National Institute for Health and Care Excellence

Draft

# Neonatal infection: antibiotics for prevention and treatment

## NICE guideline: methods

*NICE guideline <number>*

*Methods*

*December 2020*

*Draft for Consultation*

*Evidence reviews were developed by the NICE guideline updates team*

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 All NICE guidance is subject to regular review and may be updated or withdrawn.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

## What this guideline covers

This method document describes the evidence reviews that were part of the 2021 update of this guideline. These evidence reviews included:

Evidence review A: Info and support

Evidence review B: Intrapartum antibiotics

Evidence review C: PPROM

Evidence review D: Risk factors for early onset

Evidence review E: Risk factors for late onset

Evidence review F: Intravascular catheters

Evidence review G: Investigations

Evidence review H: Antibiotics

Evidence review I: Antifungals

## What this guideline does not cover

This method does not cover the parts of the 2012 guideline that were not updated in 2021. The methods used to develop these sections of the guideline are described in the full version of the guideline.

# Methods

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

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

## Developing the review questions and outcomes

The 13 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline updates team and refined and validated by the guideline committee.

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. Where possible, review protocols were prospectively registered in the PROSPERO register of systematic reviews.

### Searching for evidence

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

### Selecting studies for inclusion

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

The reviews undertaken for this guideline all made use of the priority screening functionality with the EPPI-reviewer systematic reviewing software. This uses a machine learning algorithm (specifically, an SGD classifier) to take information on features (1, 2 and 3 word blocks) in the titles and abstract of papers marked as being ‘includes’ or ‘excludes’ during the title and abstract screening process, and re-orders the remaining records from most likely to least likely to be an include, based on that algorithm. This re-ordering of the remaining records occurs every time 25 additional records have been screened.

For most reviews in this guideline, priority screening was used to highlight the most relevant records earlier in the search but was not used as a method to stop abstract

screening early. Consequently, the whole abstract database was searched for most review questions.

For two reviews (Review G – Investigations and Review H – antibiotics), priority screening was used to prioritise the articles that were most relevant to the review and to allow screening to be stopped early.

Research is currently ongoing as to what are the appropriate thresholds where reviewing of abstract can be stopped, assuming a defined threshold for the proportion of relevant papers it is acceptable to miss on primary screening. As a conservative approach until that research has been completed, the following rules were adopted during the production of this guideline:

- At least 50% of the identified abstract (or 1,000 records, if that is a greater number) were always screened.
- After this point, screening was only terminated if at least 500 additional abstracts were screened without a single new include being identified.
- A random 10% sample of the studies remaining in the database when the threshold was reached were additionally screened, to check if a substantial number of relevant studies were not being correctly classified by the algorithm, with the full database being screened if concerns were identified.

For review G (Investigations), 2796 abstracts (64% of the database) were screened and for review H (antibiotics), 2949 abstracts (60% of the database) were screened before the stopping criteria was met. As an additional check to ensure this approach did not miss relevant studies, the included studies lists of systematic reviews were searched to identify any papers not identified through the primary search.

# Incorporating published systematic reviews

For all review questions where a literature search was undertaken looking for a particular study design, systematic reviews containing studies of that design were also included. All included studies from those systematic reviews were screened to identify any additional relevant primary studies not found as part of the initial search.

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

## Quality assessment

Individual systematic reviews that were considered as a direct source of data were quality assessed using the ROBIS tool, with each 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 individual systematic review 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.

**Table 1: 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 |

| Quality | Applicability | Use of published evidence synthesis |
|---|---|---|
| | | 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. |
| Low | Any | The published evidence synthesis was not used as a source of data and a full literature review was completed. |
| Any | Not applicable | |

For most reviews, no additional studies were identified from systematic reviews. For one review (review D - early onset risk factors), 2 additional studies which examined the effectiveness of prognostic models for early-onset infection were found from systematic reviews. The primary studies were reviewed and included in this review. For review I (antifungals), most results were taken directly from 2, high quality, fully applicable, systematic reviews which examined the effectiveness of antifungals as prophylaxis in very low birthweight and preterm babies.

