the European Association of Psychosomatic Medicine (EAPM). Wroclaw Poland. 169(no pagination). [5]Poirrier JE, Bergemann R. MSR26 The Use of Copilot, a Generative Artificial Intelligence Tool, as VBA Programming Assistant in Excel-Based Health Economic Models. Value in Health 2023;26(12 Supplement):S398. [6]Reason T, Rawlinson W, Langham J, et al. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. PharmacoEconomics - Open 2024;8(2):191-203. [7]Chhatwal J YI, Balta D, Ermis T, Tenkin S, Samur S, Ayer T. Can Large Language Models Generate Conceptual Health Economic Models? 2024-05, ISPOR 2024. Volume 27. Atlanta, GA, USA: Value in Health, 2024. [8]Signorovitch J, Llop C, Song Y, et al. MSR155 Identifying Relevant Clinical Regulatory and Health Technology Assessment (HTA) Precedents via Artificial Intelligence (AI). Value in Health 2023;26(12 Supplement):S423. [9]Reason T, Rawlinson W, Langham J, et al. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. PharmacoEconomics - Open 2024;8(2):191-203. [10]Aballea S, Maszke S, Shchepetnov A, et al. HTA347 Exploration of the Factors Driving Decisions of the Transparency Commission Based on an Analysis of HAS Meeting Transcripts Using Natural Language Processing. Value in Health 2023;26(12 Supplement):S387.