Prioritising Pre-Submission Investment: Associations Between Health Economic Model Characteristics and NICE TA Critique Burden and Recommendation Outcomes

HTA55

Gareth Wynne-McHardy[1], MBChB, Amanda Hansson Hedblom[1], MSc, Anna McCormick[1], DPhil, Gerdi Strydom[1], MBA, PharmD

[1]Valid Insight (Part of the Bioscript Group), Macclesfield, Cheshire, United Kingdom

Objectives

To categorise and quantify health economic model critique in NICE appraisals, and to analyse the direction and magnitude of any association between critique categories and NICE recommendations.

Methods

We reviewed the committee papers for 96 NICE technology appraisals published between 01/01/2025 and 07/05/26 (TA1027 to TA1153).[1] One committee paper was analysed for each appraisal. The most recent paper was analysed when papers from more than one committee meeting were available; otherwise, the only available paper was analysed.

A structured and auditable protocol was implemented to extract ten predefined binary critique domains from each appraisal. A critique domain is an area of the submission in which the EAG can raise a concern. A domain scored 1 when the concern was raised, and 0 otherwise.

The ten critique domains were:

- **Comparator relevance:** whether comparators reflect NHS practice or are appropriate for decision-making
- **Clinical evidence:** quality, maturity, applicability or robustness of trial or observational data
- **Extrapolation:** methods and plausibility of long-term predictions beyond observed data
- **Treatment effect assumptions:** how effects are applied or kept over time
- **Utility:** derivation, selection or application of quality-of-life values
- **Costing:** estimation, inclusion or application of cost inputs
- **Model structure:** health states, transitions or conceptual framework
- **Comparative efficacy estimation:** methods to estimate comparative efficacy across studies (ITCs)
- **Uncertainty analysis:** whether uncertainty is explored and presented adequately
- **External validity:** agreement of model outputs with external data, clinical expectations or real-world evidence

The **core issues score** counts how many of the critique domains the EAG raised in an appraisal. ***Note:*** *only critique domains with incidence ≤80% across all appraisals were considered, given that critiques found almost always cannot show which appraisals were criticised more.*

ITC use and surrogate endpoint use were also recorded as model structural characteristics. These were analysed independently of critique domains.

The recommendation outcome was sourced from the final NICE guidance: 41 recommended, 46 recommended with restrictions (including managed access and split decisions), 5 not recommended and 4 terminated (no decision). Recommendation tests compared recommended without restrictions (41) against all other appraisals with a decision (51).

Statistical analysis

The Mann-Whitney U test compared the core issues score of appraisals with and without each item. Its effect size, the rank-biserial r, runs from 0 (no difference) to 1 (every appraisal with the item has a higher score than every appraisal without it).

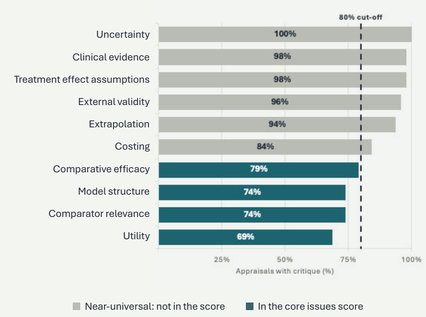
Each domain was removed from the score before its own test, and for ITC use, the comparative efficacy critique was removed.

The OR divides the odds of a recommendation with the critique by the odds without it, meaning an OR below 1 indicates fewer recommendations. The q-value is the p-value after the Benjamini-Hochberg false discovery rate correction.[2] Results with q<0.05 and an effect at or above the pre-set minimum (r≥0.20) are marked *. ORs use Fisher's exact p without correction (* when p<0.05) and a log-method 95% CI.

Sensitivity analyses were conducted with no cost comparisons and with STAs only.

Results

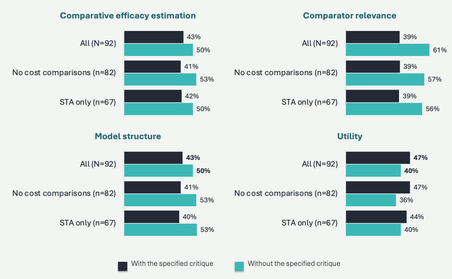
Critique domain frequency per appraisal



Six domains were excluded from the core issues score because they were above the 80% cut-off.

