Off-the-Shelf LLM, Human Expert, or Purpose-Built Agent? A Controlled Validation Hackathon Benchmarking Three Approaches to Quality Control of Health Economic Models

MSR181

Tushar Srivastava, Hanan Irfan, Thaison Tong, Shilpi Swami

ConnectHEOR, London, UK. Email: tushar.srivastava@connectheor.com

connectHEOR

> **On two HTA-ready models with 40 seeded errors and a five-hour limit, a purpose-built validator agent detected 38, two senior health economists 22 and an off-the-shelf LLM 18.**

1 BACKGROUND

NICE appraisals have reported technical errors in almost every company cost-effectiveness model [1], and the usual safeguard is a manual, expert-executed checklist review. Cost-effectiveness models underpin **health technology assessment (HTA)** and encode assumptions in tens of thousands of interdependent cells, so errors are common.[2,3]

- **Quality control (QC)** instruments such as the TECH-VER checklist systematise verification, but they are structured, expert-executed and slow.[4]
- **Large language models (LLMs)** are proposed to compress this effort, yet hallucination and inconsistent numerical accuracy are documented, and HTA bodies require validated AI methods.[5,6]

Evidence gap: it is untested whether a general-purpose LLM can perform HTA-grade QC, or whether experts or purpose-built tools are needed.

2 OBJECTIVES

We benchmarked three QC approaches against a known set of seeded errors under one controlled, time-limited protocol, comparing each arm on **sensitivity, precision and time**. **Recalculation-dependent logic errors** were the stratum of principal interest, as the defects least visible in a static workbook and most consequential for a model's results.

Seeding a known set of errors gives an unambiguous denominator, so sensitivity is measured rather than estimated from reviewer consensus.

Off-the-shelf LLM
Frontier model, no domain layer

Human experts
Two senior health economists

AI validator agent
Purpose-built, in-house

3 METHODS

We ran a controlled fault-injection benchmark as a five-hour hackathon. Production-grade models were used rather than simplified test workbooks, because their scale and interdependency are part of what makes QC hard, and all three arms worked from identical copies of the seeded workbooks.

1 Two models
Rare disease Markov (~150 000 cells); metastatic endometrial cancer partitioned survival (~120 000 cells)

→

2 Seed 40 errors
20 per model: numeric or transcription, broken references, recalculation-dependent logic, hygiene

→

3 Run three arms
Off-the-shelf LLM (GPT 5.2), two senior health economists and the AI QC agent, within the same five hours

→

4 Blinded scoring
A blinded panel scored every flag for sensitivity, precision and time

- Seeding errors by type let the analysis distinguish approaches that read a workbook from those that interrogate it: 1) numeric or transcription errors 2) broken references 3) logic/conceptual errors and 4) hygiene.
- The off-the-shelf LLM had no execution layer and was prompted to perform QC on the workbook as presented.
- The five-hour limit reflects a realistic working session of the QC agent and deliberately constrains the human arm, whose accuracy would rise with more time.
- The panel was blinded to arm and [was independent of agent development], removing the obvious route by which scorer expectation could favour the purpose-built system.

4 RESULTS

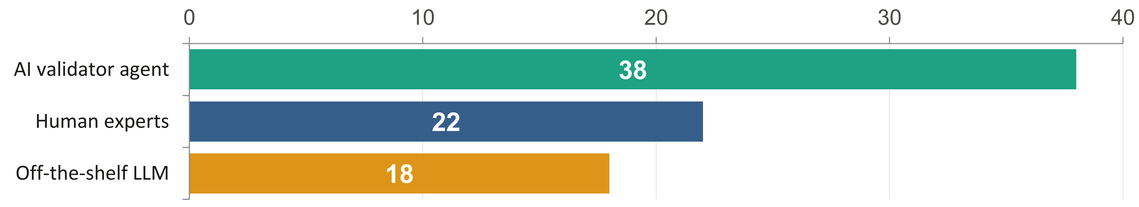
Within the five-hour window the agent found **38 of 40** seeded errors (95%), the human experts 22 (55%) and the off-the-shelf LLM 18 (45%); the human result reflects the five-hour limit, not reviewer skill. Only the agent caught the issues around treatment waning calculation in subgroups and survival-input calculation.

95%
Agent sensitivity (38 of 40 errors)

9 of 10
Logic errors found by the agent

90% vs 31%
Precision: agent versus off-the-shelf LLM

Figure 1: Seeded errors detected by arm (of 40)



The agent completed its pass in about 2 hours per model. The human result shows what best-practice review reaches in five hours, not a ceiling on what senior economists can find.

Table 1: Detection and precision by arm

Arm	Found (of 40)	Logic (of 10)	Precision
AI validator agent	**38**	**9**	**90%**
Human experts	22	5	85%
Off-the-shelf LLM	18	2	31%

The gap was widest on logic errors, which appear only when an input is perturbed and the model recomputed. LLM false positives, including hallucinated references, cut its precision to 31%.

5 DISCUSSION

An off-the-shelf LLM reasons over a **static workbook** and cannot change a parameter and observe the recomputed result, which is exactly what logic errors require.

- This matches evidence that general-purpose models perform poorly on real spreadsheet tasks, and that AI help degrades outside a model's capability frontier.[7,8]
- The implication is a division of labour: the agent as a first-line layer, with senior reviewers adjudicating and remaining accountable.[5,6,9]
- **Two misses of 40:** a gamma rather than beta distribution on one minor probabilistic input (logic) and a missing source note for one cost input (hygiene); neither changed the deterministic base-case ICER.
- Limitations: seeded errors may not mirror natural ones, and only two models and one LLM configuration were tested.

AI, artificial intelligence; HTA, health technology assessment; LLM, large language model; NICE, National Institute for Health and Care Excellence; QC, quality control.

6 CONCLUSION

On two production-grade models with a five-hour limit, a purpose-built, execution-capable validator detected more seeded errors than both an off-the-shelf LLM and two senior health economists, especially on logic errors. An off-the-shelf LLM alone is not enough for HTA-grade validation: it misses errors that appear only on recalculation while generating the heaviest false-positive burden.

- **Execution matters:** only the agent recalculated, finding 9 of 10 logic errors.
- **Next steps:** test natural (unseeded) errors, more models and other LLM configurations.

Pair the agent as a fast first-line layer with senior human adjudication, keeping AI assistive rather than autonomous.

REFERENCES

1. Radeva D, et al. Int J Technol Assess Health Care. 2020;36.
2. Eddy DM, et al. Value Health. 2012;15.
3. Chilcott J, et al. Health Technol Assess. 2010;14.
4. Buyukkaramikli NC, et al. PharmacoEconomics. 2019;37.
5. Fleurence RL, et al. Value Health. 2025;28.
6. NICE. Use of AI in evidence generation: position statement. 2024.
7. Ma Z, et al. NeurIPS Datasets and Benchmarks Track. 2024.
8. Dell'Acqua F, et al. Organization Science. 2026;37.
9. Srivastava T, et al. Adv Ther. 2025;42.

Financial disclosure:
Authors are employees of ConnectHEOR Limited. No external funding received. No conflict of interest to declare.

Poster presented at ISPOR Europe 2026 in Vienna, Austria, 8-11 November 2026