National Institute for Health and Care Excellence

Final

Pneumonia: diagnosis and management

NICE guideline: methods

NICE guideline NG250

Methods

September 2025

Final

Evidence reviews developed by the National Institute for Health and Care Excellence

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN: 978-1-4731-7194-7

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence commissioned the Guideline Development Team to bring together and update the NICE guidelines on:

- pneumonia in adults: diagnosis and management (CG191)
- pneumonia (community-acquired): antimicrobial prescribing (NG138)
- pneumonia (hospital-acquired): antimicrobial prescribing (NG139).

The remit for this new update is to provide NICE guidance on the diagnosis and management of pneumonia.

What this guideline covers

This guideline update and amalgamation covers recommendations and research recommendations on risk assessment tools, lung ultrasound, virtual wards, biomarkers, microbiological tests, antibiotic duration for children, adjunctive corticosteroids, non-invasive respiratory support, follow-up chest x-rays, and information for parents and carers of children with pneumonia. Please see the guideline scope for more information.

What this guideline does not cover

All other recommendations in this guideline were developed using the methods outlined in the methods chapter of the 2014 full guideline, available here: Evidence | Pneumonia in adults: diagnosis and management | Guidance | NICE.

Methods

This guideline was developed using the methods described in the Developing NICE guidelines: the manual.

Declarations of interest were recorded according to the NICE conflicts of interest policy.

Developing the review questions and outcomes

The 11 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] (PICO[S]) for reviews of interventions
- Population, index test(s), reference standard and outcome for reviews of diagnostic and prognostic test accuracy
- Population, prognostic tool(s), outcomes, measures, study type(s) for reviews of outcome prediction tools

Full literature searches, critical appraisals and evidence reviews were completed for all review questions. Details of these elements are found in the review protocols for each review (see Appendix A of each relevant review). Where protocol deviations have been made, these will be reported in the 'Methods' section of the individual review.

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. Where possible, review protocols were prospectively registered in the PROSPERO register of systematic reviews. Protocols are included in appendix A in each evidence review along with the PROPSERO registration number where available.

Searching for evidence

Evidence was searched for each review question using the methods specified in the Developing NICE guidelines: the manual. Full details of search strategies, databases searched, and numbers of studies identified can be found in appendix B of each individual review.

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example 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. Where resources allowed, 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

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 and appropriate critical appraisal tools were used to assess study quality and applicability.

Where studies used mixed populations of patients with respiratory illness, a cut off rule of >=75% pneumonia patients was used to select a studies for inclusion.

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, which in all cases was the ROBIS checklist. 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 1, not the quality of evidence contained within it, which was either directly extracted from the synthesis or assessed in the usual way as outlined in the section on 'Appraising the quality of evidence'. These assessments were used as a guide, but final decisions on whether to use a synthesis were discussed with the committee and exceptions could be made on a case by case basis.

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

- Low risk of bias – 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. The review is thorough and well conducted.
- Moderate risk of bias – 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. The review is well conducted but has some limitations.
- High risk of bias – It is possible that relevant and important studies have been missed by the review. There are errors or methodological limitations in the way the review was conducted that may impact upon it's use.

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 1. When published evidence syntheses were used instead of undertaking a new review, the analyses and GRADE conducted by the study authors were extracted directly. This means that there may be minor variations in how the analysis method was used and how GRADE was applied to the findings, compared to standard NICE methods. When published evidence syntheses were used only as a source of primary studies, or when GRADE was not applied to syntheses by the study authors, data from these evidence syntheses were quality assessed and presented in GRADE 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 1: Criteria for using published evidence syntheses as a source of data

Risk of bias	Applicability	Use of published evidence synthesis
Low	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.
Low	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. Network meta-analyses was considered in situations where there were at least 3 treatment alternatives, but no network meta-analyses were required in this update. When there were 2 treatment alternatives, pairwise meta-analysis was used to compare interventions. RCT and non-randomised comparative studies data were pooled separately.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.3. 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 where sufficient information was reported in the included papers. Absolute risks were 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).

A pooled mean difference was calculated for continuous outcomes (using the inverse variance method) when the same scale was used to measure an outcome across different studies. Where different studies presented continuous data measuring the same outcome but using different numerical scales (e.g. a 0-10 and a 0-100 visual analogue scale), these outcomes were all converted to the same scale before meta-analysis was conducted on the mean differences.

