EE457

Exploring LLMs in the Conceptual and Functional Construction of Health Economic Models: A Case Study on Alzheimer's Diagnostic Cost-Effectiveness Model

Emilija Veljanoska[1], Agota Szende[2]
[1]Fortrea, Market Access Consulting & HEOR, Munich, Germany
[2]Fortrea, Market Access Consulting & HEOR, Leeds, UK

Introduction

Large language models (LLMs), such as ChatGPT-4, are transforming the way knowledge is processed and applied.

Their use in structural definition, functional logic, and parameter assignment within health economic (HE) modeling remains largely untested.

If applied in these domains, LLMs may support standardized and reproducible HE model construction; researching their application is important due to their implications for model automation and policy-relevant outputs.

Objective

This study investigates how an LLM can contribute to the conceptual and functional development of a cost-effectiveness model in Excel. The LLM was tested on a diagnostic use case in Alzheimer's disease (AD).

Methods

This study followed a two-phase approach:

1) Targeted literature review

A targeted literature search (Jan 2019 – April 2025, Embase) was conducted to identify previous and/or existing applications of LLMs in HE modeling; the search was disease- and LLM-agnostic.

2) Structured case study

ChatGPT-4 was prompted to:

- propose a disease progression model structure for AD
- generate an Excel layout to represent the model structure
- retrieve parameter values, including incidence, costs, utilities, and diagnostic performance metrics

All outputs were compared against a published benchmark model (PMID: 40054769)[1] to assess structural and parameter alignment; this model simulates the cost-effectiveness of ^{18}F-flutemetamol positron emission tomography (PET) vs. cerebrospinal fluid (CSF) testing in suspected AD.

Results

1) Targeted literature review

The targeted literature search identified five studies applying LLMs in HE modeling.

Among these, one study[2] was selected for comparison; it implemented a model using a code-based platform and reported metrics on accuracy, error rates, and level of human input required for code refinement and debugging.

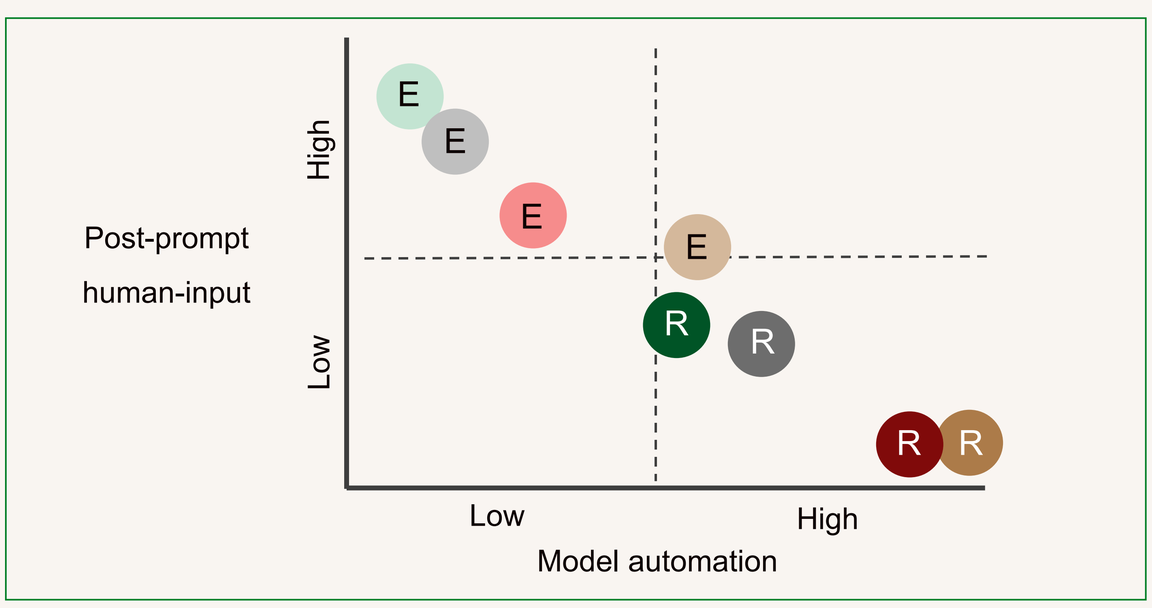
This study demonstrated that LLMs could assist with patient survival data extraction and R-scripted model replication; however, it still required human oversight for complex logic[2].

2) Structured case study:

Excel functionality

The LLM produced a static model layout, including sheet headers and structural placeholders, in Excel.

However, the functional implementation was limited compared to R, as the LLM did not implement formulas, lookup logic, named ranges, or computational functionality (**Figure 1**).



