Generative Artificial Intelligence for Automated Health Economic Model Development: A Proof-of-Concept Framework

EE197

Anubhav Patel, John R. Cook | Peritia, Morrisville, NC, USA

ISPOR US 2026 | Philadelphia, PA, USA

Keywords: Artificial Intelligence · Cost-Effectiveness · Health Economic Modeling · Automation · Large Language Models

BACKGROUND

Health economic (HE) models require extensive manual programming, deep domain expertise, and significant development time — often weeks to months per model.

Large language models (LLMs) offer an opportunity to automate and standardize model development, reducing build time while maintaining accuracy.

This proof-of-concept evaluates whether generative artificial intelligence (AI) can auto-generate cost-effectiveness (CE) models directly from published journal articles—across Markov, partitioned survival model (PSM), and individual patient data (IPD) microsimulation structures, without manual programming by the modeler.

To our knowledge, no prior AI-assisted HE modelling framework has been systematically validated against published results across these model types.

OBJECTIVES

1. Evaluate feasibility of generative AI to auto-generate and execute CE models directly from published journal articles, without manual programming by modeler.
2. Reduce model development time while maintaining accuracy relative to published results.
3. Improve reproducibility through AI-generated parameter extraction and model templates.
4. Demonstrate applicability across five CE models spanning three model structures: three-state Markov, PSM, and IPD microsimulation.

METHODS

Study Design

A proof-of-concept study in which an LLM was provided a published paper as input. AI read the paper, extracted all parameters, and auto-generated models in both Excel and R. This process was repeated across five publications spanning three model structures: three-state Markov, partitioned survival, and individual patient data microsimulation. Claude (Anthropic, claude.ai) was the LLM used.

No manual programming by the modeler was required, except for the IPD microsimulation model. Although the AI generated the Visual Basic for Applications (VBA) macro code, the user was required to insert the code into the Excel workbook, representing a single manual step necessitated by Excel's macro security architecture.

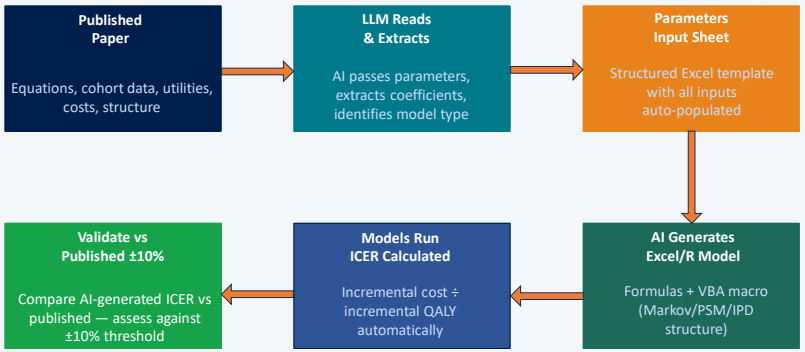
Ref #	Author (Year)	Model Structure	Disease (Country)
1	Takenaka (2021)	Markov 1	Atopic Dermatitis (Japan)
2	De Benedetto (2023)	Markov 2	KPC-Kp BSI (Italy)
3	Sikirica (2012)	Markov 3	ADHD (USA)
4	Aguiar-Ibáñez (2022)	PSM	MSI-H/dMMR CRC (USA)
5	Hoerger (2023)	IPD	T2D (USA)

ADHD: attention-deficit/hyperactivity disorder, KPC-Kp: Klebsiella pneumoniae carbapenemase–producing Klebsiella pneumoniae, MSI-H/dMMR CRC: microsatellite instability–high/deficient mismatch repair colorectal cancer, T2D: type 2 diabetes, USA: United States of America

Validation Approach

AI-generated ICERs were compared against published peer-reviewed values. For the three Markov models, independent human replication was also performed by a modeler, and build time was recorded for both AI and human approaches to enable direct comparison. A validation threshold of ±10% deviation from the published ICER was pre-specified a priori as a pragmatic criterion for replication success, consistent with established principles of model validation and external consistency in health economic modelling (Eddy et al., 2012). Build time was measured from paper input to functional model output, including sensitivity analyses.