# Evidence synthesis and meta-analyses of pair-wise data

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome. For continuous outcomes analysed as mean differences, where change from baseline data were reported in the trials and were accompanied by a measure of spread (for example standard deviation), these were extracted and used in the meta-analysis. Where measures of spread for change from baseline values were not reported, the corresponding values at study end were used and were combined with change from baseline values to produce summary estimates of effect. These studies were assessed to ensure that baseline values were balanced across the treatment groups; if there were significant differences at baseline these studies were not included in any meta-analysis and were reported separately. Where there were differences in populations or interventions, meta-analyses were not performed, and instead the results of individual studies were presented.

# Evidence of effectiveness of interventions

## Quality assessment

Individual RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias 2.0 Tool. Other studies were quality assessed using the ROBINS-I 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, 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 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.

## Methods for combining intervention evidence

Meta-analyses of interventional data were conducted with reference to the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al. 2011).

Where outcomes measured the same underlying construct but used different instruments/metrics, data were analysed using standardised mean differences (Hedges' g).

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).

Fixed- and random-effects models (der Simonian and Laird) were fitted for all syntheses, 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 one or both of the following conditions was met:

- 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.

- The presence of significant statistical heterogeneity in the meta-analysis, defined as $I2 \geq 50\%$.

In situations where subgroup analyses were conducted, pooled results and results for the individual subgroups are reported when there was evidence of between group heterogeneity, defined as a statistically significant test for subgroup interactions (at the 95% confidence level). Where no such evidence as identified, only pooled results are presented.

Meta-analyses were performed in Cochrane Review Manager V5.3.

## Minimal clinically important differences (MIDs) for intervention reviews

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal clinically important difference thresholds relevant to this guideline. 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 MID could be defined from their experience. MIDs identified through this process were intended to be used to inform discussions on the clinical importance of effects and the precision of effect estimates. No published MIDs were found through this process and the committee did not wish to pre specify consensus MIDs for any outcome. The clinical importance of effects was judged by the committee taking into account evidence across all outcomes and absolute effect estimates. These discussions are documented in the committee discussion section of each evidence review.

## Modified GRADE for pairwise meta-analyses of interventional evidence

GRADE was used to assess the quality of evidence for the selected outcomes as specified in 'Developing NICE guidelines: the manual'. Data from randomised controlled trials, non-randomised controlled trials and cohort studies (which were quality assessed using the Cochrane risk of bias tool or ROBINS-I) were initially rated as high quality while data from other study types were initially rated as low quality. Ratings were subsequently downgraded once for each serious source of uncertainty and twice of each very serious source of uncertainty, as outlined in the table below.

This evidence review for this guideline was conducted using a modified version of the GRADE approach to rate the certainty of evidence in systematic reviews. This is part of a pilot project being undertaken by NICE, to examine the assessment of certainty of evidence in systematic reviews. Instead of using predefined MIDs to assess imprecision in GRADE tables, imprecision was assessed qualitatively during committee discussions. These discussions involved consideration of published MIDs where they exist, but the committee were also encouraged to make judgements of imprecision based on the 95% confidence intervals and sample sizes reported in the GRADE tables. The committee were not aware of any published MIDs for any of the outcomes in the intervention reviews and so the discussions were based on the width of confidence intervals and whether they crossed the line of no effect. This should enable judgements of clinical importance to be made in the context of wider decision making, taking into account evidence across all outcomes and analyses, including health economic analyses.

Committee discussions regarding the clinical importance of effects was recorded in the 'imprecision and clinical importance of effects' section of the evidence review. In particular, this included consideration of whether the whole effect of a treatment (which may be felt across multiple independent outcome domains) would be likely to be clinically meaningful, rather than simply whether each individual sub outcome might be meaningful in isolation. The impact of imprecision on the recommendations was presented in the 'quality of the evidence' section of the committee discussion in the evidence review.