For continuous outcomes analysed as mean differences, change from baseline values were used in the meta-analysis if they were accompanied by a measure of spread (for example standard deviation). Where change from baseline (accompanied by a measure of spread) were not reported, the corresponding values at the timepoint of interest were used. If only a subset of trials reported change from baseline data, final timepoint values were combined with change from baseline values to produce summary estimates of effect.

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. 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 was clearly not met, 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 $I^2 \geq 50\%$.

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).

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:

- **Positive likelihood ratios** describe how many times more likely positive index test results 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 index test results 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])
- **Sensitivity** is the probability that the index test results will be positive in a person with the condition.
  - sensitivity = TP/(TP+FN)
- **Specificity** is the probability that the index test results will be negative in a person without the condition.
  - specificity = TN/(FP+TN)

Meta-analysis of diagnostic accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.1 (Deeks et al. 2022).

Where five or more studies were available, a bivariate model was fitted using the `mada` package in R v3.4.0, which accounts for the correlations between positive and negative likelihood ratios, and between sensitivities and specificities. Where sufficient data were not available (2-4 studies), separate independent pooling was performed for positive likelihood ratios, negative likelihood ratios, sensitivity and specificity, using R. This approach is conservative as it is likely to somewhat underestimate test accuracy, due to failing to account for the correlation and trade-off between sensitivity and specificity (see Deeks 2010).

Random-effects models (der Simonian and Laird 1986) were fitted for all syntheses, as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010).

Data synthesis for predictive accuracy data

For the purpose of this guideline predictive accuracy data are classified as any data in which an index 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 develop a condition or outcome of interest at some time after the observation of the index feature 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 go on to develop the condition or outcome of interest) and false positives and true negatives (in people who do not).

When deciding whether data should be synthesised or presented separately, heterogeneity in the population, index feature and outcome to be predicted were considered to determine whether data could be meaningfully combined. When it was decided that data could be meaningfully combined, the same methods were used when synthesising predictive accuracy data as those described for synthesising diagnostic accuracy data.

Appraising the quality of evidence

Intervention studies (relative effect estimates)

RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias Tool version 2.0. Non-randomised controlled trials and cohort studies were quality assessed using the ROBINS-I tool. Other study types (for example controlled before and after studies) were assessed using the preferred option specified in appendix H of Developing NICE guidelines: the manual. Evidence on each outcome for each individual study was classified into one of the following 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.
- Critical risk of bias (ROBINS-I only) - It is very 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, 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. For all reviews except review B, no appropriate published MIDs were identified so the Guideline Committee were asked to prospectively specify any outcomes where they felt a consensus clinical decision threshold could be defined from their experience. Again a consensus clinical decision threshold was not identified so default MIDs were used in all reviews except review B. For review B on Hospital at home, published MIDs were identified and applied for all quality of life outcomes assessed using the SF-36.

For continuous outcomes expressed as a mean difference where no other clinical decision threshold was available, a clinical decision threshold of 0.5 times the median standard deviations of the comparison group arms was used (Norman et al. 2003). For continuous outcomes expressed as a standardised mean difference where no other clinical decision threshold was available, a clinical decision threshold of 0.5 standard deviations was used. For relative risks and hazard ratios, a default clinical decision threshold for dichotomous outcomes of 0.8 to 1.25 was used. Where possible, odds ratios were converted to risk ratios before presentation to the committee to aid interpretation.

GRADE for intervention studies analysed using pairwise analysis

GRADE was used to assess the certainty of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials, non-randomised controlled trials and cohort studies (which were quality assessed using the Cochrane risk of bias tool 2.0 or ROBINS-I) were initially rated as high quality while data from other study types were initially rated as low quality. The quality of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 2It was noted that many of these criteria have changed as a result of NICE methods harmonisation work, but because those changes were established part way through development of this guideline, original criteria have been adopted across all reviews for consistency.

Table 2: Rationale for downgrading quality of evidence for intervention studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias (combine serious and critical as high risk for ROBINS-I), the outcome was downgraded two levels.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.

