National Institute for Health and Care Excellence

Advanced breast cancer: diagnosis and management

NICE guideline CG81

Methods

March 2026

Draft for consultation

Disclaimer

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and/or their carer or guardian.

Local commissioners and/or providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN:

Contents

Development of the guideline

Remit

The methods outlined in this document relate to the update of Advanced breast cancer: diagnosis and management (CG81) published in 2026.

This methods document does not cover recommendations published as part of the original guideline in 2009. Methods for these recommendations are available as part of the original evidence review.

To see “What this guideline covers” and “What this guideline does not cover”, see the guideline scope for Advanced breast cancer: diagnosis and management.

Methods

This evidence review was developed using the methods and process described in Developing NICE guidelines: the manual.

Declarations of interest were recorded according to NICE's conflicts of interest policy.

Developing the review questions and outcomes

The 3 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline development team and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- Population, Intervention, Comparator and Outcome and Study type (PICOS) for reviews of interventions
- Population, index test(s), reference standard and outcome for reviews of diagnostic and predictive accuracy

Full literature searches, critical appraisals and evidence reviews were completed for all review questions.

Reviewing research evidence

Review protocols

Review protocols were developed with the guideline committee to outline the inclusion and exclusion criteria used to select studies for each evidence review.

Searching for evidence

Evidence was searched for each review question using the methods specified in Developing NICE guidelines: the manual.

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example, previous versions of the guideline or studies identified by committee members) were uploaded into EPPI reviewer software (version 5) and de-duplicated. Titles and abstracts were assessed for possible inclusion using the criteria specified in the review protocol. At least 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

Priority screening was not used for any reviews in this guideline and therefore the full database was screened for each review.

The full text of potentially eligible studies was retrieved and assessed according to the criteria specified in the review protocol. A standardised form was used to extract data from included studies. There were no instances that it was thought necessary to contact study investigators for missing data.

Incorporating published evidence syntheses

If published evidence syntheses were identified sufficiently early in the review process (for example, from the surveillance review or early in the database search), they were considered for use as the primary source of data, rather than extracting information from primary studies. Syntheses considered for inclusion in this way were quality assessed to assess their suitability using the appropriate checklist, as outlined in Table 1. Note that this quality assessment was solely used to assess the quality of

the synthesis in order to decide whether it could be used as a source of data, as outlined in Table 2, not the quality of evidence contained within it, which was assessed in the usual way as outlined in the section on 'Appraising the quality of evidence'.

Table 1: Checklists for published evidence syntheses

Type of synthesis	Checklist for quality appraisal
Systematic review of quantitative evidence	ROBIS

Each published evidence synthesis was classified into one of the following three groups:

- High quality – It is unlikely that additional relevant and important data would be identified from primary studies compared to that reported in the review, and unlikely that any relevant and important studies have been missed by the review.
- Moderate quality – It is possible that additional relevant and important data would be identified from primary studies compared to that reported in the review, but unlikely that any relevant and important studies have been missed by the review.
- Low quality – It is possible that relevant and important studies have been missed by the review.

Each published evidence synthesis was also classified into one of three groups for its applicability as a source of data, based on how closely the review matches the specified review protocol in the guideline. Studies were rated as follows:

- Fully applicable – The identified review fully covers the review protocol in the guideline.
- Partially applicable – The identified review fully covers a discrete subsection of the review protocol in the guideline (for example, some of the factors in the protocol only).
- Not applicable – The identified review, despite including studies relevant to the review question, does not fully cover any discrete subsection of the review protocol in the guideline.

The way that a published evidence synthesis was used in the evidence review depended on its quality and applicability, as defined in Table 2. When published evidence syntheses were used as a source of primary data, data from these were quality assessed and presented in GRADE/CERQual tables in the same way as if data had been extracted from primary studies. In questions where data was extracted from both systematic reviews and primary studies, these were checked to ensure none of the data had been double counted through this process.

Table 2: Criteria for using published evidence syntheses as a source of data

Quality	Applicability	Use of published evidence synthesis
High	Fully applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search or data analysis. Searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted.
High	Partially applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search and data analysis for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted. For other sections not covered by the evidence synthesis, searches were undertaken as normal.
Moderate	Fully applicable	Details of included studies were used instead of undertaking a new literature search. Full-text papers of included studies were still retrieved for the purposes of data analysis. Searches were only done to cover the period of time since the search date of the review.
Moderate	Partially applicable	Details of included studies were used instead of undertaking a new literature search for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. For other sections not covered by the evidence synthesis, searches were undertaken as normal.

