Chaining Specialised AI Agents in a Governed Multi-Agent Workflow for Validating, Explaining, and Reporting an Oncology Cost-Effectiveness Model

MSR182

Tushar Srivastava, Hanan Irfan, Shilpi Swami

ConnectHEOR, London, UK

Email: tushar.srivastava@connectheor.com

> **In one oncology partitioned survival model, a chain of validator, explainer and report-writer agents with human sign-off between stages detected 21 of 22 seeded errors and produced a first-draft technical report in under one working day, against 6–8 weeks for the usual manual process (team estimate).**

1 BACKGROUND

Validating a company cost-effectiveness model, explaining it and writing the technical report typically takes **6–8 weeks of senior time** (team estimate), and every handover is another place for an error to enter. Company models are built in Excel, and NICE appraisals have reported technical errors in almost all models. [1]

- **Verification and validation:** analysts check every model by hand, using TECH-VER for technical verification and AdViSHE for validation.[2,3]
- **Reporting:** analysts then explain the model and write the technical report.
- **Opportunity:** AI agents can now read workbooks, run checks and draft text.

Evidence gap: as far as we know, published evaluations test one AI agent on one task. In a chain, each agent inherits upstream errors, a leading cause of multi-agent failure.[4]

2 OBJECTIVES

We built and tested **three AI agents (model validator, explainer and report writer) chained under human-cleared gates** to validate, interpret and document a completed cost-effectiveness model in advanced breast cancer.

1. Chain
Test whether multi-agent AI workflows in HEOR can hand work to each other effectively, acting as a team, with human review.

2. Accuracy
Compare the AI agents' accuracy in quality checking the model, explaining the model and drafting evidence for it

3. Gains and governance
Measure time saved per deliverable and the human checks needed to keep handoffs defensible.

3 METHODS

The artefact was a completed **partitioned survival model** in HER2-negative germline BRCA-mutated advanced breast cancer, evaluated from a National Health Survey (NHS) and Personal Social Services (PSS) perspective. The workflow ran as a staged chain with a **human-cleared gate** between stages driven by various specialized AI agent.

1. Validator agent → Runs static and dynamic checks to QC and grades each finding by severity.

2. Human-cleared gate → A senior health economist signs off the model before it moves on.

3. Explainer agent → Answers the report writer's input questions from the model with justifications

4. Report writer agent → Drafts a full technical report from the explainer's answers.

5. Consistency review → Human reviewers check the report against the model.

↺ Failing models return for correction until QC passes.

↺ The validated model and the report are re-loaded into the explainer for review.

- **Validator:** static checks interrogate formulas and references, dynamic checks alter inputs to expose hidden faults, and quality control (QC) must pass before handoff.
- **Explainer:** answers the report writer's input-form questions from the workbook, not from training data, so each answer is checkable.[5]
- **Report writer:** drafts the technical report from the explainer's answers; the explainer then checks report statements back against the model.
- **Evaluation:** senior health economists adjudicated every checkpoint; efficiency was elapsed time and reviewer effort per deliverable versus the manual process.

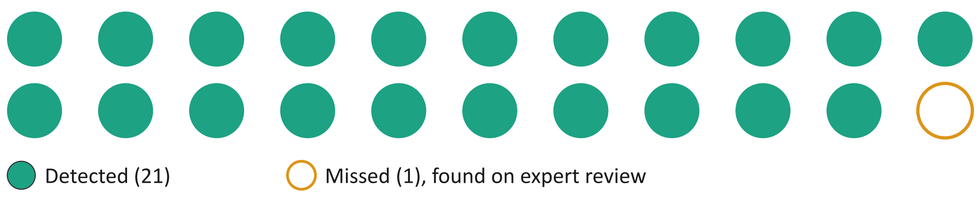
4 RESULTS

The chain took the model from validation to first-draft report in **under one working day** (vs 6–8 weeks), while the validator caught **21 of 22 seeded errors (95%)**.

>90% less validation time, at under three hours per cycle

88% of report writer input-form fields answered without manual entry

<1 day to a first-draft report, down from six to eight weeks



Detected faults included a **state-utility mapping fault** which however let the workbook compute without complaint.

Deliverable	Before	With agent chain
Validation cycle	Manual pass	**Under 3 hours (>90% less time)**
Input-form preparation	Manual entry	**≈65% less effort**
Reviewer onboarding	20 hours	**6 hours**
First-draft report	6–8 weeks	**Under 1 working day**

- **Explainer:** cut input-preparation effort by about 65% and reviewer onboarding from 20 to 6 hours.
- **Report writer:** drafted the full technical report in under one working day, down from 6–8 weeks.

5 DISCUSSION

The agents performed best where they could check their work against the workbook: **validating formulas, answering model questions and drafting text**. The one missed error needed **judgement on whether a plausible result was right**, which no agent could anchor to the model.

- **Why gates matter:** errors compound along a chain because downstream agents treat upstream output as fact [4], and reviewers tend to accept automated findings.[6]
- **Implications:** time saved frees senior staff for judgement, in line with NICE's view that AI should augment experts [7]; AI use is reported per ELEVATE-GenAI.[8]
- **Limitations:** one model and indication, seeded rather than natural errors, and no concurrent control for efficiency.

AI, artificial intelligence; BRCA, breast cancer gene; HER2, human epidermal growth factor receptor 2; NHS, National Health Service; NICE, National Institute for Health and Care Excellence; QC, quality control.

6 CONCLUSION

A **validation-gated chain** of specialised agents, bounded by expert adjudication, caught 95% of seeded errors in one oncology model and cut first-draft reporting from an estimated 6–8 weeks to under a day, without displacing methodological judgement.

The one missed error shows why the human gate must stay: multi-agent orchestration works as **auditable decision support**, not autonomous delivery.

The human-cleared gate, the audit trail and named accountability must be treated as part of the system.

REFERENCES

1. Chilcott J, et al. Health Technol Assess. 2010;14.
2. Büyükkaramikli NC, et al. PharmacoEconomics. 2019;37.
3. Vemer P, et al. PharmacoEconomics. 2016;34.
4. Cemri M, et al. arXiv:2503.13657. 2025.
5. Lewis P, et al. Adv Neural Inf Process Syst. 2020;33.
6. Goddard K, et al. J Am Med Inform Assoc. 2012;19.
7. NICE. Use of AI in evidence generation: position statement. 2024.
8. Fleurence RL, et al. Value Health. 2025;28.

Financial disclosure: All authors are employees of ConnectHEOR Limited. No external funding was received. The authors have no conflict of interest to declare.

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