GRADE criteria	Reasons for downgrading quality
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: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.4% and 66.7%, the outcome was downgraded one level, or if data on the outcome was only available from one study. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels. Where I^2 is 80% or above, data may be too heterogeneous to meaningfully pool. This will be considered on a case by case basis.
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 crosses both lines of the MID.
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.

For outcomes that were originally assigned a quality rating of 'low' (when the data was from observational studies that were not appraised using the ROBINS-I checklist), the quality of evidence for each outcome was upgraded if any of the following three conditions were met and the risk of bias for the outcome was rated as 'no serious':

- Data from studies showed an effect size sufficiently large that it could not be explained by confounding alone.
- Data showed a dose-response gradient.
- Data where all plausible residual confounding was likely to increase our confidence in the effect estimate.

Diagnostic accuracy studies

Individual diagnostic accuracy studies were quality assessed using the QUADAS-2 tool and 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.

GRADE for diagnostic accuracy evidence

Evidence from diagnostic accuracy studies was initially rated as high-quality and downgraded according to the standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 4 below.

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

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

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.

If studies could not be pooled in a meta-analysis, GRADE assessments were undertaken for each study individually and reported as separate lines in the GRADE profile.

Using likelihood ratios as the primary outcomes

The following schema (Table 3), adapted from the suggestions of Jaeschke et al. (1994), was used to interpret the likelihood ratio findings from diagnostic test accuracy reviews.

Table 3: Interpretation of likelihood ratios

Value of likelihood ratio	Interpretation
LR ≤ 0.1	**Very large** decrease in probability of disease
0.1 < LR ≤ 0.2	**Large** decrease in probability of disease
0.2 < LR ≤ 0.5	**Moderate** decrease in probability of disease
0.5 < LR ≤ 1.0	**Slight** decrease in probability of disease
1.0 < LR < 2.0	**Slight** increase in probability of disease
2.0 ≤ LR < 5.0	**Moderate** increase in probability of disease
5.0 ≤ LR < 10.0	**Large** increase in probability of disease
LR ≥ 10.0	**Very large** increase in probability of disease

The committee were consulted to set 2 clinical decision thresholds for each measure: the likelihood ratio above (or below for negative likelihood ratios) which a test would be recommended, and a second below (or above for negative likelihood ratios) which a test would be considered of no clinical use. These were used to judge imprecision (see below). If the committee were unsure which values to pick, then the default values of 2 for LR+ and 0.5 for LR- were used based on Table 6, with the line of no effect (being 1.0) as the second clinical decision line in both cases.

Table 4: Rationale for downgrading quality of evidence for diagnostic accuracy data

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
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: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.3% and 66.7%, the outcome was downgraded one level, or if data on the outcome was only available from one study. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels.
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 likelihood ratios 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.

Predictive accuracy studies

Individual prognostic studies that did not assess or develop a prediction model were quality assessed using the QUIPS checklist. Studies that developed or assessed a prediction model were assessed using the PROBAST checklist. 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 outcome to be predicted.
- Partially indirect – Important deviations from the protocol in one of the population, index feature and/or outcome to be predicted.
- Indirect – Important deviations from the protocol in at least two of the population, index feature and/or outcome to be predicted.

Modified GRADE for predictive accuracy data

GRADE has not been developed for use with predictive accuracy data, therefore a modified approach was applied using the GRADE framework. Evidence from cohort, cross sectional or case-control studies was initially rated as low-quality, and then assessed according to the same criteria as described in the section on standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 6 below.

The choice of primary outcome for decision making was determined by the committee and GRADE assessments were undertaken based on these outcomes. In some cases, this was area under the curve or c-statistics, but in other reviews (e.g. review H) likelihood ratios were considered primary outcomes .

When interpreting calibration statistics (calibration slope and intercept) the reviews on risk assessment tools, calibration is understood as the degree to which predictions made by a model, on average, agree with the overall outcomes observed in the sample. As a rule of thumb, if the slope is close to 1 and the intercept is close to 0, then there is good overall calibration.

Where likelihood ratios were used as the primary outcomes, see the above section on ‘Using likelihood ratios as the primary outcomes.’