**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.<br>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.<br>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.<br>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.<br>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.<br>N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study.<br>Not serious: If the $I^2$ was less than 33.3%, the outcome was not downgraded.<br>Serious: If the $I^2$ was between 33.3% and 66.7%, the outcome was downgraded one level.<br>Very serious: If the $I^2$ was greater than 66.7%, the outcome was downgraded two levels. |
| Imprecision | This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes. |

The quality of evidence for each outcome was upgraded if any of the following three conditions were met:

- Data from non-randomised studies showing an effect size sufficiently large that it cannot be explained by confounding alone.
- Data showing a dose-response gradient.
- Data where all plausible residual confounding is likely to increase our confidence in the effect estimate.

## 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.

# Methods for combining direct and indirect evidence (network meta-analysis) for interventions

Hierarchical Bayesian Network Meta-Analysis (NMA) was performed using WinBUGS version 1.4.3. The models used reflected the recommendations of the NICE Decision Support Unit's Technical Support Documents (TSDs) on evidence synthesis, particularly TSD 2 ('A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials'; see http://www.nicedsu.org.uk). The WinBUGS code provided in the appendices of the TSDs was used without substantive alteration to specify synthesis models.

Three separate chains with different initial values were used. Results were assessed for convergence to determine the length of 'burn-in' period required by examining the 'bgdiag' and 'history' plots. Results were reported summarising at least 10,000 samples from the posterior distribution of each model, having run and discarded the 'burn-in' iterations. The MC error was assessed to check that it was sufficiently small (less than 5% of the standard deviation of the posterior distribution for each parameter) and additional samples were summarised if this was the case.

Non-informative prior distributions were used in all models. Unless otherwise specified, trial-specific baselines and treatment effects were assigned Normal (0, 10000) priors, and the between-trial standard deviations used in random-effects models for dichotomous outcomes were given Uniform (0, 5) priors. These are consistent with the recommendations in TSD 2 for dichotomous outcomes. For the 1 continuous outcome (length of stay) we used Uniform (0, 10) priors, which is substantially wider than the expected variance for this outcome.

Fixed-effect and random-effects models were explored for each outcome, with the final choice of model based on the total residual deviance and deviance information criterion (DIC): if DIC was at least 3 points lower for the random-effects model, it was preferred; otherwise, the fixed-effect model was considered to provide an equivalent fit to the data in a more parsimonious analysis, and was preferred.

Inconsistency between direct and indirect evidence was assessed by inspecting NMA outputs compared with pairwise direct results. We also fitted 'inconsistency models' and compared the posterior mean deviance contribution for each data point to the analogous value from the main ('consistency') models to identify areas of inconsistency (see TSD 4 'Inconsistency in networks of evidence based on randomised controlled trials'; http://www.nicedsu.org.uk). Datapoints with an absolute difference in deviance of 0.5 or greater between the 2 models were deemed worthy of additional investigation, to see if any reason for underlying heterogeneity could be established.

## Modified GRADE for network meta-analyses

A modified version of the standard GRADE approach for pairwise interventions was used to assess the quality of evidence across the network meta-analyses undertaken. While most criteria for pairwise meta-analyses still apply, it is important to adapt some of the criteria to take into consideration additional factors, such as how each 'link' or pairwise comparison within the network applies to the others. As a

result, the following was used when modifying the GRADE framework to a network meta-analysis. It is designed to provide a single overall quality rating for an NMA, which can then be combined with pairwise quality ratings for individual comparisons (if appropriate), to judge the overall strength of evidence for each comparison.