Methods of combining evidence

Data synthesis for intervention studies

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome. When there were 2 treatment alternatives, pairwise meta-analysis was used to compare interventions.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane RevMan (web version). A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event, and a pooled incidence rate ratio was calculated for dichotomous outcomes reporting total numbers of events. Both relative and absolute risks were presented, with absolute risks calculated by applying the relative risk to the risk in the comparator arm of the meta-analysis (calculated as the total number events in the comparator arms of studies in the meta-analysis divided by the total number of participants in the comparator arms of studies in the meta-analysis). If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into the web version of Cochrane Review Manager.

No evidence presented as continuous outcomes was identified as part of this update.

Random effects models were fitted when significant between-study heterogeneity in methodology, population, intervention or comparator was identified by the reviewer in advance of data analysis. This decision was made and recorded before any data analysis was undertaken. For all other syntheses, fixed- and random-effects models were fitted, with the presented analysis dependent on the degree of heterogeneity in the assembled evidence. Fixed-effects models were the preferred choice to report, but in situations where the assumption of a shared mean for fixed-effects model were clearly not met, even after appropriate pre-specified subgroup analyses were conducted, random-effects results are presented. Fixed-effects models were deemed to be inappropriate if there was significant statistical heterogeneity in the meta-analysis, defined as I2 more than or equal to 40%.

Results from subgroup analyses (rather than the full analysis) were assessed using GRADE only when statistically significant subgroup differences were identified ($p < 0.05$). In these cases, where the full analysis displayed significant statistical heterogeneity, but the pre-specified subgroup analyses were less heterogeneous (with $I2 < 40\%$), the results from these subgroups were reported using fixed effects models.

Data synthesis for diagnostic accuracy data

In this guideline, diagnostic test accuracy (DTA) data are classified as any data in which a feature – be it a symptom, a risk factor, a test result or the output of some algorithm that combines many such features – is observed in some people who have the condition of interest at the time of the test and some people who do not. Such data either explicitly provide, or can be manipulated to generate, a 2x2 classification of true positives and false negatives (in people who, according to the reference standard, truly have the condition) and false positives and true negatives (in people who, according to the reference standard, do not).

Where multiple observer interpretations of the same images were reported in a study, all interpretations were included.

The 'raw' 2x2 data can be summarised in a variety of ways. Those that were used for decision making in this guideline were as follows:

- **Sensitivity** is the probability that the feature will be positive in a person with the condition.
    - sensitivity = TP/(TP+FN)
- **Specificity** is the probability that the feature will be negative in a person without the condition.
    - specificity = TN/(FP+TN)
- **Positive likelihood ratios** describe how many times more likely positive features are in people with the condition compared to people without the condition. Values greater than 1 indicate that a positive result makes the condition more likely.
    - LR^+ = (TP/[TP+FN])/(FP/[FP+TN])

- **Negative likelihood ratios** describe how many times less likely negative features are in people with the condition compared to people without the condition. Values less than 1 indicate that a negative result makes the condition less likely.
    - LR^- = (FN/[TP+FN])/(TN/[FP+TN])

Where studies did not provide 2x2 data, a Microsoft excel calculator was used to convert diagnostic accuracy measures (sensitivity, specificity, diagnostic accuracy, predictive values or likelihood ratios) to the raw 2x2 data.

Meta-analysis of diagnostic accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.0 (Deeks et al. 2023).

Where three or more studies were available for all included strata, a bivariate model was fitted using the using the metaDTA app (https://crsu.shinyapps.io/MetaDTA/), which accounts for the correlations between positive and negative likelihood ratios, and between sensitivities and specificities (alternative approaches were required where convergence wasn't achieved).

MetaDTA is based on the glmer function in the lme-4 package in R which runs a bivariate meta-analysis using a generalised linear mixed model (GLMM), as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.

Where an analysis only contained 2 studies, regardless of whether the data converged in MetaDTA, meta-analysis was not completed because a minimum of 3 studies are needed to estimate the parameters needed for bivariate meta-analysis. Sensitivity and specificity forest plots were obtained from Cochrane RevMan (version 8.16.0), and likelihood ratios were obtained using a Microsoft Excel calculator. The range of point estimates were presented in the GRADE table as a pooled effect estimate could not be obtained.