AI MODEL GENERATION — How the AI Builds CE Models



RESULTS — MODEL VALIDATION (5 Models)

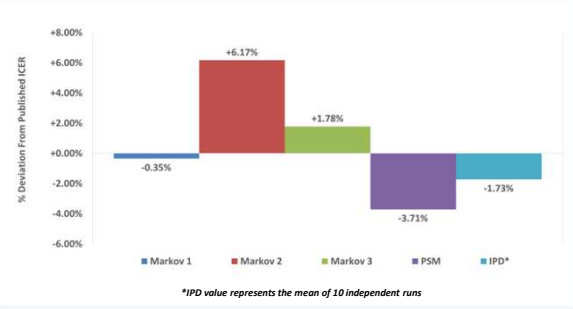
Model	Published ICER	AI Replication*		Human Replication		Replication Time Savings (%)**
		ICER	Deviation (%)	ICER	Deviation (%)	
Markov 1 (¥/QALY)	3,923,633	3,909,855	− 0.35%	3,608,670	− 8.03%	30%
Markov 2 (€/QALY)	32,317	34,311	+ 6.17%	34,311	+ 6.17%	40%
Markov 3 ($/QALY)	31,660	32,225	+ 1.79%	30,759	− 2.85%	36%
PSM ($/QALY)	6,984	6,725	−3.71%	NA	NA	NA
IPD† ($/QALY)	9,103	Min: 7,221 Max: 10,086	−20.67% 10.80%	NA	NA	NA

** Results were identical for both the Excel and R model replications*

*** For the three Markov models, mean build time was 2-2.5 weeks for AI versus 3-4 weeks for human replication, representing a 30–40% reduction. Build time savings was not captured for the PSM and IPD microsimulation models.*

† ***IPD Microsimulation Accuracy:*** *Given the stochastic nature of IPD microsimulation, the AI-replicated model was run ten times using different random number seeds (N=10,000 per run). Eight of ten runs produced ICERs within ±10% of the published value of $9,103/QALY. Run-to-run variability in IPD microsimulation is an expected function of sample size and does not reflect the instability unique to the AI-replicated model, but rather to the stochastic nature of individual patient simulation.*

ICER VALIDATION — Published vs AI-Replicated



**IPD value represents the mean of 10 independent runs*

DISCUSSION

- **Findings:** AI accurately reproduced ICERs across 5 CE models spanning 4 disease areas, 3 countries, and 3 model structures — including IPD microsimulation, the most complex HE model type.
- **Strengths:** Models spanned the full complexity spectrum — Markov, PSM, and IPD microsimulation — strengthening generalizability. Independent human replication provided a direct head-to-head comparator for Markov models. Framework used a standard commercial LLM with no custom tools, ensuring reproducibility by any HEOR team.
- **Current limitation:** Build time comparison was conducted for 3-state Markov models only — human replication time was comparable for these. Timing data for the PSM and IPD microsimulation were not captured; the ~30–40% build time reduction should not be generalized across all model structures.
- **Future:** AI-assisted replication could serve as a rapid starting point for early-stage model development, where inputs are subsequently adjusted to reflect the specific characteristics of a product in development. It could also support competitive intelligence by enabling efficient exploration of alternative settings, assumptions, and payer perspectives across multiple scenarios.

CONCLUSION

- Build time was reduced ~30–40% with the Markov models while maintaining accuracy across diverse therapeutic areas
- Expert oversight remains essential — while AI reduces mechanical burden, robust model development still requires technical implementation and validation alongside clinical judgement
- This proof-of-concept establishes a foundation for broader AI integration in HEOR practice
- This proof-of-concept demonstrates that AI can reliably support HE model replication; application to de novo model development and alternative model structures requires further evaluation

REFERENCE

1. Takenaka et al. J Cutan Immunol Allergy 2021; 4: 100–108 (DOI: 10.2165/11634240)
2. De Benedetto et al. Microorganisms 2023; 11: 1102 (PMC10222869)
3. Sikirica et al. Pharmacoeconomics 2012; 30(8): e1–e15 (DOI: 10.2165/11634240)
4. Aguiar-Ibáñez et al. J Med Econ 2022; 25(1): 469–480 (PMC9012785)
5. Hoerger TJ et al. Value Health 2023; 26(9): 1372–1380 (PMC11017333)