Why Are HTA Reviewers Still Hunting for Coding Errors? An AI Validator Agent as a First-Line QA Layer for Submitted Models

MSR124

Tushar Srivastava, Hanan Irfan, Thaison Tong, Shilpi Swami

ConnectHEOR, London, UK

Email: tushar.srivastava@connectheor.com

> **In a simulated first-line HTA verification of one Markov model with 40 seeded errors, an AI validator agent identified 39, including every critical structural error, and returned a severity-ranked issue list in under four hours against several days manually; simulated reviewers kept 92% of its classifications.**

1 BACKGROUND

Health technology assessment (HTA) teams must confirm that a company model does what its documentation claims before judging whether it fits the decision problem. Technical errors are common: a review of NICE single technology appraisals completed in 2017 found errors in **almost all** of them.[1]

- Good practice recommendations, AdViSHE and TECH-VER formalise verification [2–4], but guidance cannot supply reviewer hours.
- Submission volumes and assessor capacity are widely reported as under strain [add reference], so time spent on arithmetic is not spent on comparators or extrapolation.

Evidence gap: generative AI has been proposed across the HTA evidence pipeline [5], yet, as far as we are aware, published applications sit mainly on the submitter's side, not the agencies.

2 OBJECTIVES

We tested whether an AI model validator agent, deployed as a **first-line quality-control (QC) layer**, could speed up verification of a company-submitted model while preserving reviewer control.

1. SPEED
Test whether the agent cuts time to a severity-ranked issue list compared with multi-day manual checking of the same model.

2. DETECTION
Quantify how many of a known set of errors it detects, especially critical structural errors, and how often its flags are real.

3. CONTROL
Establish whether speed gains preserve reviewer control, measured as the share of classifications retained.

3 METHODS

We simulated an HTA body's first-line verification, with the agent in the assessor's seat rather than the modellers. It worked from the Excel workbook and documentation of a **submission-quality six-state Markov model** in early triple-negative breast cancer.

1. Ingest model and documentation → 2. Write a sheet-by-sheet summary → 3. Reviewer approves the summary → 4. Check against the approved map → 5. Rank issues by severity → 6. Reviewer retains or reclassifies

↻ **Reviewer in control: findings loop back to the reviewer, who can correct the map or overturn a flag**

Approving the summary is the first governance control: every later flag points to a location the reviewer has already seen.

- We used a simulation because it fixes the number of errors in advance; in a live appraisal, the true error count is unknown, so detection rates cannot be measured.
- We seeded 40 errors graded by difficulty across transition logic, discounting, state-utility mapping, cohort accounting and cross-referencing, the high-risk areas in TECH-VER.[4]
- The agent's output was compared with multi-day manual checking of the same model, and simulated reviewers judged each flag, retaining or overturning its severity.
- Outcomes were errors detected, time to a severity-ranked issue list, positive predictive value (PPV) and the share of classifications retained.

4 RESULTS

The agent identified **39 of 40 seeded errors**, including every critical structural error. First-pass verification took under four hours per model against multi-day manual checking, and false positives were rare.

97.5% of seeded errors identified

~80% less time to a severity-ranked issue list

94% positive predictive value



Outcome	Result	Why it matters
Critical structural errors	**All identified**	These are the errors that can move an ICER
Time to ranked issue list	**Under 4 h vs multi-day**	Reviewers start with findings that could change the result
Positive predictive value	**94%**	The flags reviewers worked through were mostly real
Classifications retained	**92% of flags**	Overriding remained a live option, not a formality

Detection held across graded difficulty, so it did not depend on errors being conspicuous. The single miss was a face-validity judgement, not a computational error.

Because the output was a ranked list rather than a list of anomalies, the time saving was usable: hygiene-level items come last, or not at all if the submission is returned first. The reclassified minority shows reviewers stayed in control.

5 DISCUSSION

The agent carried out the **full structural audit**, tracing every transition, cohort, discount factor and cross-sheet link in the workbook, and caught every critical error within hours. Experts are left free for the judgements only they can make.

- **Where it fell short:** it is one miss was a face-validity error (is a trajectory clinically plausible?), which needs clinical judgement.
- **Why speed matters:** finding defects in hours, not a week, allows earlier clarification questions under fixed national appraisal timelines such as NICE's.
- **Fit with HTA guidance:** reviewers can see and override every flag, as NICE and Canada's Drug Agency expect.[6,7]
- **Limitations:** one simulated model, seeded rather than real errors, and simulated reviewers.

AI, artificial intelligence; NICE, National Institute for Health and Care Excellence; QA, quality assurance.

6 CONCLUSION

As a proof of concept, the validator is a **first-line QC enabler** for HTA reviewers, not a replacement for their critique: in a simulated review it caught every critical structural error and returned a ranked issue list in under four hours.

It could let a reviewer verify several models in parallel, frees assessors for methodological judgement and leaves a line-by-line audit trail.

With 94% of flags confirmed as real by simulated reviewers, reviewers spend little time dismissing false alarms.

Reviewers reclassified 8% of flags, so severity stays a human decision. The next step is prospective evaluation with an HTA reviewer on naturally occurring errors.

REFERENCES

1. Radeva D, et al. Int J Technol Assess Health Care. 2020;36.
2. Eddy DM, et al. Value Health. 2012;15.
3. Vemer P, et al. PharmacoEconomics. 2016;34.
4. Büyükkaramikli 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. Canada's Drug Agency. Position statement on the use of AI. 2025.

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

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