An Agentic AI Model Validator for Automated Quality Control of Excel-Based Budget Impact Models

MSR235

Tushar Srivastava, Hanan Irfan, Syed Salleh

ConnectHEOR, London, UK

Email: tushar.srivastava@connectheor.com

connectHEOR

> **Agentic AI validator caught 96% of budget impact model seeded errors in under 90 minutes, so senior reviewers can focus on the assumptions that matter.**

1 BACKGROUND

A budget impact model can be arithmetically consistent and still wrong by millions: in one of our test models, a seeded population double-count overstated five-year budget impact by about 14%, and no single formula was wrong.

Budget impact analysis (BIA) estimates the financial consequences of adopting a new intervention for a defined budget holder, and it is a common requirement in reimbursement submissions.[1,2]

- Model verification tools are mature, yet **quality control (QC)** of Excel-based **budget impact models (BIMs)** remains manual, slow and inconsistently documented.[3,4]
- BIM defects concentrate in population funnels, market share and uptake, discounting conventions and cost-year indexing.

Gap: errors left in place can materially distort projected expenditure, and any AI-assisted evidence must be justified, transparent and validated.[5,6]

2 OBJECTIVES

We evaluated an agentic AI **Model Validator Agent** designed as an auditable, first-line QC layer for Excel-based BIMs, benchmarked against senior human validators. *Together, the three aims define how far the agent can safely be relied upon.*

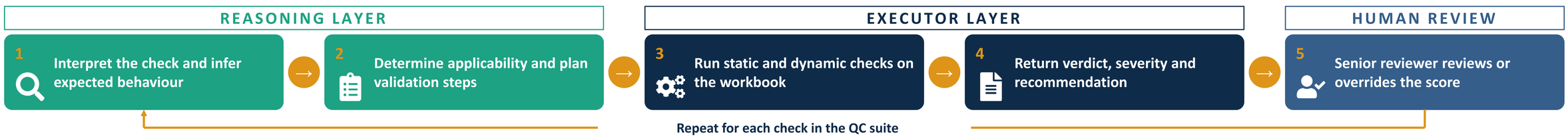
Accuracy
Measure detection of seeded errors of graded difficulty under controlled fault injection.

Speed
Measure QC completion time per model and compare it with manual review.

Boundaries
Characterise failure modes, so the limit of reliance is stated rather than assumed.

3 METHODS

For every check, a **reasoning layer** interprets the check, infers how a correct model should behave, determines applicability and plans validation. An **executor layer** then runs static checks that read the workbook and dynamic checks that perturb inputs or switch scenarios in the Excel model to test the model against its own logic.



Each check passes through reasoning, execution and human review, so every verdict arrives with its evidence and stays overridable. The agent performs technical verification; face validity and conceptual validation stay with the modeller.

- **Outputs:** each check returns a pass or fail verdict, supporting observations, a severity of critical, moderate or minor, and a corrective recommendation, aggregated into a model score that a senior reviewer can reclassify.
- **Models:** two production-grade, anonymised 5-year BIMs, one in endometrial cancer (~10 worksheets, ~60 000 cells) and one in chronic obstructive pulmonary disease with a mixed prevalent and incident population (~13 worksheets, ~80 000 cells).
- **Benchmark:** 24 errors of graded difficulty were seeded across population funnels, market-share displacement, discounting, annualisation, cost-year indexing and scenario-toggle logic; senior human validators reviewed the same workbooks, and outcomes were detection and QC time.

4 RESULTS

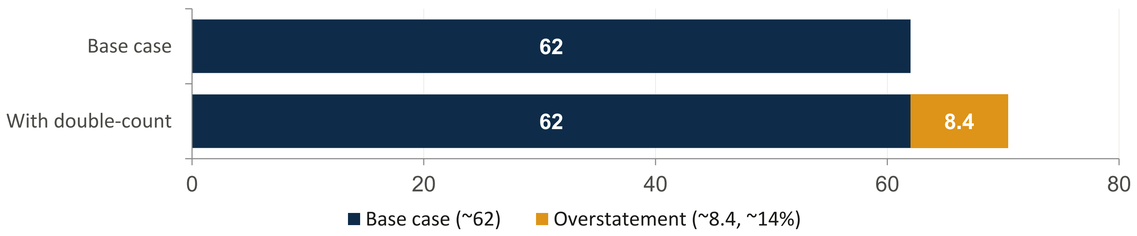
The agent detected **23 of 24 seeded errors** across the two models, a sensitivity of 96%. Automated QC completed in under 90 minutes per model, over 90% faster than manual review, and a few false positives needed adjudication.

23/24 seeded errors detected (96% sensitivity)

<90 min of automated QC per model

>90% faster than manual review

Sample error detected: Five-year budget impact (GBP million) calculation



One double-count in the incidence funnel overstated five-year budget impact by about GBP 8.4 million (~14%).

This was the **largest quantitative error,** set against a base case near GBP 62 million. This error was seeded; in a live model, an error of this size would change the budget impact a payer sees.

What the agent found in the seeded errors

Seeded error	Finding	Outcome
Scenario toggle	Identical inputs gave non-zero incremental impact	**Detected (critical)**
Uptake linkage	0% uptake failed to zero projected spend	**Detected (critical)**
Market share	Shares did not sum to 100% under toggles	**Detected**
Population funnel	Incidence double-count (~GBP 8.4m)	**Detected**
Inflation indexing	Inconsistency in an appendix table	**Missed**

Critical faults surfaced only on recalculation; the single miss was an appendix-table inconsistency.

These errors discovered by the agent are generally hard to find by inspecting individual formulas. The single miss was minor: an isolated appendix inconsistency that does not feed into the model's budget impact outputs.

5 DISCUSSION

The model validator AI agent excelled wherever correctness can be **verified by recomputation**, the linked logic that drives every budget projection.

- It also caught faults that surface only when the model is run, not just read.
- Mechanical checking becomes **auditable and reproducible**, freeing senior reviewers for face validity, epidemiology and key assumptions.
- This aligns with HTA guidance and prior work positioning AI as a governed, assistive component under human accountability.[5-8]
- **Next steps:** extend testing beyond two single-organisation models to prospective, unseeded builds and targeted handling of isolated, low-link content.

6 CONCLUSION

An agentic, reasoning-based validator delivered **accurate, rapid and auditable** first-pass QC for Excel-based BIMs with seeded errors, detecting population, uptake and aggregation faults that materially affect projected expenditure.

- Running the model exposed faults that **formula review alone does not.**
- Next step: **prospective testing** on unseeded builds, more models and more organisations.

It complements, rather than replaces, senior expert review.

REFERENCES

1. Mauskopf JA, et al. Value Health. 2007;10.
2. Sullivan SD, et al. Value Health. 2014;17.
3. Eddy DM, et al. Value Health. 2012;15.
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. Srivastava T, et al. Adv Ther. 2025;42.
8. Reason T, et al. PharmacoEconomics Open. 2024;8.

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