Random-effects models (der Simonian and Laird) were fitted for all syntheses, as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.0 (Deeks et al. 2023).

Appraising the quality of evidence

Intervention studies (relative effect estimates)

RCTs were quality assessed using the Cochrane Risk of Bias Tool. Risk of bias for single studies were conducted once for objective outcomes, once for subjective outcomes, and once for adverse events.

Where systematic reviews were used as a source of evidence for RCTs but they use the Cochrane Risk of Bias Tool 1 for risk of bias, the judgements were taken from that review and converted to Cochrane risk of bias Tool 2 judgements so that all RCTs were assessed in the same way. Descriptions of the approach taken are written in the methods specific to the review.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, intervention, comparator and/or outcomes in the study and how directly these variables could address the specified review question. Studies were rated as follows:

- Direct – No important deviations from the protocol in population, intervention, comparator and/or outcomes.
- Partially indirect – Important deviations from the protocol in one of the following areas: population, intervention, comparator and/or outcomes.
- Indirect – Important deviations from the protocol in at least two of the following areas: population, intervention, comparator and/or outcomes.

Minimally important differences (MIDs) and clinical decision thresholds

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal clinically important difference thresholds relevant to this guideline that might aid the committee in identifying clinical decision thresholds for the purpose of GRADE. Identified MIDs were assessed to ensure they had been developed and validated in a methodologically rigorous way, and were applicable to the populations, interventions and outcomes specified in this guideline. In addition, the Guideline Committee were asked to prospectively specify any outcomes where they felt a consensus clinical decision threshold could be defined

from their experience. In particular, any questions looking to evaluate non-inferiority (that one treatment is not meaningfully worse than another) required a clinical decision threshold to be defined to act as a non-inferiority margin.

Clinical decision thresholds were used to assess imprecision using GRADE and aid interpretation of the size of effects for different outcomes. Clinical decision thresholds that were used in the guideline are given in Table 3.

Table 3: Identified clinical decision thresholds

Outcome	Clinical decision threshold	Source
Quality of life FACT-G total	3 to 7 points	Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004 Sep;57(9):898-910. doi: 10.1016/j.jclinepi.2004.01.012. PMID: 15504633.
Quality of life FACT-B total	7 to 8 points	Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004 Sep;57(9):898-910. doi: 10.1016/j.jclinepi.2004.01.012. PMID: 15504633.
Quality of life TOI (trial outcome index) of FACT-B	5 to 6 points	Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004 Sep;57(9):898-910. doi: 10.1016/j.jclinepi.2004.01.012. PMID: 15504633.
Quality of life BCS of FACT-B	2 to 3 points	Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004 Sep;57(9):898-910. doi: 10.1016/j.jclinepi.2004.01.012. PMID: 15504633.
Quality of life EORTC QLQ-C30	11 points for improvement Minus 8 points for deterioration	Kawahara, T., Taira, N., Shiroiwa, T. et al. Minimal important differences of EORTC QLQ-C30 for metastatic breast cancer patients: Results from a randomized clinical trial. Qual Life Res 31, 1829–1836 (2022). https://doi.org/10.1007/s11136-021-03074-y
Quality of life	1 point	Den Oudsten, B.L., Zijlstra, W.P. & De Vries, J. The minimal clinical important difference in the World Health

WHOQOL-100		Organization Quality of Life instrument—100. Support Care Cancer 21, 1295–1301 (2013). https://doi.org/10.1007/s00520-012-1664-8

GRADE for intervention studies analysed using pairwise analysis

GRADE was used to assess the quality of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials was initially rated as high quality. The quality of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 4. These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 4: Rationale for downgrading quality of evidence for intervention studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	• Not serious (don't downgrade): less than 50% overall weighting some concerns/high risk of bias • Serious (downgrade 1 level): more than 50% some concerns/high risk of bias • Very serious (downgrade 2 levels): more than 50% high risk of bias.
Indirectness	• Not serious (don't downgrade): less than 50% of overall weighting partially direct or indirect. • Serious (downgrade 1 level): more than 50% of overall weighting partially direct or indirect. • Very serious (downgrade 2 levels): more than 50% of overall weighting indirect
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. • Not serious (don't downgrade) I2 = less than 40%; • Serious (downgrade 1 level) I2 = 40-60%; • Very serious (downgrade 2 levels) I2 = more than 60%.
Imprecision	If an MID other than the line of no effect was defined for the outcome, the outcome was downgraded once if the 95% confidence interval for the effect size crossed one line of the MID, and twice if it crossed both lines of the MID. If the line of no effect was defined as an MID for the outcome, it was downgraded once if the 95% confidence interval for the effect size crossed

	the line of no effect (i.e. the outcome was not statistically significant), and twice if the sample size of the study was sufficiently small that it is not plausible any realistic effect size could have been detected. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to clinically equivalent scenarios.
Publication bias	Where 10 or more studies were included as part of a single meta-analysis, a funnel plot was produced to graphically assess the potential for publication bias. When a funnel plot showed convincing evidence of publication bias, or the review team became aware of other evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