Six domains had critique in 84% to 100% of appraisals, and 74 of 96 (77%) had all six of these near-ubiquitous critiques. Uncertainty critique was in every appraisal. The other four domains (69% to 79%) make up the core issues score.

Recommendation rate by core critique domain



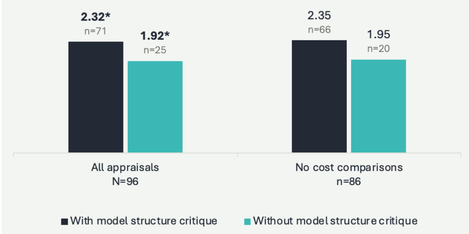
Sets: with a decision (N=92); no cost comparisons (n=82); STA only (n=67)

Bars: % of appraisals recommended without restrictions

With final NICE outcomes, no core critique domain was independently significantly associated with recommendation without restrictions (all p≥0.091).

p values: Fisher's exact test, with against without the critique (not corrected for multiple tests). Recommended means recommended without restrictions. The six near-universal domains could not be tested one by one, because only 0 to 14 of the 92 appraisals had no critique in them. Their sum had no link with the decision (p=0.876).

Model structure critique is associated with a higher critique burden (not corrected for multiple tests)

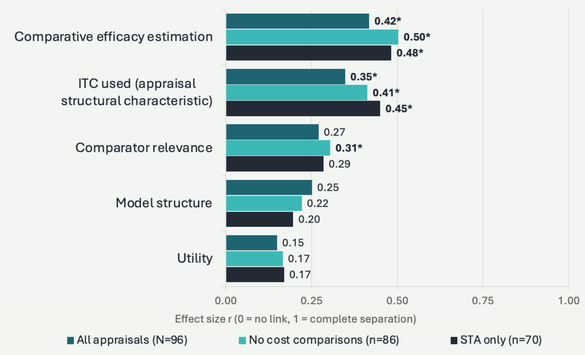


Bars: mean core issues score without model structure (0 to 3).
* Mann-Whitney U p<0.05, not corrected for multiple tests (q=0.056).

Appraisals with model structure critique had a higher core issues score (mean 2.32 against 1.92, with model structure removed from the score, r=0.25, p=0.045, N=96). The result was not significant after correction for multiple tests (q=0.056), no cost comparisons (r=0.22, p=0.103) or in STAs only (r=0.20, p=0.171). 71 of 96 appraisals (74%) had this critique. With final NICE outcomes, it was not associated with recommendation without restrictions (43% against 50%, OR 0.74, 95% CI 0.29 to 1.89, p=0.635, N=92).

The critique was as common in partitioned survival models (30 of 41, 73%) as in state transition models (31 of 40, 78%).

Core critique domain or appraisal structural characteristic correlation with the core issues score

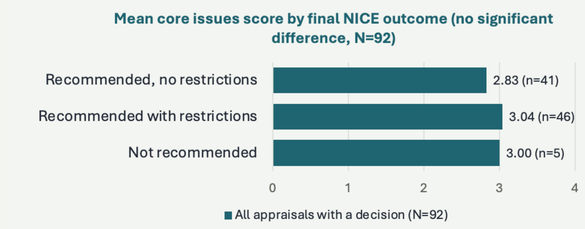


* significant: q<0.05 and r≥0.20 (five tests per set).

Each bar indicates whether appraisals with an item tend to have higher critique scores than appraisals without it.

The item's own critique is removed from the score, so it cannot inflate its own result. Example: r=0.50 indicates that in one appraisal with the item and one without, the one with the item has the higher score 75% of the time (ties count as half). In all 96 appraisals (70 STA, 10 cost comparisons, 16 others), comparative efficacy critique and ITC use were significant. In the absence of cost comparisons, comparator relevance was also significant. Of the 10 cost comparisons, 8 used an ITC and 9 had comparative efficacy critique, but only 2 had utility critique, against 64 of the other 86 appraisals. Surrogate endpoint use was not associated with critique burden (r=0.00).

ITC use is associated with greater critique burden



Sets: all appraisals (N=96); with a decision (N=92)

Appraisals with an ITC had a higher core issues score: median 4 against 2.

