1.1 Priority screening

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. In every review, all of the identified abstracts were screened. 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.

1.2 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.

1.2.1 Quality assessment

Individual systematic reviews were quality assessed using the ROBIS or AMSTAR risk assessment 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.

1.2.2 Using systematic reviews as a source of data

If systematic reviews were identified as being sufficiently applicable and high quality, and were identified sufficiently early in the review process (for example, from the surveillance review or early in the database search), they were used as the primary source of data, rather than extracting information from primary studies. The extent to which this was done depended on the quality and applicability of the review, as defined in Table 1. When systematic reviews were used as a source of primary data, and unpublished or additional data included in the review which is not in the primary studies was also included. Data from these systematic reviews was then quality assessed and presented in GRADE/CERQual

tables as described below, 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 cross-referenced to ensure none of the data had been double counted through this process.

Table 1: Criteria for using systematic reviews as a source of data

Quality	Applicability	Use of systematic review
High	Fully applicable	Data from the published systematic review 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.
High	Partially applicable	Data from the published systematic review 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. For other sections not covered by the systematic review, 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 systematic review, searches were undertaken as normal.

1.3 Evidence synthesis and meta-analyses

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. For continuous outcomes analysed as standardised mean differences, where only baseline and final time point values were available, change from baseline standard deviations were estimated, assuming a correlation coefficient of 0.5.

1.4 Evidence of effectiveness of interventions

1.4.1 Quality assessment

Individual RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias 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.

1.4.2 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 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. 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). Both relative and absolute risks were presented, with absolute risks calculated by applying the relative risk to the pooled risk in the comparator arm of 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 $I^2 \geq 40\%$.

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.

Meta-analyses were performed in Cochrane Review Manager v5.3.

1.4.3 Minimal clinically important differences (MIDs)

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. In particular, any questions looking to evaluate non-inferiority (that one treatment is not meaningfully worse than another) required an MID to be defined to act as a non-inferiority margin. No MIDs were found through this process and used to assess imprecision in the guideline.

For standardised mean differences where no other MID was available, an MID of 0.2 was used, corresponding to the threshold for a small effect size initially suggested by Cohen et al. (1988). For relative risks where no other MID was available, a default MID interval for dichotomous outcomes of 0.8 to 1.25 was used.

When decisions were made in situations where MIDs were not available, the 'Evidence to Recommendations' section of that review should make explicit the committee's view of the expected clinical importance and relevance of the findings. In particular, this includes 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.

1.4.4 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 (2014)'. Data from all study designs was initially rated as high quality considering the best design for answering each research question. For example; in reviews that assessed risk factors, cohort studies were not downgraded and all started off as high-quality. The quality of the evidence for each outcome was subsequently downgraded or not from this initial point, based on the criteria given in Table 2

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, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies at high and low risk of bias.
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. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between direct and indirect studies.
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. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study.

GRADE criteria	Reasons for downgrading quality
	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. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies with the smallest and largest effect sizes.
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. If the line of no effect was defined as an MID for the outcome, it was downgraded once if the 95% confidence interval for the effect size crossed the line of no effect (i.e. the outcome was not statistically significant), and twice if the sample size of the study was sufficiently small that it is not plausible any realistic effect size could have been detected. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to clinically equivalent scenarios.

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.

1.4.5 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.

1.4.6 Evidence statements

Evidence statements for pairwise intervention data are classified in to one of four categories:

- Situations where the data are only consistent, at a 95% confidence level, with an effect in one direction (i.e. one that is 'statistically significant'), and the magnitude of that effect is most likely to meet or exceed the MID (i.e. the point estimate is not in the zone of equivalence). In such cases, we state that the evidence showed that there is an effect.
- Situations where the data are only consistent, at a 95% confidence level, with an effect in one direction (i.e. one that is 'statistically significant'), but the magnitude of that effect is most likely to be less than the MID (i.e. the point estimate is in the zone of equivalence). In such cases, we state that the evidence could not demonstrate a meaningful difference.
- Situations where the data are consistent, at a 95% confidence level, with an effect in either direction (i.e. one that is not 'statistically significant') but the confidence limits are smaller than the MIDs in both directions. In such cases, we state that the evidence demonstrates that there is no difference.
- In all other cases, we state that the evidence could not differentiate between the comparators.

1.5 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 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 schema above has the effect of setting a minimal important difference for positive likelihoods ratio at 2, and a corresponding minimal important difference 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.

1.5.1 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 two groups:

- Low risk of bias – Evidence of non-serious bias in zero or one domain.