Diagnostic accuracy studies

Individual diagnostic accuracy studies were quality assessed using the QUADAS-2 tool. Each individual study was classified into one of the following three groups:

- Low risk of bias – The true effect size for the study is likely to be close to the estimated effect size.
- Moderate risk of bias – There is a possibility the true effect size for the study is substantially different to the estimated effect size.
- High risk of bias – It is likely the true effect size for the study is substantially different to the estimated effect size.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, index features and/or reference standard in the study and how directly these variables could address the specified review question. Studies were rated as follows:

- Direct – No important deviations from the protocol in population, index feature and/or reference standard.
- Partially indirect – Important deviations from the protocol in one of the population, index feature and/or reference standard.
- Indirect – Important deviations from the protocol in at least two of the population, index feature and/or reference standard.

- Where a meta-analysis could not be conducted because only 2 studies were included in the analysis, individual studies reporting sensitivity and specificity had equal weighting. Where ≥50% of the studies had some concerns for risk of bias, the evidence was downgraded by one level and where ≥50% of the studies had high risk of bias the evidence was downgraded by two levels.

GRADE for diagnostic accuracy evidence

Evidence from diagnostic accuracy studies was initially rated as high quality and then downgraded according to the standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 5 below.

The choice of primary outcome for decision making was determined by the committee and GRADE assessments were undertaken based on these outcomes.

In all cases, the downstream effects of diagnostic accuracy on patient-important outcomes were considered. This was done explicitly during committee deliberations and reported as part of the discussion section of the review detailing the likely consequences of true positive, true negative, false positive and false negative test results. In reviews where a decision model is being carried (for example, as part of an economic analysis), these consequences were incorporated here in addition.

Using sensitivity and specificity as the primary outcomes

GRADE assessments were only undertaken for sensitivity and specificity but results for positive and negative likelihood ratios are also presented alongside those data.

The committee were consulted to set 2 clinical decision thresholds for each measure: the value above which a test would be recommended, and a second below which a test would be considered of no clinical use. These values were used to judge imprecision (see below).

If studies could not be pooled in a meta-analysis, GRADE assessments were undertaken on the body of evidence that was considered for meta-analysis, rather than on individual estimates or median estimates.

Sensitivity and specificity were rated as high, moderate or low based on the following:

- High: Point estimate is greater than or equal to the upper clinical decision-making threshold. Upper clinical decision-making thresholds were 90% for sensitivity and 80% for specificity
- Moderate: Point estimate greater than or equal to the lower clinical decision-making threshold but lower than the upper clinical decision-making threshold.
- Low: Point estimate is less than the lower clinical decision-making threshold. Lower clinical decision-making thresholds were 70% for sensitivity and 60% for specificity.
- Sensitivity and specificity were interpreted using the percentage point estimate. For sensitivity, the value of the point estimate determines the proportion that is ruled out and for specificity, the value of the point estimate determines the proportion that are ruled in as having the condition.

These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 5: Rationale for downgrading quality of evidence for diagnostic accuracy data

GRADE criteria	Reasons for downgrading quality
Risk of bias	• Not serious (don't downgrade): less than 50% overall weighting some concerns/high risk of bias • Serious (downgrade 1 level): more than 50% some concerns/high risk of bias • Very serious (downgrade 2 levels): more than 50% high risk of bias. Weight of the study was equal to the study sample as a proportion of the total number of participants contributing to the outcome.
Indirectness	• Not serious (don't downgrade): less than 50% of overall weighting partially direct or indirect. • Serious (downgrade 1 level): more than 50% of overall weighting partially direct or indirect. • Very serious (downgrade 2 levels): more than 50% of overall weighting indirect Weight of the study was equal to the study sample as a proportion of the total number of participants contributing to the outcome.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across