Using area under the curve or c-statistics as the primary outcomes

The following schema (Table 5) was used to interpret the area under the curve and c-statistic findings from predictive accuracy reviews, with each category of classification accuracy representing a clinical decision threshold. When judging

imprecision, 95% confidence intervals that crossed one threshold of classification accuracy were downgraded once, and ones that crossed 2 thresholds of classification accuracy were downgraded twice (for example a 95% CI of 0.68 to 0.82 would cross from adequate to excellent classification accuracy so would be downgraded twice).

Table 5: Interpretation of c-statistics

Value of c-statistic	Interpretation
c-statistic <0.6	**Poor** classification accuracy
0.6 ≤ c-statistic < 0.7	**Adequate** classification accuracy
0.7 ≤ c-statistic < 0.8	**Good** classification accuracy
0.8 ≤ c-statistic < 0.9	**Excellent** classification accuracy
0.9 ≤ c-statistic < 1.0	**Outstanding** classification accuracy

Table 6: Rationale for downgrading quality of evidence for predictive accuracy data

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels. If there is no meta-analysis, indirectness is based on the applicability of the study.
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: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.3% and 66.7%, the outcome was downgraded one level, or if data on the outcome was only available from one study. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels.
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 area under the curve and c-statistics as the primary outcome' for a description of how clinical decision thresholds were agreed.

GRADE criteria	Reasons for downgrading quality
	If there is no meta-analysis, a single result may also be downgraded as serious if the sample size is small (<500) or very serious if the sample size is very small (<250).
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.

Reviewing economic evidence

Inclusion and exclusion of economic studies

Literature reviews seeking to identify published cost–utility and cost effectiveness analyses of relevance to the issues under consideration were conducted for all questions. In each case, the search undertaken for the clinical review was modified, retaining population and intervention descriptors, but removing any study-design filter and adding a filter designed to identify relevant health economic analyses. In assessing studies for inclusion, population, intervention and comparator, criteria were always identical to those used in the parallel clinical search; only cost–utility analyses were included. Costing and US studies were excluded due to limited information or not being representative of the UK healthcare system. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Appraising the quality of economic evidence

Economic studies identified through a systematic search of the literature were appraised using a methodology checklist (applicability and limitations checklist) designed for economic evaluations (NICE guidelines manual; 2014). This checklist is not intended to judge the quality of a study per se, but to determine whether an existing economic evaluation is useful to inform the decision-making of the committee for a specific topic within the guideline.

There are 2 parts of the appraisal process. The first step is to assess applicability (that is, the relevance of the study to the specific guideline topic and the NICE reference case); evaluations are categorised according to the criteria in Table 7.

Table 1 Applicability criteria

Level	Explanation
Directly applicable	The study meets all applicability criteria, or fails to meet one or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness
Partially applicable	The study fails to meet one or more applicability criteria, and this could change the conclusions about cost effectiveness
Not applicable	The study fails to meet one or more applicability criteria, and this is likely to change the conclusions about cost effectiveness. These studies are excluded from further consideration

In the second step, only those studies deemed directly or partially applicable are further assessed for limitations (that is, methodological quality); see categorisation criteria in Table 8.

Table 8: Methodological criteria

Level	Explanation
Minor limitations	Meets all quality criteria, or fails to meet one or more quality criteria but this is unlikely to change the conclusions about cost effectiveness
Potentially serious limitations	Fails to meet one or more quality criteria and this could change the conclusions about cost effectiveness
Very serious limitations	Fails to meet one or more quality criteria and this is highly likely to change the conclusions about cost effectiveness. Such studies should usually be excluded from further consideration

Where relevant, a summary of the main findings from the systematic search, review and appraisal of economic evidence is presented in an economic evidence profile alongside the clinical evidence.

Health Economic Modelling

As well as reviewing the published economic literature for each review question, as described above, original economic analysis was prioritised in selected areas. Priority areas for new health economic analysis were agreed by the committee. However, due to a lack of available effectiveness data no original health economic modelling was undertaken.

References

Deeks, J. J., Bossuyt, P. M., Leeflang, M. M., & Takwoingi, Y. (Eds.). (2023). *Cochrane handbook for systematic reviews of diagnostic test accuracy*. John Wiley & Sons.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177-188.

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/

NICE guidelines manual (2014) Developing-nice-guidelines-the-manual-2014-edition-pdf-6596134525.

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.