Model shell (brown); formula implementation (grey); model traces (green); output generation (red)

Figure 1. LLM performance in HE modeling: Excel (E) vs. R

Model structure

ChatGPT-4, guided by structured prompts and model specifications, produced an initial disease progression model consistent with the benchmark, including mild cognitive impairment (MCI) with and without AD pathology, dementia, and death (**Figure 2**).

An expanded structure, introducing MCI with unknown etiology as a decision node, was considered based on centiloid thresholds; however, this refinement was not retained by ChatGPT-4 due to limited biomarker data and structural simplification.

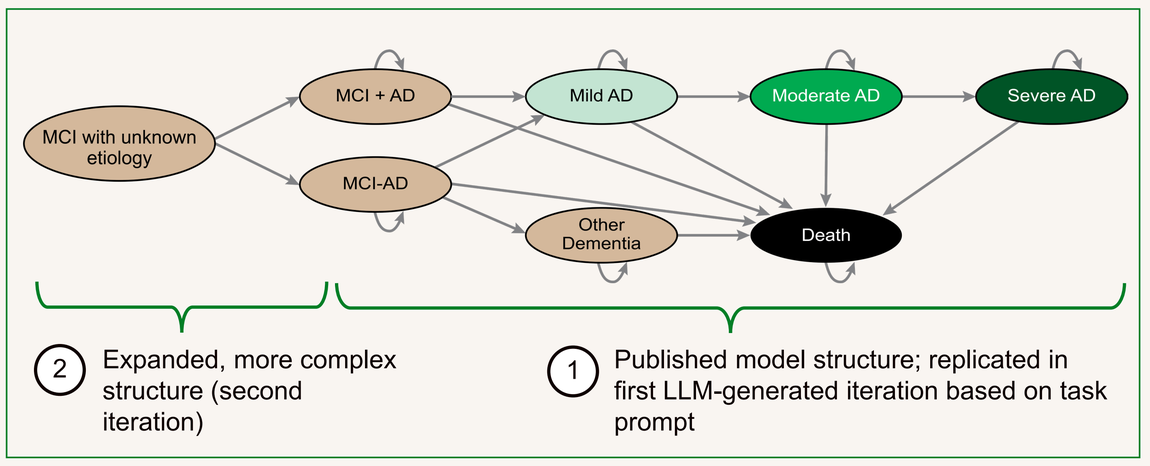


Figure 2. Disease progression model structure: LLM vs. published model

Parameter accuracy

LLM-generated estimates for costs/utilities substantially deviated from the benchmark, whilst diagnostic accuracy remained closer; however, outputs varied across iterations, with inconsistent parameter shifts relative to the published values (**Table 1**).

Table 1. Parameter deviations from benchmark: LLM vs. published model

	Published value	LLM-generated value	Δ
MCI incidence (per 1,000 person-years)	41	21	Underestimated by ~50%
Health state costs	MCI AD: $1,308 Mild AD: $1,534 Moderate/severe AD: $2,132	MCI AD: $2,355 Mild AD: $2,401 Moderate/severe AD: $4,051	Overestimated by up to 2×
Utilities	MCI AD: 0.73 Mild AD: 0.66 Moderate AD: 0.47 Severe AD: 0.32	MCI AD: 0.80 Mild AD: 0.72 Moderate AD: 0.50 Severe AD: 0.35	Within 1.1× of benchmark
Sensitivity/specificity	Sensitivity PET: 91% Specificity PET: 90% Sensitivity CSF: 76% Specificity CSF: 77%	Sensitivity PET: 92% Specificity PET: 90% Sensitivity CSF: 83% Specificity CSF: 81%	Deviation ranged from 1–9%

Conclusions

LLMs like ChatGPT-4 demonstrate early promise in assisting with the conceptual development of HE models and generating preliminary parameter estimates.

LLMs have limited ability to implement HE models in spreadsheet-based tools like Excel, whereas index- and script-based structures used in code-based platforms like R align more closely with LLM architectures.

ChatGPT-4 was tested as an external tool without built-in Excel integration; AI systems with embedded spreadsheet functionality may improve functional implementation in HE modeling and merit further evaluation.

Until LLMs can reliably generate functional logic and computation; human expertise remains vital for robust and credible HE modeling.

References

1. Veljanoska E, Gomes S, Szende A, Cabra A, Munter-Young R, Gordois A. The Value of Positron Emission Tomography for Confirmation of Alzheimer's Disease in the Era of Amyloid-Targeting Therapies. Value Health. 2025;28(6):829-838. doi:10.1016/j.jval.2025.02.011

2. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. Pharmacoecon Open. 2024;8(2):191-203. doi:10.1007/s41669-024-00477-8

Fortrea