**Table 3: Rationale for downgrading quality of evidence for intervention studies**

| GRADE criteria | Reasons for downgrading quality |
|---|---|
| Risk of bias | Not serious: If fewer than 33.3% of the studies in the network meta-analysis were at moderate or high risk of bias, the overall network was not downgraded.<br>Serious: If greater than 33.3% of the studies in the network meta-analysis were at moderate or high risk of bias, the network was downgraded one level.<br>Very serious: If greater than 33.3% of the studies in the network meta-analysis were at high risk of bias, the network was downgraded two levels. |
| Indirectness | Not serious: If fewer than 33.3% of the studies in the network meta-analysis were partially indirect or indirect, the overall network was not downgraded.<br>Serious: If greater than 33.3% of the studies in the network meta-analysis were partially indirect or indirect, the network was downgraded one level.<br>Very serious: If greater than 33.3% of the studies in the network meta-analysis were indirect, the network was downgraded two levels. |
| Inconsistency | N/A: Inconsistency was marked as not applicable if there were no links in the network where data from multiple studies (either direct or indirect) were synthesised.<br>For network meta-analyses conducted under a Bayesian framework, the network was downgraded one level if the DIC for an inconsistency model was more than 3 points higher than the corresponding consistency model. |
| Imprecision | This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes.. |

# Diagnostic test accuracy evidence

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 are as follows:

- **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])
- **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)

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

**Table 4: 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 schema above has the effect of setting a clinical decision threshold for positive likelihoods ratio at 2, and a corresponding clinical decision threshold for negative likelihood ratios at 0.5. Likelihood ratios (whether positive or negative) falling between these thresholds were judged to indicate no meaningful change in the probability of disease.

## Quality assessment

Individual studies were quality assessed using the QUADAS-2 tool, which contains four domains: patient selection, index test, reference standard, and flow and timing. 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.

## Methods for combining diagnostic test accuracy evidence

Meta-analysis of diagnostic test accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010).

Where applicable, diagnostic syntheses were stratified by:

- Presenting symptomatology (features shared by all participants in the study, but not all people who could be considered for a diagnosis in clinical practice).
- The reference standard used for true diagnosis.

Where five or more studies were available for all included strata, 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 Microsoft Excel. 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) were fitted for all syntheses, as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010).

## GRADE for diagnostic test accuracy evidence

The choice of primary outcome for decision making was determined by the committee and GRADE assessments were undertaken using the appropriate method from those listed below.

In all cases, following completion of the GRADE table, the downstream effects of these tests on patient- important outcomes were considered. This could be 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. Alternatively, in reviews where a decision model is being carried (for example, as part of an economic analysis), these consequences may be incorporated here instead.

### Using likelihood ratios as the primary outcomes

GRADE assessments were only undertaken for positive and negative likelihood ratios, as the clinical decision thresholds used to assess imprecision were based on these outcomes, but results for sensitivity and specificity are also presented alongside those data.

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: Rationale for downgrading quality of evidence for diagnostic questions using likelihood ratio measures.

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 , with the line of no effect as the second clinical decision line in both cases.

**Table 5: Rationale for downgrading quality of evidence for diagnostic questions using likelihood ratio measures.**

| 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.<br>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.<br>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.<br>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.<br>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.<br>N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study.<br>Not serious: If the $I^2$ was less than 33.3%, the outcome was not downgraded.<br>Serious: If the $I^2$ was between 33.3% and 66.7%, the outcome was downgraded one level.<br>Very serious: If the $I^2$ was greater than 66.7%, the outcome was downgraded two levels. |

Upgrading of evidence was carried out with caution, and was considered when:

- Test accuracy was extremely high (and the associated confidence intervals also only include extremely high accuracy estimates)
- All plausible confounding would act to reduce test accuracy

### 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 (prognostic) accuracy evidence

In this guideline, predictive (or prognostic) accuracy 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 go on to develop the condition of interest 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 develop the condition) and false positives and true negatives (in people who, according to the reference standard, do not). This category would include studies classed as prediction models under the TRIPOD statement, provided the data were reported a 2x2 classification data.

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

- **Positive likelihood ratios** describe how many times more likely positive features are in people who develop the condition compared to people who do not. 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 who develop the condition compared to people who do not. 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 feature will be positive in a person who goes on to develop the condition.
    - sensitivity = TP/(TP+FN)
- **Specificity** is the probability that the feature will be negative in a person who does not go on to develop the condition.
    - specificity = TN/(FP+TN)

The following schema, adapted from the suggestions of Jaeschke et al. (1994), was used to interpret the findings from predictive accuracy evidence.

