From Proof-of-Concept to Product: A Maturity Framework for the PoC-to-Production Journey of AI Tools in HEOR

MSR109

Hanan Irfan[1], **Tushar Srivastava**[1], Kunal Hriday[2]

[1]ConnectHEOR, London, UK; [2]ConnectHEOR, Delhi, India.
Email: tushar.srivastava@connectheor.com

> **In HEOR, a working demonstration is not a working product: AI tools become reliable only when they pass explicit gates for reliability, reproducibility and ownership, with accountable people signing off at each stage.**

1 BACKGROUND

Health economics and outcomes research (HEOR) teams increasingly build their own artificial intelligence (AI) tools for evidence synthesis, economic modelling and technical reporting.[1] A proof-of-concept (PoC) can be ready in days and look convincing, but it runs on clean inputs and controlled runs

- These outputs may feed health technology assessment (HTA) submissions, where agencies expect transparent, reproducible methods and hold named people, not tools, accountable.[2,3]
- The PoC is a small fraction of a working system and needs rigorous monitoring and validation to match the complexity of real HEOR use.[4]

The gap: no framework defines when an in-house AI tool is ready to be relied on for HTA-grade work, what it must prove beyond a one-off demonstration, or who signs off that decision.

2 OBJECTIVES

This work set out to make the PoC-to-production journey **explicit and repeatable** for AI tools in HEOR, through three aims.

1. Develop a reusable maturity framework that maps the PoC-to-production journey through three gated stages.
2. Define what each stage must prove, so that capability shown in a demonstration is not mistaken for viability.
3. Specify the engineering-readiness dimensions that govern each transition, in terms portable across very different tools.

Each aim moves the framework from structure, to proof, to explicit criteria.

3 METHODS

We ran a **structured framework-development exercise**, drawing lessons from six AI-enabled HEOR tools at PoC stage meant to be used for evidence synthesis, economic modelling and technical reporting. We analysed each tool on its own; everyone had been built first as PoC and eventually moved towards scalable repeated use.

1. Review built tools → Evidence synthesis, economic modelling and technical reporting

2. Contrast conditions → PoC success versus dependable repeated use

3. Group break points → Scale-up failures become product-viability dimensions

4. Arrange stages → Proof-of-concept, validated tool, governed production system

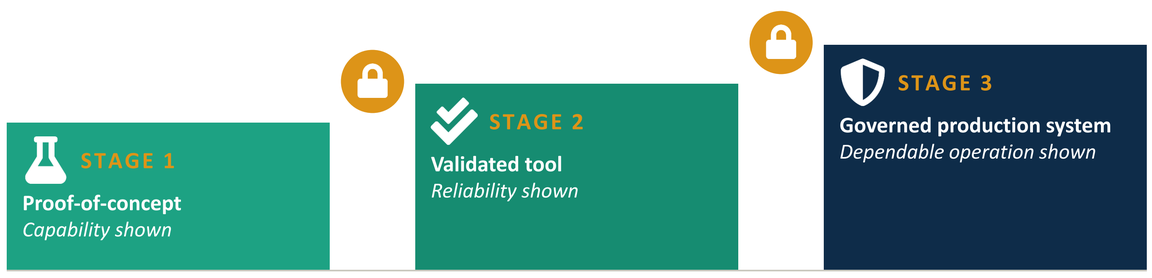
5. Define gates → Engineering-readiness criteria for each transition

Criteria were kept conceptual, so the framework stays portable across tools, at the price of requiring judgement at each gate.

- **Comparison:** for each tool, the conditions under which its PoC succeeded were contrasted with those required for dependable repeated use.
- A demonstration runs once on material the demonstrator selects, with an expert watching; repeated use runs many times on whatever arrives, often unsupervised.
- Points where tools broke during scale-up were grouped into product-viability dimensions, as a failure in one tool is a property of that class of tool.
- **Outcome: Scale-up** criteria covered representative inputs, robustness to messy inputs, edge cases and drift [5], reproducibility, observability, integration and ownership.

4 RESULTS

Product viability from PoC is earned through **three gated stages**, not granted by a single demonstration. Each stage proves a different property (capability, then reliability, then dependable operation), and no tool can skip a stage by excelling at the one before.



Stage 1 is real, but a curated one-off demo says nothing about the next input.

Stage	Gate to advance
Stage 1	Curated, single-run success is rejected as evidence of readiness. It is mere technical possibility.
Stage 2	Reliable performance on representative, unseen inputs, predictable edge-case behaviour and reproducible outputs across runs and versions.
Stage 3	Dependable operation in real workflows: monitored for drift, maintainable, observable and owned, with human oversight where outputs feed downstream analysis.

Stage 2 tests the tool; Stage 3 tests the system around it.

- **Reliability:** the tool is tested on inputs the builder did not choose, including messy cases such as scanned PDFs, and re-run with fixed prompts and model versions.
- **Human in the loop:** experts double-review a sample of outputs against a reference standard, resolve any disagreements and sign off before results are used.
- **Audit trail:** every output is logged with its input, prompt, model version and reviewer, so any result can be traced and re-run.
- **Safe operation:** a named owner checks the tool for drift against a fixed benchmark, re-tests it after each update and decides when to retire it.
- **The gate decides, not the demo:** at every transition, the question is whether the stated conditions hold, not how promising the tool looks.

5 DISCUSSION

A PoC runs under conditions the builder controls, which **suppress the very properties** that decide viability. A product demonstration is evidence about a different question.

- The suggested framework sits alongside production-readiness rubrics [6], staged maturity models for machine learning in healthcare [7] and lifecycle risk frameworks.[8]
- Most of the work sits after the demonstration, so naming it early changes how HEOR tool-building is planned and resourced.
- Because progression is a gate rather than a score, each transition decision is explicit and can be recorded and revisited.
- **Limitations:** it draws on one organisation's tools, has not been evaluated prospectively, and industry-wide consistency is untested.

AI, artificial intelligence; HEOR, health economics and outcomes research; HTA, health technology assessment; NICE, National Institute for Health and Care Excellence; NIST, National Institute of Standards and Technology; PoC, proof-of-concept.

6 CONCLUSION

- The framework separates **three questions**: can the tool do the task, does it do so reliably on inputs it has not seen, and can it be relied on in routine work under a named owner.
- A staged, gated progression turns ad hoc tool-building into a **repeatable engineering discipline**.
- It offers a proposed **shared standard** for when a tool is ready to rely on; it is drawn from one organisation's tools and has not been tested elsewhere.
- Most effort comes after the demonstration, so **resource it from the outset**.
- Next: **prospective, multi-organisation evaluation** of the framework.

REFERENCES

1. Fleurence RL, et al. Value Health. 2025;28.
2. NICE. Position statement ECD11. 2024.
3. Flemyng E, et al. JBI Evid Synth. 2025;23.
4. Sculley D, et al. Adv Neural Inf Process Syst. 2015;28.
5. Finlayson SG, et al. N Engl J Med. 2021;385.
6. Breck E, et al. IEEE Big Data. 2017.
7. Li Y, et al. J Med Internet Res. 2025;27.
8. Tabassi E. NIST AI 100-1. 2023.

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