GRADE criteria	Reasons for downgrading quality
	studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. Ranges of test performance are determined by the 2 clinical decision thresholds set for each measure (sensitivity and specificity. • Not serious (don't downgrade): ≥80% of point estimates are within the same range of test performance. • Serious (downgrade 1 level): <80% of point estimates are within the same range of test performance, and the point estimates are separated by 1 decision making threshold. • Very serious (downgrade 2 levels): <80% of point estimates are within the same range of test performance, and the point estimates are separated by 2 decision making thresholds. Where there are apparent differences in effect size due consideration was given to the appropriateness of pooling studies.
Imprecision	If the 95% confidence interval for the outcome crossed one of the clinical decision thresholds, the outcome was downgraded one level. If the 95% confidence interval spanned both thresholds, the outcome was downgraded twice. See the section on 'Using sensitivity and specificity as the primary outcome' for a description of how clinical decision thresholds were agreed.
Publication bias	If the review team became aware of evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability. Funnel plot analyses were not carried out as this method is not recommended by the Cochrane Collaboration (Macaskill et al., 2023) to detect publication bias for DTA studies. Macaskill et al. (2023) explained that statistical tests to detect funnel plot asymmetry are designed primarily for use on reviews of randomised trials and that applying such tests in systematic reviews of test accuracy is likely to result in publication bias being incorrectly indicated by the test far too often. They also recommended exploring heterogeneity in test accuracy as participant and study characteristics may be associated with study size as well as test accuracy. Heterogeneity have been considered already as part of the GRADE assessment within this review.

1

Reviewing economic evidence

The committee is required to make decisions based on the best available evidence of both clinical effectiveness and cost effectiveness. Guideline recommendations should be based on the expected costs of the different options in relation to their expected health benefits (that is, their 'cost effectiveness') rather than the total implementation cost. However, the committee will also need to be increasingly confident in the cost effectiveness of a recommendation as the cost of implementation increases. Therefore, the committee may require more robust evidence on the effectiveness and cost effectiveness of any recommendations that are expected to have a substantial impact on resources; any uncertainties must be offset by a compelling argument in favour of the recommendation. The cost impact or savings potential of a recommendation should not be the sole reason for the committee's decision.3

Health economic evidence was sought relating to the key clinical issues being addressed in the guideline. Health economists:

- Undertook a systematic review of the published economic literature.
- Undertook new cost-effectiveness analysis in priority areas.

Literature review

The health economists:

- Identified potentially relevant studies for each review question from the health economic search results by reviewing titles and abstracts. Full papers were then obtained.
- Reviewed full papers against prespecified inclusion and exclusion criteria to identify relevant studies (see below for details).
- Critically appraised relevant studies using economic evaluations checklists as specified in the NICE guidelines manual.
- Extracted key information about the studies' methods and results into health economic evidence tables (which can be found in appendices to the relevant evidence reports).

- Generated summaries of the evidence in NICE health economic evidence profile tables (included in the relevant evidence report for each review question) – see below for details.

Inclusion and exclusion criteria

Full economic evaluations (studies comparing costs and health consequences of alternative courses of action: cost–utility, cost-effectiveness, cost–benefit and cost–consequences analyses) and comparative costing studies that addressed the review question in the relevant population were considered potentially includable as health economic evidence.

Studies that only reported cost per hospital (not per patient), or only reported average cost effectiveness without disaggregated costs and effects were excluded. Literature reviews, abstracts, posters, letters, editorials, comment articles, unpublished studies and studies not in English were excluded. Studies published before 2010 and studies from non-OECD countries or the USA were also excluded, on the basis that the applicability of such studies to the present UK NHS context is likely to be too low for them to be helpful for decision-making.

Remaining health economic studies were prioritised for inclusion based on their relative applicability to the development of this guideline and the study limitations. For example, if a high quality, directly applicable UK analysis was available, then other less relevant studies may not have been included. Where exclusions occurred on this basis, this is noted in the relevant evidence report.

For more details about the assessment of applicability and methodological quality see Table 6 below and the economic evaluation checklist (appendix H of the NICE guidelines manual 2014) and the health economics review protocol, which can be found in each of the evidence reports.

When no relevant health economic studies were found from the economic literature review, relevant UK NHS unit costs related to the compared interventions were presented to the committee to inform the possible economic implications of the recommendations.