**Table 6: 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 schema above has the effect of setting a clinical decision threshold for positive likelihoods ratio at 2, and a corresponding clinical decision threshold for negative likelihood ratios at 0.5. Likelihood ratios (whether positive or negative) falling between these thresholds were judged to indicate no meaningful change to probability of disease.

## Quality assessment

Individual studies reporting clinical prediction models were quality assessed using the PROBAST tool, which contains five domains: participant selection, predictors, outcome, sample size and participant flow, analysis (Wolff et al. 2018). Cohort studies reporting other predictive accuracy data were quality assessed using the QUIPs checklist. Each individual study was classified into one of the following three groups based on an assessment of the overall risk of bias:

- 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, predictive 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, predictive feature and/or reference standard.
- Partially indirect – Important deviations from the protocol in one of the population, predictive feature and/or reference standard.
- Indirect – Important deviations from the protocol in at least two of the population, predictive feature and/or reference standard.

## Methods for combining predictive accuracy evidence

Where applicable, predictive accuracy syntheses were stratified by:

- Presenting symptomatology (features shared by all participants in the study, but not all people in the full relevant clinical population).

- The length of time between the measurement of the predictive feature and the final outcome.
- The reference standard used for categorising true positives.

Where five or more studies were available for all included strata, 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 Microsoft Excel. This approach is likely to somewhat underestimate test accuracy (see Deeks 2001).

Random-effects models (der Simonian and Laird) were fitted for all syntheses, due to the expected level of between study heterogeneity in prognostic reviews.

## Modified GRADE for predictive accuracy evidence

GRADE has not been developed for use with predictive accuracy studies; therefore a modified approach was applied using the GRADE framework. GRADE assessments were only undertaken for positive and negative likelihood ratios, as the clinical decision thresholds used to assess imprecision were based on these outcomes.

Cohort studies were initially rated as high-quality evidence if well conducted, and then downgraded according to the standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table  below.

**Table 7:  Rationale for downgrading quality of evidence for evidence reporting 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.<br>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.<br>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.<br>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.<br>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.<br>N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study.<br>Not serious: If the $I^2$ was less than 33.3%, the outcome was not downgraded.<br>Serious: If the $I^2$ was between 33.3% and 66.7%, the outcome was downgraded one level.<br>Very serious: If the $I^2$ was greater than 66.7%, the outcome was downgraded two levels. |
| Imprecision | If the 95% confidence interval for sensitivity crossed one of the clinical decision thresholds, the outcome was downgraded one level, as the data were deemed to be imprecise. If the 95% confidence interval spanned both thresholds, the outcome was downgraded twice, as suffering from very serious imprecision. Specificity was assessed for imprecision in a similar manner using the 2 previously defined clinical decision thresholds. |

### 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.

## Other prognostic evidence

Other prognostic studies were also included if they reported outcomes of c-statistics. Hazard ratios were also included in the review protocols for some reviews, but no data were found for these outcomes. These studies were also quality assessed using the PROBAST checklist (in the case of studies reporting clinical prediction models) or the QUIPs checklist (in the case of other prognostic studies), as in the predictive accuracy section above.

### Assessing c-statistics

C-statistics were assessed in a similar manner to likelihood ratios using the categories in Table below.

**Table 8 Interpretation of c-statistics (Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression: John Wiley & Sons; 2013.)**

| 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** |

Meta-analyses were carried out using the metamisc package in R v3.4.0, which confines the analysis results to between 0 and 1 matching the limited range of values that c-statistics can take. Random effects meta-analysis was used when the $I^2$ was 50% or greater.

In any meta-analyses where some (but not all) of the data came from studies at high risk of bias, a sensitivity analysis was conducted, excluding those studies from the analysis. Results from both the full and restricted meta-analyses are reported. Similarly, in any meta-analyses where some (but not all) of the data came from indirect studies, a sensitivity analysis was conducted, excluding those studies from the analysis.

