Can a Model Validator Agent Expedite Your Submissions to HTA Bodies via a Continuous Validation Loop During Model Development?

MSR125

Tushar Srivastava, Hanan Irfan, Shilpi Swami

ConnectHEOR, London, UK

Email: tushar.srivastava@connectheor.com

> **Run at four milestones during the build of one Markov cost-effectiveness model rather than as a final check, an AI validator agent detected 23 of 24 seeded errors, and critical issues fell from 12 to 0 before sign-off; a validation pass took about 3–4 hours against about 3–4 days manually (team estimate).**

1 BACKGROUND

Quality control (QC) of Excel cost-effectiveness models (CEMs) **underpins the credibility** of health technology assessment (HTA) submissions, yet it remains manual, slow and usually concentrated at the end of development, when errors are costliest to fix.

- A methodological review produced a taxonomy of model errors and showed how rarely development processes are designed to prevent them.[1]
- A review of National Institute for Health and Care Excellence (NICE) appraisals found technical errors in **almost every company model**, most often transcription, logic and computational errors.[2]

Evidence gap: generative AI is proposed as an assistive HTA layer under human oversight [3], but its use as a **continuous layer** during development is untested.

2 OBJECTIVES

We evaluated whether an AI model validator agent, used as a **continuous in-development QC layer** rather than a final gate, could shorten the build-and-validate cycle of a complex CEM before HTA submission while keeping expert oversight.

1 DETECTION
Measure detection against 24 deliberately seeded errors of graded difficulty across four mechanism classes.

2 SEVERITY
Characterise the issues surfaced and their severity across successive model versions, up to pre-sign-off.

3 EFFICIENCY
Record the effect on validation effort, rework and time to sign-off readiness against an equivalent manual pass.

3 METHODS

The case study was a **submission-grade Markov CEM** in early or locally advanced triple-negative breast cancer (M0, Stage II–III), built to the NICE reference case [4] with 6 health states, monthly cycles, a lifetime horizon, ~185 000 cells and 14 VBA modules: too large for exhaustive manual checking within a development schedule.

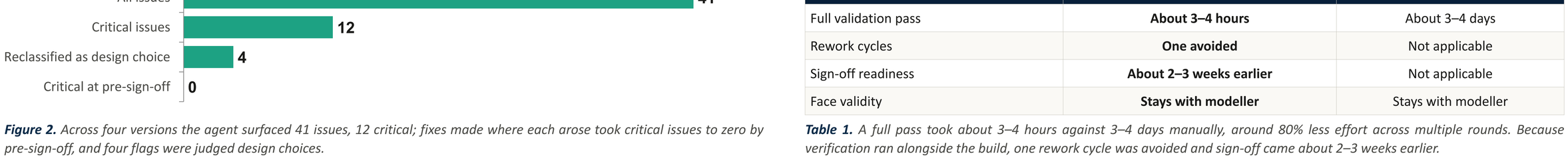


Figure 1. *The loop ran on each new model version at four milestones, so each finding was fixed and re-verified in the milestone where it arose.*

- Deployment was the variable tested: the agent checked the **live build at each of 4 development milestones** up to pre-sign-off, not the finished model.
- Findings were **fixed in the milestone that raised them**, and a **selective re-run** of the affected checks confirmed each fix.
- Static checks confirm formulas match intent; only recalculation exposes **state-dependent faults**, and static spreadsheet reasoning stays hard for language models.[5]
- Detection was tested with **24 seeded errors** across transition logic, discounting, utility mapping and cohort accounting; outcomes were tracked per version.

4 RESULTS

The agent detected **23 of 24 seeded errors (96%)**, including two dynamic-logic faults visible only on recalculation: transition probabilities exceeding one under a subgroup toggle, which moved more than the whole cohort out of a state, and a hazard ratio applied beyond its waning cycle, which kept the treatment effect running past its intended point.

96% of seeded errors detected (23 of 24)

12 → 0 critical issues surfaced, falling to zero by pre-sign-off

~80% less effort per full validation pass

Issues flagged across four model versions (n)

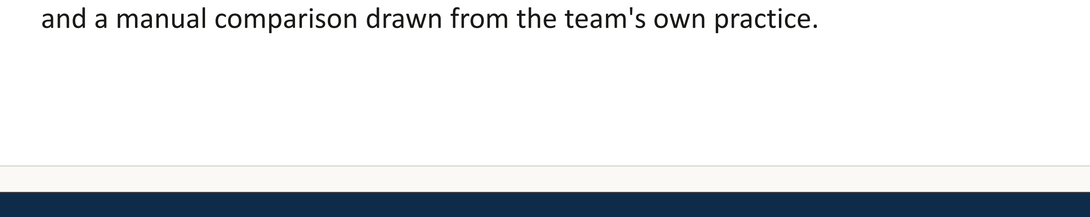


Figure 2. *Across four versions the agent surfaced 41 issues, 12 critical; fixes made where each arose took critical issues to zero by pre-sign-off, and four flags were judged design choices.*

Validator agent vs an equivalent manual pass

Measure	Validator agent	Manual pass
Full validation pass	**About 3–4 hours**	About 3–4 days
Rework cycles	**One avoided**	Not applicable
Sign-off readiness	**About 2–3 weeks earlier**	Not applicable
Face validity	**Stays with modeller**	Stays with modeller

Table 1. *A full pass took about 3–4 hours against 3–4 days manually, around 80% less effort across multiple rounds. Because verification ran alongside the build, one rework cycle was avoided and sign-off came about 2–3 weeks earlier.*

- Because verification ran **alongside the build**, fixes landed at the milestone where each error was introduced, and **one rework cycle was avoided**.

5 DISCUSSION

Detection follows from the **architecture**: a fault spanning a toggle, a multiplier and a cycle counter sits in no single formula and appears only on recalculation.

- The value lies in **timing**: an error found at pre-sign-off costs a rework cycle, but at the milestone where it arose it costs an afternoon.
- This matters under fixed clocks, such as NICE clarification turnarounds.[6]
- Severity classes with explicit rationale give assessors an audit trail, in line with NICE's call for justified, transparent AI use with human involvement.[7]
- Limitations: one model and one team, seeded errors that may not match natural ones, and a manual comparison drawn from the team's own practice.

6 CONCLUSION

Embedding an overridable, auditable validator agent **throughout development** removed every critical error before sign-off and cut each validation pass from days to hours (team estimate for the manual pass), while methodological judgement and face validity stayed **human**.

Fixing each error at the milestone where it arose avoided a rework cycle and, by team estimate, brought sign-off **2–3 weeks** forward. Testing across further disease areas and model types should confirm whether these gains generalise.

REFERENCES

1. Chilcott J, et al. Health Technol Assess. 2010;14.
2. Radeva D, et al. Int J Technol Assess Health Care. 2020;36.
3. Fleurence RL, et al. Value Health. 2025;28.
4. NICE. Health technology evaluations: the manual (PMG36). 2022.
5. Ma Z, et al. Adv Neural Inf Process Syst. 2024;37.
6. European Commission. Joint clinical assessments, Regulation (EU) 2021/2282. 2025.
7. NICE. Use of AI in evidence generation (ECD11). 2024.

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