When AI Gets HEOR Wrong: A Taxonomy of Silent Failure Modes in AI-Assisted Health Economic Work and How to Detect Them

MSR47

Tushar Srivastava[1], Hanan Irfan[1], Kunal Hriday[2], Shilpi Swami[1]

[1]ConnectHEOR, London, UK; [2]ConnectHEOR, Delhi, India.

Email: tushar.srivastava@connectheor.com

> **The most dangerous AI failures in HEOR are the ones that look right: they survive face validity and are caught mainly by independent recomputation, source tracing and pre-specified checks.**

1 BACKGROUND

Artificial intelligence (AI) is now used across the **health economics and outcomes research (HEOR)** workflow, from screening to economic modelling, yet the literature is dominated by demonstrations of capability.[1]

- The National Institute for Health and Care Excellence (NICE) expects AI use to be justified, reported transparently and its new risks mitigated.[2]
- Technical errors are already common in models submitted to NICE, with transcription, logic and computational errors the most frequent.[3]
- AI often changes the character of these errors rather than their existence: outputs arrive **fluent, complete and confident**, removing the cues reviewers rely on.

Gap: far less work reports how AI-assisted HEOR fails, or which control catches each failure.

2 OBJECTIVES

The study set out to turn scattered failure cases into an actionable taxonomy, with three aims.

1. Synthesise failure cases observed in controlled studies of AI-assisted HEOR into a **structured taxonomy** of failure modes.
2. Characterise each mode by its mechanism, its position in the workflow and its consequence for a submission, noting whether it **survives a standard face-validity check**.
3. Pair each mode with its **detection signature** and the human or automated control that reliably catches it.

3 METHODS

We pooled failure cases from completed in-house studies spanning the AI-assisted HEOR workflow, rather than reporting each study separately. Each failure was classified on **four axes**, then mapped to the detection method and the human or automated control that reliably catches it.

1 Pool failure cases

- Cost-effectiveness and budget impact model validation
- Discrepancies between off-the-shelf model vs human expert vs purpose-built agent
- Model interrogation for assessors
- Systematic literature review screening and extraction

→

2 Classify on four axes

- Mechanism
- Survival of a face-validity check
- Position in the workflow
- Consequence for a submission

→

3 Map each mode to a control

- **Detection:** how the failure is found, e.g., recomputation or source tracing
- **Control:** who applies it, a named reviewer or an automated check
- **Output:** every mode has a named check before sign-off

- Pooling was deliberate: a mechanism that recurs across several tasks is a property of the technology, not of one task.
- The strands placed the AI in different roles, from interrogating an existing model to producing, reviewing and extracting evidence at volume.
- Survival of a face-validity check carried most weight: a failure caught on first reading is a nuisance, but one that survives is a risk.
- Workflow position shows how far an error can propagate, and consequence for a submission makes a technical defect decision-relevant.

4 RESULTS

Failures grouped into recurring modes. The highest-risk modes shared one property: they **survive face validity**, so they are caught not by reading the output but by independent recomputation, source tracing and pre-specified checks.

Failure mode	What goes wrong	What catches it
Fabrication	Plausible but non-existent references, comparators or values appear; fluency makes this the most damaging mode.	Source tracing: verification against the source
Recalculation-dependent logic error	Invalid results, such as transition probabilities > 1 under a subgroup, pass static inspection.	Independent recomputation: forced recalculation
Confident mis-extraction	Correct-looking but wrong inputs are pulled from source documents and propagate silently.	Source tracing: comparison with the source
Drift	Outputs change as the underlying model updates, breaking earlier validation.	Validation that carries a date and a model version
Automation complacency	Errors are missed because a fluent output discouraged scrutiny by the reviewer.	Verification against an external referent, not closer reading

Table 1. Every failure mode has a named detection route, and none is caught by re-reading the output.

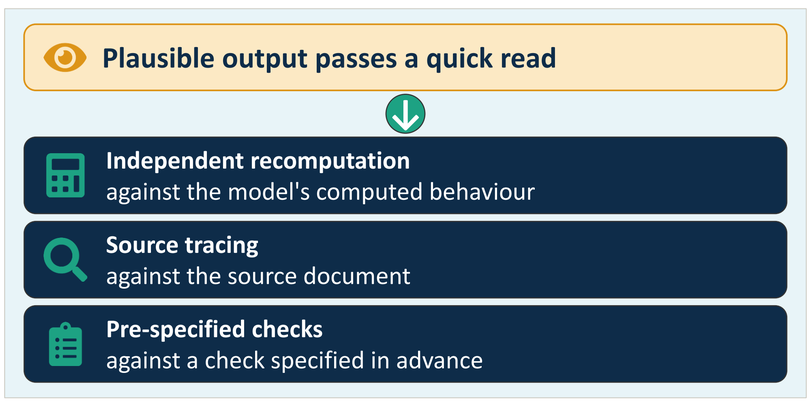


Figure 1. Only an external comparison catches the highest-risk modes.

Fabrication and mis-extraction pass inspection because the values sit in the range a reviewer expects. Logic errors are not visible in the model; they appear only when it computes under the relevant parameters.

5 DISCUSSION

LLMs often optimise for **plausibility** rather than correspondence with an external referent, so their errors are, by construction, the ones that look right. Review effort should therefore be **reallocated rather than increased**.

- Fabricated references are a documented instance, with high rates reported in model-generated medical content.[4]
- Automation bias [5] and model drift [6] are documented, so validation of an AI-assisted workflow must carry a date and a model version.
- The modelling field already has verification instruments, TECH-VER and AdViSHE [7,8], that support external checks.
- The cases come from one organisation and the taxonomy is descriptive, so it does not estimate how often each mode occurs.

6 CONCLUSION

AI-assisted HEOR fails in **identifiable, recurring ways**, and the most dangerous failures are the ones that look right.

- Each failure mode is paired with a **named check**, so validation goes where errors hide.
- In our cases the effective checks were **recomputation, source tracing and pre-specified tests** rather than closer reading.
- HTA bodies gain a **shared vocabulary** for asking how AI was used and verified.

Named reviewers verify and remain accountable for any AI-assisted output.

REFERENCES

1. Fleurence RL, et al. Value Health. 2025;28.
2. NICE. Position statement ECD11. 2024.
3. Radeva D, et al. Int J Technol Assess Health Care. 2020;36.
4. Bhattacharyya M, et al. Cureus. 2023;15.
5. Goddard K, et al. J Am Med Inform Assoc. 2012;19.
6. Chen L, et al. Harvard Data Sci Rev. 2024;6.
7. Buyukkaramikli NC, et al. PharmacoEconomics. 2019;37.
8. Vemer P, et al. PharmacoEconomics. 2016;34.

Financial disclosure: The 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