NICE National Institute for Health and Care Excellence

Developing best practice for generative AI in health economic evaluation: evidence from a systematic review and stakeholder engagement

Steve Sharp*[1], Kusal Lokuge*[1], Jamie Elvidge[1]

* Equal first author contribution

[1] National Institute for Health and Care Excellence

Background

Generative AI (GenAI) has emerged in the current decade as a paradigm shifting technology with potential to transform the process of health economic evaluation (HEE), a resource-intensive element of health technology assessment (HTA). This NICE HTA innovation Laboratory (HTA Lab) project aimed to develop a set of early best practice principles for the use of GenAI in HEE, informed by a systematic review, exploratory use cases, and stakeholder engagement through a series of workshops.

Methods

To inform this work, the HTA Lab conducted:

- A systematic literature review of GenAI use in economic modelling[1], (updated Oct 2025). The review focused on early GenAI use cases with studies selected using predefined criteria.
- Collaborative projects with Value Analytics Labs and Estima Scientific to test GenAI tools on 3 different economic modelling use cases: **country adaptation**, **validation** and **replication-construction**. Each use case followed structured workflows with human oversight for selecting parameters, refining prompts and quality assurance.
- Three workshops (Mar–Jul 2025) with stakeholders from HTA agencies, external assessment groups, industry, and ISPOR to explore GenAI use in HEE. Each workshop had distinct aims, ranging from exploring applications and barriers to reviewing use-case testing and refining best practice principles.

A set of early best practice principles was developed to supplement the NICE AI position statement[2] by triangulating findings from the above activities.

Results

Updated systematic review

We identified 38 eligible studies (9 journal publications; 29 conference abstracts). These included 22 primary case studies, 2 qualitative studies, 8 reviews, and 6 expert opinion pieces. The evidence base is rapidly evolving (**13 new studies** identified in a **6-month update**).

The primary case studies covered **model: conceptualisation** (3), **replication** (7), **adaptation** (3), **construction** (2), **updating** (1), **optimisation** (2), **reporting and dissemination** (4) and **final appraisal extraction** (1).

Despite limited evidence, **model construction, updating, optimisation** and **adaptation** use cases showed potential for future use in practice. **Model replication,** a gateway use case in validating GenAI tools, was commonly reported.

Further research across use cases (including **model validation** which had no studies) is needed to advance knowledge and support application.

Stakeholder workshops

The workshops engaged over 60 stakeholders with economics and data science expertise.

Workshop 1 identified high perceived value in automating tasks such as data extraction and model replication, but flagged concerns around accuracy, privacy, and oversight.

Workshop 2 reviewed use-case testing on adapting and validating economic models, and initiated best practice discussions, highlighting reproducibility and human oversight.

Workshop 3 focused on testing model construction and refining best practice principles. Across workshops, stakeholders emphasised transparency, accountability, and tailored training. Feedback informed a draft framework of 19 best practice principles to support responsible GenAI use in health economic modelling.

Collaborative use cases

Model adaptation: GenAI adapted a US Alzheimer's model to NICE's reference case, automatically identifying UK-specific data sources and updating code. Human input guided parameter selection and interpretation. Limitations included reliance on expert oversight and lack of comparable UK models for accuracy checks.

Model validation: GenAI was tested against NICE's asthma model QA process, successfully replicating 12/14 checks. Discrepancies were traced to human errors. This highlights its potential to reduce the burden of routine QA tasks while maintaining accuracy.

Model replication-construction: GenAI replicated and partially constructed the NICE Safe prescribing model using R scripting. Despite some deviations, conclusions were consistent. This may support faster model development and easier maintenance.

Best practice principles

Purpose, scope and responsibilities

Assess opportunity cost and demonstrate value
Include human oversight
Maintain accountability

Transparency, reproducibility and explainability

Provide clear documentation
Enable replication and auditability
Embed ethical data use

Technical standards and interoperability

Apply validation and quality assurance
Evaluate system compatibility

Ethics, bias and regulatory compliance

Assess and mitigate bias
Comply with regulatory and legal requirements
Safeguard privacy and security

Education, training and awareness

Provide tailored training
Support multidisciplinary collaboration
Communicate findings clearly

General principles

Support continuous improvement
Engage stakeholders at all stages
Promote responsible use

References

1. Sharp S, Lokuge K, Elvidge J et al. (2025) Systematic literature review of the use of generative AI in health economic evaluation. Available at https://www.medrxiv.org/content/10.1101/2025.04.25.25326412v2 Accessed on 15/10/25
2. NICE (2024) Use of AI in evidence generation: NICE position statement. https://www.nice.org.uk/position-statements/use-of-ai-in-evidence-generation-nice-position-statement Accessed on 15/10/25

Steve Sharp
Scientific Adviser
National Institute for Health and Care Excellence
Email: steve.sharp@nice.org.uk

Kusal Lokuge
Senior Analyst
National Institute for Health and Care Excellence
Email: kusal.lokuge@nice.org.uk

nice.org.uk