NICE health economic evidence profiles

NICE health economic evidence profile tables were used to summarise cost and cost-effectiveness estimates for the included health economic studies in each evidence review report. The health economic evidence profile shows an assessment of applicability and methodological quality for each economic study, with footnotes indicating the reasons for the assessment. These assessments were made by the health economist using the economic evaluation checklist from the NICE guidelines manual. It also shows the incremental costs, incremental effects (for example, quality-adjusted life years [QALYs]) and incremental cost-effectiveness ratio (ICER) for the base case analysis in the study, as well as information about the assessment of uncertainty in the analysis. See Table 6 for more details.

When a non-UK study was included in the profile, the results were converted into pounds sterling using the appropriate purchasing power parity.

Table 6: Content of NICE health economic evidence profile

Item	Description
Study	Surname of first author, date of study publication and country perspective with a reference to full information on the study.
Applicability	An assessment of applicability of the study to this guideline, the current NHS situation and NICE decision-making:[(a)] • Directly applicable – the study meets all applicability criteria, or fails to meet 1 or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness. • Partially applicable – the study fails to meet 1 or more applicability criteria, and this could change the conclusions about cost effectiveness. • Not applicable – the study fails to meet 1 or more of the applicability criteria, and this is likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review.
Limitations	An assessment of methodological quality of the study:[(a)] • Minor limitations – the study meets all quality criteria, or fails to meet 1 or more quality criteria, but this is unlikely to change the conclusions about cost effectiveness. • Potentially serious limitations – the study fails to meet 1 or more quality criteria, and this could change the conclusions about cost effectiveness. • Very serious limitations – the study fails to meet 1 or more quality criteria, and this is highly likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review.

Item	Description
Other comments	Information about the design of the study and particular issues that should be considered when interpreting it.
Incremental cost	The mean cost associated with one strategy minus the mean cost of a comparator strategy.
Incremental effects	The mean QALYs (or other selected measure of health outcome) associated with one strategy minus the mean QALYs of a comparator strategy.
Cost effectiveness	Incremental cost-effectiveness ratio (ICER): the incremental cost divided by the incremental effects (usually in £ per QALY gained).
Uncertainty	A summary of the extent of uncertainty about the ICER reflecting the results of deterministic or probabilistic sensitivity analyses, or stochastic analyses of trial data, as appropriate.

(a) Applicability and limitations were assessed using the economic evaluation checklist in appendix H of the NICE guidelines manual[3]

Undertaking new health economic analysis

As well as reviewing the published health economic literature for each review question, as described above, new health economic analysis was undertaken by the health economist in all areas.

The following general principles were adhered to in developing the cost-effectiveness analyses:

- Methods were consistent with the NICE reference case for interventions with health outcomes in NHS settings.
- The committee was involved in the design of the model, selection of inputs and interpretation of the results.
- Model inputs were based on the systematic review of the clinical literature supplemented with other published data sources where possible.
- When published data were not available committee expert opinion was used to populate the model.
- Model inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.
- The model was peer-reviewed by another health economist at NICE.

Full methods and results of the cost-effectiveness analysis are described in a separate economic analysis report.

Cost-effectiveness criteria

NICE sets out the principles that committees should consider when judging whether an intervention offers good value for money. In general, an intervention was considered to be cost effective (given that the estimate was considered plausible) if either of the following criteria applied:

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more clinically effective compared with all the other relevant alternative strategies), or
- the intervention cost less than £20,000 per QALY gained compared with the next best strategy.

If the committee recommended an intervention that was estimated to cost more than £20,000 per QALY gained, or did not recommend one that was estimated to cost less than £20,000 per QALY gained, the reasons for this decision are discussed explicitly in 'The committee's discussion of the evidence' section of the relevant evidence report, with reference to issues regarding the plausibility of the estimate or to factors set out in NICE methods manuals.

In the absence of health economic evidence

When making recommendations in areas not in the scope of the health economic analysis and where no relevant published evidence was identified the committee made a qualitative judgement about cost effectiveness by considering expected differences in resource use between options and relevant UK NHS unit costs, alongside the results of the review of clinical effectiveness evidence.

The UK NHS costs reported in the guideline are those that were presented to the committee and were correct at the time recommendations were drafted. They may have changed subsequently before the time of publication. However, we have no reason to believe they have changed substantially.

References

Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current

Follmann D, Elliott P, Suh I, Cutler J (1992) Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology 45:769–73

Fu R, Vandermeer BW, Shamliyan TA, et al. (2013) Handling Continuous Outcomes in Quantitative Synthesis In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK154408/

Norman G., Sloan JA., Wyrwich KW. (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41(5):582-92.