Evaluating Generative AI in Replicating Health Economic Models: A Case Study on Ulcerative Colitis

MSR30

Sumeyye Samur[1], Jakob Langer[2], Emir Gursel[1], I. Fatih Yildirim[1], Turgay Ayer[1,3], Jagpreet Chhatwal[1,4], Ipek Ozer Stillman[5]

1 Value Analytics Labs, Boston, MA, USA; 2 Takeda Pharmaceuticals International AG, Zurich, Switzerland; 3 Georgia Institute of Technology, Atlanta, GA, USA; 4 Massachusetts General Hospital Institute for Technology Assessment, Harvard Medical School, Boston, MA, USA; 5 Takeda Pharmaceuticals U.S.A., Inc., Boston, MA, USA

Value Analytics Labs

KEY FINDINGS

- This study highlights the potential and current limitations of Generative AI to replicate a complex HTA-ready cost-effectiveness model based on a publication and a technical report—a model that captures the multifaceted nature of ulcerative colitis, including treatment sequencing.
- Generative AI shows strong potential to replicate health economic models when supported by detailed and standardized documentation, with its performance closely tied to the transparency and clarity of input definitions and model assumptions
- Improving standardized reporting practices within the HEOR field is needed to enable Generative AI to better support stakeholders during HTA processes.

BACKGROUND

- Health economic modeling plays a critical role in evidence generation for healthcare decision-making.
- Generative AI has the potential to accelerate model development, reduce manual effort, and increase reproducibility.
- However, the feasibility of using AI to replicate published health economic models remains underexplored.

OBJECTIVE

- To evaluate the feasibility and accuracy of Generative AI in replicating a published health economic model for ulcerative colitis as described in text, i.e., without access to the actual model.
- To assess AI performance across different levels of source detail (publication vs. a detailed technical report).

METHODS

Model Selection

Target: Markov model with a treatment sequence for ulcerative colitis published by Salcedo et al.[1]

Replication Experiments

Experiment 1: Publication-Based Replication

- Used ValueGen.AI[2], a GPT-4-based platform integrating LangChain[3], CrewAI[4], and OpenAI[5] libraries.
- Extracted model structures, health states, transition probabilities, costs, and utilities.
- Reconstructed the model using R's heemod package.

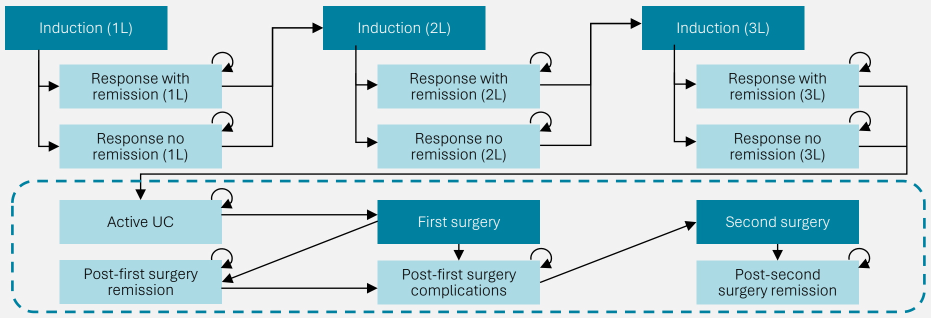
Experiment 2: Technical Report-Based Replication

- Applied the same extraction pipeline to a more detailed technical report version of the model.
- Compared AI outputs across both iterations.

Evaluation Criteria

- Accuracy in identifying:
  - Health states
  - Transition probabilities
  - Costs
  - Utilities
  - Treatment lines

Figure 1: *Schematic of the original model structure as published by Salcedo et al.* [1]


Induction (1L)
Response with remission (1L)
Response no remission (1L)
Induction (2L)
Response with remission (2L)
Response no remission (2L)
Induction (3L)
Response with remission (3L)
Response no remission (3L)
Active UC
Post-first surgery remission
First surgery
Post-first surgery complications
Second surgery
Post-second surgery remission

RESULTS

We evaluated the performance of Generative AI in replicating key health economic model components from the publication and the technical report across five core modeling domains: health states, transition probabilities, costs, utilities, and treatment pathways.

1. Health States

Key finding: Generative AI performed better in identifying health states when parsing the technical report, especially for capturing surgery-related and death states, but continued to hallucinate or misclassify some non-existent states due to ambiguous phrasing and insufficient descriptions in both the publication and technical report document. (Table 1)

Table 1: *Comparison of Large Language Model (LLM) performance for the publication vs the technical report on health states*

