Governance in Practice: An Audit-Trail and Validation-Log Standard for AI-Generated HEOR Deliverables

MSR177

Shilpi Swami, Tushar Srivastava

ConnectHEOR, London, UK

Email: shilpi.swami@connectheor.com

> **We propose a log entry at every checkpoint of an AI-assisted SLR, cost-effectiveness model or dossier that records the model, settings, inputs and prompt version, how validation was shown, and who signed off, so that an independent reviewer can repeat and challenge the work.**

1 BACKGROUND

"We have a human in the loop" does not tell a reviewer who checked a dossier figure, against what standard, or with which model version. AI now contributes directly to the deliverables that carry health economics and outcomes research (HEOR) evidence into reimbursement decisions.[1]. Governance has stayed mostly high-level and principle-based.

- Health technology assessment (HTA) bodies such as NICE expect AI use to be justified, transparently reported and backed by evidence of risk management.[2,3]
- ELEVATE-GenAI sets out what a study assisted by a large language model must disclose, but not where records are made or who clears them.[4]
- The EU AI Act, GxP data integrity and model risk management in banking (SR 11-7) already rely on logged, signed checkpoints.[5-7]

Gap: none of this guidance settles the operational layer, and we are not aware of an equivalent standard for HEOR deliverables.

2 OBJECTIVES

We aimed to develop an **audit-trail and validation-log standard** that sets out, for any AI-assisted HEOR deliverable, where checks must happen and what each check must record.

1 WHERE
Set out where checks must happen across the stages of the SLR, cost-effectiveness model and dossier.

2 WHAT
Fix, at each check, what is recorded, how validation is shown and who signs off.

3 REPEAT
Accumulate a record from which an independent reviewer could repeat and challenge the work.

The objectives build from where checks happen to a record others can test.

3 METHODS

We took core HEOR deliverables in which AI is now used such as the SLR, cost-effectiveness model and dossier, and broke each into its natural stages. The **deliverable**, not the AI tool, was the unit of analysis, because governance attaches to the artefact that reaches a committee.

Select deliverables → SLR, cost-effectiveness model and dossier

Break into natural stages → Checks land where a review already exists

Place checkpoints → Wherever an output must be approved, plus final sign-off

Define the minimum record → Enough for an independent reviewer to repeat and challenge

Checkpoints sit where an output becomes the input to a downstream step, because an unchecked error there becomes load-bearing for everything after it.

- A **checkpoint** was placed wherever an AI-assisted output must be checked and approved before the work moves on, including the final sign-off before release.
- At each checkpoint we defined the **minimum record** an independent reviewer would need to repeat and challenge the work, a higher bar than disclosure alone.
- Working from natural stages means checks land where a review already exists, instead of creating a compliance process parallel to the scientific one.
- HTA AI position statements set the oversight expectations, ISPOR guidance the reporting fields, and standard audit practice the control pattern.[2–4]

4 RESULTS

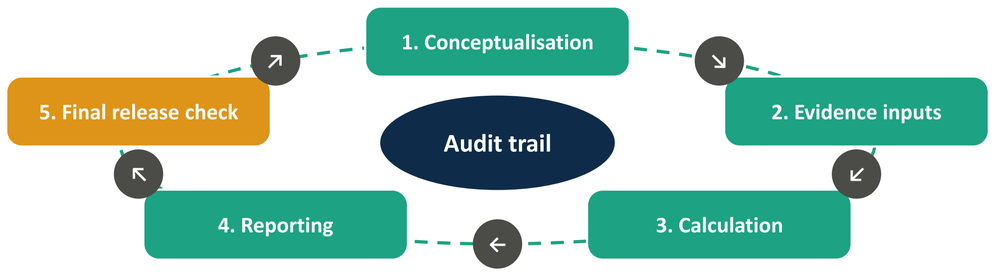
At every checkpoint, a **permanent log entry** records three things: what is recorded, how validation is shown and who signs off. Together the entries form the **audit trail**, closed by a final release check with clear re-validation triggers.

One permanent log entry at every checkpoint

Three things recorded in each entry

Named human accountability as the binding control

Log entry part	What the entry holds
What is recorded	Model and version, settings, inputs, prompt version, and a timestamped copy of the AI output.
How validation is shown	Method, reference standard, extent checked, pass thresholds set in advance, errors by severity and how each was resolved, and the named reviewer.
Who signs off	Set by how critical the step is: analyst, independent senior health economist or accountable lead, with separate quality control for the most critical steps.



Recording the model alone is not enough: the same model with different settings or a different prompt version can produce different outputs. For reproducibility, all three must be captured together

The trail accretes in the order the work was done. Re-validation triggers mark when an entry stops being current and its check must be repeated.

5 DISCUSSION

A principle such as human oversight has no specific location in a workflow, so its absence is hardly detectable. Decomposing a deliverable into stages turns it into a bounded set of stages, each of which either has an entry or does not.

- Where the trail exists, a dossier figure traces back through its checkpoint to the model, settings, inputs, prompt version and output, and to its named reviewer.
- This is closer to what HTA processes ask than non-standardised metrics for a tool, and it demonstrates justification and transparency rather than asserting them.[2,4]
- Anchoring the standard to the deliverable that HTA bodies assess, and grading sign-off by criticality, lets it scale down as readily as up.
- We report an internally developed design that has not had external consensus review or measured adoption; its value depends on tooling that captures provenance and on local agreement about checkpoints.

AI, artificial intelligence; HEOR, health economics and outcomes research; HTA, health technology assessment; NICE, National Institute for Health and Care Excellence; SLR, systematic literature review.

6 CONCLUSION

- Treating AI governance as a logged, step-by-step audit trail within each HEOR deliverable moves teams to the **evidenced oversight HTA bodies expect**.
- **Named human accountability** remains the binding control.
- Checkpoints sit at **existing review steps**, so governance builds on the scientific workflow instead of running beside it.
- **Re-validation triggers** keep the trail current when models, prompts or inputs change.

Real-world adoption studies are the next step and would also test whether reviewers outside the producing team can use the trail.

REFERENCES

1. Fleurence RL, et al. Value Health. 2025;28(2).
2. NICE. Position statement ECD11. 2024.
3. Canada's Drug Agency. Position statement on AI. 2025.
4. Fleurence RL, et al. Value Health. 2025;28(11).
5. Regulation (EU) 2024/1689 (AI Act). 2024.
6. MHRA. GXP data integrity guidance. 2018.
7. Federal Reserve; OCC. SR 11-7. 2011.

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