Without comparative efficacy critique, the median core issues scores were 3 and 2. ITC use was associated with the core issues score in all appraisals (r=0.35, q=0.007), no cost comparisons (r=0.41, q=0.001) and in STAs only (r=0.45, q=0.005). Comparative efficacy critique stayed under the 80% cut-off (79.2%, 76 of 96) only because not every appraisal used an ITC. If all appraisals had used an ITC, it would have exceeded the 80% cut-off.

The core issues score did not differ between appraisals recommended without restrictions and other appraisals with a decision (r=0.05, p=0.656, not corrected for multiple tests). Mean scores: 2.83 recommended, 3.04 recommended with restrictions, 3.00 not recommended.

Updates to the analysis following the abstract

18 significant associations between appraisal characteristics and critique domains were noted in the abstract. These were produced in an earlier version of the analysis in which the core issues score also counted ITC use and surrogate endpoint use (core issue score 0 to 6). During this prior analysis, each item also stayed in the score during its own test, introducing circularity.

The 18 associations were of four types:

- **9 linked** a critique domain or appraisal structural characteristic with critique burden (the core issues score or the total of all ten domains), for example, comparative efficacy critique (r=0.956) and ITC use (r=0.763)
- **3 linked** two critique domains or appraisal structural characteristics, for example, ITC use and comparative efficacy critique (V=0.539)[3]
- **5 linked** appraisal type, model type or therapy area with each other or with utility critique, for example, appraisal type and model type (V=0.545)
- **1 linked** utility critique with the recommendation (V=0.663, n=22)

The current analysis uses one committee paper per appraisal. The most recent committee paper was used when there were several per appraisal, otherwise the only paper available was used (N=96). The score counts only the four core critique domains (0 to 4). ITC use and surrogate endpoint use are tested separately, and each domain is removed from the score before its own test, so effect sizes were smaller. A correlation of 0.42 was found for comparative efficacy critique and of 0.35 for ITC use. Outcomes now come from the final NICE guidance.

Conclusions

Methodological critique in NICE appraisals from January 2025 to May 2026 reflects both the underlying model complexity and the appraisal context. ITC use and comparative efficacy critique are associated with a higher overall critique burden, as is model structure critique (not after correction for multiple tests). The relationship between critique burden and final NICE recommendation was not found to be significant, an issue which a small sample size of Not Recommended submissions is likely to have contributed to. These findings identify comparative efficacy estimations and model structure as drivers of critique burden. Organisations may benefit from:

- mapping the NICE precedent to understand how similar technologies and therapy areas fared, and which model structures the committee accepted;
- using clinical opinion to support the model structure and the choice of comparators;
- using a robust ITC method with defensible assumptions.

To further the analyses conducted here, future work could extend the date range of the included appraisals and use papers from every committee meeting, to increase robustness. Additionally, the model critique domains evaluated may be further refined to increase granularity.

Limitations

The study found associations only. Data came from NICE committee papers for appraisals published from 01/01/2025 to 07/05/2026, with one paper per appraisal (the most recent). Critique came from the most recent committee paper at the time point of retrieval, but the outcome came from the final guidance, and only 5 appraisals were not recommended.

References

1. NICE. https://www.nice.org.uk/. Technology appraisals TA1027 to TA1153: papers, 2025-2026. Full list of 96 appraisals in the HTA55 supplement.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.
3. Bergsma W. A bias-correction for Cramér's V and Tschuprow's T. J Korean Stat Soc. 2013;42(3):323-328. doi:10.1016/j.jkss.2012.10.002.

Abbreviations

AI: Artificial intelligence; **CI:** confidence interval; **EAG:** external assessment group; **ITC:** indirect treatment comparison; **NHS:** National Health Service; **NICE:** National Institute for Health and Care Excellence; **OR:** odds ratio; **q:** p value after Benjamini-Hochberg false discovery rate correction; **r:** rank-biserial correlation; **STA:** single technology appraisal; **TA:** technology appraisal; **V:** bias-corrected Cramér's V.

Disclosures

Funding: No financial disclosures.
AI statement: Extracted data were analysed with the support of AI. AI was also utilised for poster editing.

VALID INSIGHT®

PART OF THE BIOSCRIPT GROUP

validinsight.com

Valid Insight, Greenway House, Larkwood Way, Tytherington Business Park, Tytherington, Macclesfield, Cheshire, SK10 2XR

gareth.wynne-mchardy
@bioscriptgroup.com

ISPOR, 09 11 2026©
2026 Valid Insight Ltd