A modified version of GRADE was carried out to assess the quality of the meta-analysed c-statistics as follows:

- Imprecision - the 95% CI boundaries were examined and if they crossed 2 categories of test classification accuracy as shown in Table 8 Interpretation of c-statistics (Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression: John Wiley & Sons; 2013.)

then the study was downgraded once (imprecision rated as serious); if the boundaries crossed 3 categories then the study was downgraded twice (very serious imprecision).

- Inconsistency, indirectness and risk of bias were determined using the methods in the section on GRADE for predictive accuracy evidence.

In cases where meta-analyses could not be carried out due to the large numbers of studies without 95% CI, the following decision rules were used to assess risk of bias, indirectness, imprecision and inconsistency for each outcome:

1. Risk of bias and indirectness were calculated as normal, but using the study weight by population, rather than weight in the meta-analysis.

2. Imprecision

    a. Single study with 95% CI: the 95% CI boundaries were examined and if they crossed 2 categories of test classification accuracy then the study was downgraded once (imprecision rated as serious); if the boundaries crossed 3 categories then the study was downgraded twice (very serious imprecision).

        i. Multiple studies with 95% CI: the individual studies were rated as in a. and then if >33.3% of the studies by population weight were rated serious then the analysis was downgraded once; if > 33.33% were rated very serious the analysis was downgraded twice.

    b. Single study or multiple studies without 95% CI: the mean sample size was calculated and if this was < 250 then the analysis was downgraded twice (very serious); if it was >250, but > 500 the analysis was downgraded once (serious); if the mean was > 500 people/study then the analysis was not downgraded (not serious).

    c. Multiple studies with and without 95% CI: the studies without 95% CI were analysed as in 2c; those with 95% CI were analysed as in 2b. The results were averaged, but the number of studies in each group were also taken

into account with the result that if there were a lot more studies in one group compared to the other then that group rating would be used. In general, not serious and serious or not serious and very serious were averaged to serious; serious and very serious resulted in a very serious rating.

3. Inconsistency

    a. Single study with or without 95% CI: N/A

    b. Multiple studies with or without 95% CI: the highest and lowest point estimates were examined. If they spanned < 2 categories of c-statistic classification accuracy the analysis was rated as not serious for inconsistency; if they spanned 2 categories this was rated as serious and ≥ 3 categories was rated as very serious.

# Association studies

In this guideline, association studies are defined those reporting data showing an association of a predictor (either a single variable or a group of variables) and an outcome variable, where the data are not reported in terms of outcome classification (i.e. diagnostic/predictive accuracy). Data were reported as hazard ratios (if measured over time) or odds ratios or risk ratios (if measured at a specific time-point). Data reported in terms of model fit or predictive accuracy were not assessed using this method.

## Quality assessment

Individual cohort studies were quality assessed using the QUIPS 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.

Individual cross-sectional studies were quality assessed using the Joanna Briggs Institute critical appraisal checklist for analytical cross sectional studies (2016), which contains 8 questions covering: inclusion criteria, description of the sample, measures of exposure, measures of outcomes, confounding factors, and statistical analysis. 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.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, predictors 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, predictors and/or outcomes.
- Partially indirect – Important deviations from the protocol in one of the population, predictors and/or outcomes.
- Indirect – Important deviations from the protocol in at least two of the population, predictors and/or outcomes.

## Methods for combining association studies

Adjusted odds ratios, hazard ratios and risk ratios from multivariate models were only considered for pooling if the same set of predictor variables were used across multiple studies, the same thresholds to measure predictors were used across studies, and the studies controlled for the same confounding factors. This was not the case for any data in this guideline and so data was presented separately for individual studies.

## Modified GRADE for association studies

GRADE has not been developed for use with association studies; therefore a modified approach was applied using the GRADE framework. Data from cohort studies and cross sectional studies was initially rated as high quality, with the quality of the evidence for each outcome then downgraded or not from this initial point.

