The Scenario Analyses You Forgot Are the Ones the EAG Runs: An LLM Scenario-Generation Agent Benchmarked Against Omitted and EAG-Imposed Scenarios in NICE Oncology Appraisals

MSR63

Tushar Srivastava, Hanan Irfan, Shilpi Swami
ConnectHEOR, London, UK.
Email: tushar.srivastava@connectheor.com

> **Most of the scenario analyses EAGs imposed could have been found before submission. Across 7 NICE oncology appraisals, an LLM found 21 of 26 from the company's own model, plus 8 valid scenarios no one had run.**

1 BACKGROUND

Scenario analysis is the main way that **structural uncertainty** is made visible in a cost-effectiveness model. Probabilistic analysis handles parameter uncertainty, but judgements such as which extrapolation to trust or how long an effect persists can only be examined by rerunning the analysis under an alternative assumption.

- The NICE manual expects alternative scenarios where effects are projected beyond observed data, including ones less favourable to the technology.[1]
- Evidence assessment group (EAG) analyses are highly influential [2], and our 2022 review (MSR114) found **frequent EAG requests** for additional analyses.[3]

The gap: companies know that omissions attract EAG analyses, but they have no systematic way to find the omissions in their own model before submission.

2 OBJECTIVES

We evaluated a retrieval-grounded large language model (LLM), run as a **pre-submission EAG simulation**, against three aims.

1 FIND
Identify clinically and methodologically relevant scenario analyses omitted from company submissions.

2 BENCHMARK
Compare the proposed scenarios with the additional analyses EAGs actually requested or conducted in completed NICE oncology appraisals.

3 TRACE
Establish whether proposed scenarios are traceable to submitted model inputs and free of unsupported assumptions.

The aims move from finding omitted scenarios to proving that each one can be traced.

3 METHODS

We built a **reference standard** from 7 NICE oncology single technology appraisals (2022–2024): 41 company scenarios and 34 analyses requested or conducted by EAGs, 26 of which the company had omitted. The LLM reviewed each submitted model and evidence base, grounded in retrieved source material.[4]

1 Reference standard
41 company scenarios and 34 EAG analyses extracted from 7 appraisals.

→ **2 Retrieve**
The LLM reads each submitted model, company submission and evidence base.

→ **3 Identify**
It flags assumptions likely to be challenged at technical engagement or committee.

→ **4 Rank and describe**
Each proposal states its assumption, decision relevance, rationale and model input.

→ **5 Senior review**
Two senior health economists judge validity, materiality and traceability.

- Benchmark: the 26 EAG-imposed scenarios the company omitted, each judged material enough to run by an independent reviewer.
- Each proposal had to name its model input, since a scenario that cannot be tied to an input is not implementable.
- Outcomes: recall of the 26 scenarios, precision, unsupported-scenario rate and **novel-but-valid scenarios** that neither party had assessed.

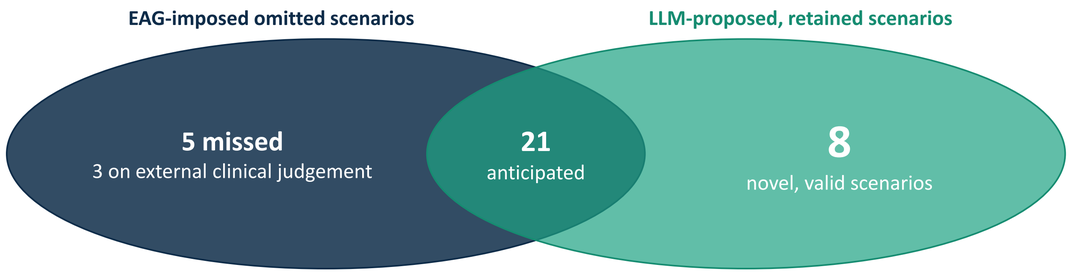
4 RESULTS

Of 71 LLM-proposed scenarios, 52 were judged valid (precision 73%). The agent anticipated **21 of the 26 EAG-imposed omitted scenarios** (recall 81%) and missed 5, of which 3 rested on **external clinical judgement**.

21/26 EAG-imposed omitted scenarios anticipated (81% recall) across 7 appraisals.

8 valid new scenarios, in 5 of 7 appraisals, that neither company nor EAG assessed.

0 unsupported scenarios retained; 6 of 71 proposals (8%) were removed at review.



21 of 26 EAG analyses were flagged from the submitted model alone (schematic, not to scale). The 5 misses began outside the submission, such as a clinician's view that a comparator does not reflect current practice.

Uncertainty domain	Where it sits in the model	Recall
Survival extrapolation	Selected distribution	**8/8 (100%)**
Time horizon	Horizon setting	**3/3 (100%)**
Treatment-effect waning	Waning rule	**4/5 (80%)**
Utilities	Utility source	**4/4 (100%)**
Treatment duration	Duration assumption	**2/3 (67%)**
External clinical judgement	Outside the submitted material	**0/3 (0%)**

Recall was 91% (21/23) where the assumption is an explicit model input and 0% (0/3) where it rested on external clinical judgement, the one domain with no input to vary.

5 DISCUSSION

The agent finds **what the model contains**. It caught 21 of the 23 scenarios tied to a model input, but none of the 3 that came from clinical opinion outside the submission.

- **Why it matters:** EAGs raise these gaps after submission, when time to respond is short; a pre-submission check moves that work earlier.
- **Where to focus:** extrapolation and waning vary most between appraisals [5], so they gain most from the check.
- **How it fits:** TECH-VER and AdViSHE [6,7] check that a model works; this predicts which assumptions reviewers will challenge.
- **Limitations:** 7 NICE oncology appraisals only; benchmark limited to documented EAG analyses; effect on ICERs not assessed.

EAG, evidence assessment group; HTA, health technology assessment; ICER, incremental cost-effectiveness ratio; LLM, large language model; NICE, National Institute for Health and Care Excellence.

6 CONCLUSION

Across 7 NICE oncology appraisals, a retrieval-grounded LLM anticipated **21 of 26 EAG-imposed omitted scenarios** from the submitted model alone and added 8 valid scenarios that neither company nor EAG had tested.

- **Earlier scrutiny:** 4 in 5 EAG challenges were visible before submission
- **Stronger validation:** all 52 retained scenarios tied to a named model input
- **Next steps:** clinical input for judgement-based gaps; other HTA bodies

LLM estimation is an assistive layer: senior health economists keep prioritisation and interpretation, as NICE expects of AI use.[8]

REFERENCES
1. NICE. Health technology evaluations: the manual (PMG36). 2022.
2. Carroll C, et al. Value Health. 2017;20.
3. Srivastava T, Hriday K. ISPOR Europe 2022, poster MSR114.
4. Lewis P, et al. Adv Neural Inf Process Syst. 2020;33.
5. Trigg LA, et al. Value Health. 2024;27.
6. Buyukkaramikli NC, et al. PharmacoEconomics. 2019;37.
7. Vemer P, et al. PharmacoEconomics. 2016;34.
8. NICE. Use of AI in evidence generation (ECD11). 2024.

Financial disclosure: Authors are employees of ConnectHEOR Limited. No external funding was received to conduct this research and no conflict of interest to declare.

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