Health State	Pub.	Technical Report	Explanation
Active Ulcerative Colitis	✓	‼ Linked to an existing state	Clearly labeled in publication, implicit description in the technical report
Induction phase	✓ (misaligned w/ treatment lines)	✓ (present, therapy linkage unclear)	Descriptions were vague or embedded in narrative, not distinctly linked to treatment phases
Response w/ remission	✓	✓	Clearly labeled in text, allowing LLMs to extract
Response w/o remission	✓	✓	Clearly labeled in text, allowing LLMs to extract
Maint. phase	X Falsely introduced	‼ Linked to an existing state	Term "maintenance" mentioned narratively, misinterpreted as a separate health state or linked to existing health states
Discon-tinuation	X Falsely introduced	✓ N/A	Word "discontinue" used in patient pathways but not as a state; LLM misclassified
Death	X Not captured	✓ Captured but not used in transitions	Not visually modeled in diagrams; implied in text, so harder for LLM to recognize
Surgery-related states	‼ Partial — only the first surgery state is captured	✓ Captured	Report had explicit state labels; publication included them only narratively
Spontaneous remission / no remission states	X Not in source; not captured	✓ Captured	Clearly labeled in the report, allowing LLMs to extract

2. Transition Probabilities

Key finding: Transition probabilities were partially captured in both sources, but time horizon recognition was consistently flawed—LLMs interpreted annual probabilities as single-cycle (biweekly) values and failed to derive transitions to death or interpret hazard ratios correctly. (Table 2)

Table 2: *Comparison of LLM performance for the publication vs the technical report on transition probabilities*

Aspect	Pub.	Technical Report	Explanation
Correctly assigned probabilities	‼ Partial — mislinked to wrong states	‼ Partial — mislinked to wrong states	Misalignment between state definitions and numeric parameters in source
Time horizon recognition	X All treated as single-cycle	X All treated as single-cycle	Temporal units like "52 weeks" not translated to model cycle (biweekly)
Death transition	X Not captured	X Not used despite death state	LLM failed to infer transitions where probabilities were not directly reported
Excess mortality due to surgery	X Not captured	X Not captured	Reported as HR, not as a probability; LLMs cannot perform statistical conversions

3. Costs

Key finding: Generative AI correctly extracted health state costs when explicitly stated, particularly surgery costs in the technical report, but failed to identify AE, drug, or administration costs due to lack of structured formatting or implied logic. (Table 3)

Table 3: *Comparison of LLM performance for the publication vs the technical report on costs*

Cost Compo nent	Pub.	Technical Report	Explanation
Health state costs	✓ For valid states; missing for hallucinated ones	‼ Linked to an existing state	LLM assumes all identified states should have a cost, even if not reported
Adverse event (AE) costs	X Not captured	X Not captured	AE costs were not explicitly linked to health states or transitions
Drug costs	X Not captured	X Not captured	Costs embedded in treatment pathway logic, not reported as unit costs
Surgery costs	X Not captured	✓ Correctly captured	Clearly labeled as "cost of surgery" in technical report
Admin./ infusion costs	X Not captured	X Not captured	Requires calculation or aggregation not present in plain text

4. Quality of Life (Utility) Inputs

Key finding: Utility values linked to explicitly labeled health states were accurately captured across both sources; however, LLMs also assigned utilities to hallucinated states, showing limitations in filtering out non-existent concepts. (Table 4)

Table 4: *Comparison of LLM performance for the publication vs the technical report on utilities*

Utility Compo nent	Pub.	Technical Report	Explanation
Active disease	✓	✓	Clearly reported and labeled
Remis sion	✓	✓	Clearly reported and labeled
Surgery	X Not captured	Captured	Report explicitly linked utility values to surgery health states
Non-existent states	‼ Assigned	‼ Assigned	LLMs apply utility values even to hallucinated states

5. Treatment Lines

Key finding: Generative AI was unable to recognize treatment sequences or transitions between therapy lines in either source, primarily due to the absence of structured or standardized pathway definitions within the source documents. (Table 5)

Table 5: *Comparison of LLM performance for the publication vs the technical report on treatment lines*

Treatment Component	Pub.	Technical Report	Explanation
1st-line vs. 2nd-line vs. 3rd-line therapies	X Not captured	X Not captured	No structured labeling in source; embedded in flowchart-style logic
Transitions between therapy lines	X Not captured	X Not captured	LLM needs explicit rule-based transitions; not inferred from narrative
Specific drug transitions	X Not captured	X Not captured	Medication switches described narratively without standardized formatting
AE-based treatment discontinuati on	X Not captured	X Not captured	Requires conditional logic that is not interpretable from text

REFERENCES

1. Salcedo, Jonathan, Daniel Hill-McManus, Chloë Hardern, Oyin Opeifa, Raffaella Viti, Ludovica Siviero, Antonio Saverio Roscini, and Gennaro Di Martino. "Cost-Effectiveness of Vedolizumab as a First-Line Advanced Therapy Versus Adalimumab Treatment Sequences for Ulcerative Colitis in Italy." PharmacoEconomics-Open 8, no. 5 (2024): 701-714.
2. ValueGen.AI, https://valuegen.ai/
3. LangChain, https://python.langchain.com/docs/how_to/qa_citations/?form=MG0AV3
4. CrewAI, https://github.com/crewAIInc/crewAI?form=MG0AV3
5. OpenAI, https://github.com/openai/openai-dotnet?form=MG0AV3

Funding: This study was funded by Takeda Pharmaceuticals U.S.A., Inc
Disclosure: Jakob Langer and Ipek Ozer Stillman are employees of Takeda and own company stock.

Contact: Sumeyye Samur, PhD, ssamur@valueanalyticslabs.com