**Table 9: Rationale for downgrading quality of evidence for association studies**

| GRADE criteria | Reasons for downgrading quality |
|---|---|
| Risk of bias | Not serious: If the outcome was from a study judged at low risk of bias the outcome was not downgraded.<br>Serious: If the outcome was from a study judged at moderate risk of bias the outcome was downgraded one level.<br>Very serious: If the outcome was from a study that was judged at high risk of bias, the outcome was downgraded two levels. |
| Indirectness | Not serious: If the outcome was from a study judged directly applicable, the overall outcome was not downgraded.<br>Serious: If the outcome was from a study judged partially applicable, the outcome was downgraded one level.<br>Very serious: If the outcome was from a study judged indirectly applicable, the outcome was downgraded two levels. |
| Inconsistency | Results were not synthesised and were presented for individual studies. Inconsistency could therefore not be assessed as was rated as 'not applicable'. |
| Imprecision | The outcome 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. |

# Qualitative evidence

## Quality assessment

Individual qualitative studies were quality assessed using the CASP qualitative checklist. Each individual study was classified into one of the following three groups:

- Low risk of bias – The findings and themes identified in the study are likely to accurately capture the true picture.
- Moderate risk of bias – There is a possibility the findings and themes identified in the study are not a complete representation of the true picture.
- High risk of bias – It is likely the findings and themes identified in the study are not a complete representation of the true picture

Each individual study was also classified into one of three groups for relevance, based on if there were concerns about the perspective, population, phenomenon of interest and/or setting in the included studies and how directly these variables could address the specified review question. Studies were rated as follows:

- Highly relevant – No important deviations from the protocol in perspective, population, phenomenon of interest and/or setting.
- Relevant – Important deviations from the protocol in one of the perspective, population, phenomenon of interest and/or setting.
- Partially relevant – Important deviations from the protocol in at least two of the perspective, population, phenomenon of interest and/or setting.

## Methods for combining qualitative evidence

Only one qualitative study was included within the reviews. If multiple qualitative studies had been identified for a single question, then information from the studies would have been combined using a thematic synthesis. Instead, the main themes were extracted from the single study and were evaluated to examine their relevance to the review question. Each relevant theme was then presented to the committee.

## CERQual for qualitative studies

CERQual was used to assess the confidence we have in the summary findings of each of the identified themes. Evidence from all qualitative study designs (interviews, focus groups etc.) was initially rated as high confidence and the confidence in the evidence for each theme was then downgraded from this initial point as detailed in Table101 Rationale for downgrading confidence in evidence for qualitative questions

**Table101 Rationale for downgrading confidence in evidence for qualitative questions**

| CERQual criteria | Reasons for downgrading confidence |
|---|---|
| Methodological limitations | Not serious: If the theme was identified in studies at low risk of bias, the outcome was not downgraded<br>Serious: If the theme was identified only in studies at moderate or high risk of bias, the outcome was downgraded one level.<br>Very serious: If the theme was identified only in studies at high risk of bias, the outcome was downgraded two levels. |

| CERQual criteria | Reasons for downgrading confidence |
|---|---|
| Relevance | High: If the theme was identified in highly relevant studies, the outcome was not downgraded<br>Moderate: If the theme was identified only in relevant and partially relevant studies, the outcome was downgraded one level.<br>Low: If the theme was identified only in partially relevant studies, the outcome was downgraded two levels. |
| Coherence | Coherence was addressed based on two factors:<br>Between study – does the theme consistently emerge from all relevant studies<br>Theoretical – does the theme provide a convincing theoretical explanation for the patterns found in the data<br>The outcome was downgraded once if there were concerns about one of these elements of coherence, and twice if there were concerns about both elements. |
| Adequacy of data | The outcome was downgraded if there was insufficient data to develop an understanding of the phenomenon of interest, either due to insufficient studies, participants or observations. |

# Health economics

The methods for the de novo models built for this guideline can be found in the appendices of their respective evidence reviews. Literature reviews seeking to identify published cost–utility 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. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Economic studies identified through a systematic search of the literature are appraised using a methodology 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 10.

**Table 11 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 |

| Level | Explanation |
|---|---|
| | 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 11 Applicability criteria

**Table 12 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.