- Moderate risk of bias – Evidence of non-serious bias in two domains only, or serious bias in one domain only.
- High risk of bias – Evidence of bias in at least three domains, or of serious bias in at least two domains.

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.

1.5.2 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 `metandi` package in STATA v13 or 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).

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.

1.5.3 Modified GRADE for diagnostic test accuracy evidence

GRADE has not been developed for use with diagnostic studies; therefore a modified approach was applied using the GRADE framework. GRADE assessments were only undertaken for positive and negative likelihood ratios, as the MIDs used to assess imprecision were based on these outcomes, but results for sensitivity and specificity are also presented alongside those data.

Cross-sectional and 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 4 below.

Table 4: Rationale for downgrading quality of evidence for diagnostic questions

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. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies at high and low risk of bias.
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. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between direct and indirect studies.
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. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. 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. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies with the smallest and largest effect sizes.
Imprecision	If the 95% confidence interval for a positive likelihood ratio spanned 2, the outcome was downgraded one level, as the data were deemed to be consistent with a meaningful increase in risk and no meaningful predictive value. Similarly, negative likelihood ratios that spanned 0.5 led to downgrading for serious imprecision. Any likelihood ratios that spanned both 0.5 and 2 were downgraded twice, as suffering from very serious imprecision. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to clinically equivalent scenarios.

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

- Data showing an effect size sufficiently large that it cannot be explained by confounding alone.
- Data where all plausible residual confounding is likely to increase our confidence in the effect estimate.

1.5.4 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.

1.5.5 Methods for combining inter-rater agreement evidence

The reliability of agreement for diagnostic data between observers was evaluated using the kappa coefficient. The measure calculates the level of agreement in classification. The general rule of thumb to follow is: if there is no agreement among the classification, then kappa ≤0; if there is complete agreement then kappa=1 (Fleiss 1971). The following schema (see Table 5), adapted from the suggestions of Fleiss, was used to interpret the level of agreement in diagnostic classification. Random-effects models (der Simonian and Laird) were fitted for all syntheses in R v3.4.0.

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.

Table 5: Interpretation of kappa coefficient

Value of kappa coefficients	Interpretation
$\kappa < 0$	No agreement
$0 < \kappa \leq 0.2$	Poor agreement
$0.2 < \kappa \leq 0.4$	Fair agreement
$0.4 < \kappa \leq 0.7$	Good agreement
$0.7 < \kappa < 1.0$	Excellent agreement
$\kappa = 1.0$	Complete agreement

1.5.6 Modified GRADE for inter-rater agreement evidence

GRADE has not been developed for use with inter-rater agreement; therefore a modified approach was applied using the GRADE framework. Data from all study types was initially rated as high quality, with the quality of the evidence for each outcome then downgraded or not from this initial point.

Table 6: Rationale for downgrading evidence for inter-rater agreement

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. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies at high and low risk of bias.
Inconsistency	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.

GRADE criteria	Reasons for downgrading quality
	Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between direct and indirect studies.
Indirectness	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. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. 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. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies with the smallest and largest effect sizes.
Imprecision	If the 95% confidence interval for the kappa coefficient spanned two of the categories in Table 5, it was downgraded one level. If the 95% confidence interval for the kappa coefficient spanned three or more of the categories in Table 5, it was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to clinically equivalent scenarios.

1.6 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/prognostic test accuracy). Depending on whether multivariable analysis was performed, data were reported as adjusted or unadjusted hazard ratios (if measured over time)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.

1.6.1 Quality assessment

Individual cohort and case-control studies were quality assessed using the CASP cohort study and case-control checklists, respectively. 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 – Evidence of non-serious bias in zero or one domain.
- Moderate risk of bias – Evidence of non-serious bias in two domains only, or serious bias in one domain only.
- High risk of bias – Evidence of bias in at least three domains, or of serious bias in at least two domains.

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.

1.6.2 Methods for combining association studies

Where appropriate, hazard ratios were pooled using the inverse-variance method, and odds ratios were pooled using the Mantel-Haenszel method. Adjusted odds ratios from multivariate models were only pooled if the same set of predictor variables were used across multiple studies and if the same thresholds to measure predictors were used across studies. However, this did not occur in practice. No odds ratios were pooled together due to different sets of predictor variables assessed across studies. Furthermore, studies adjusted for different confounders and dissimilar thresholds for measuring predictors were used across studies.

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 would need to be made and recorded before any data analysis is undertaken.
- The presence of significant statistical heterogeneity, defined as $I^2 \geq 40\%$.

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.

Meta-analyses were performed in Cochrane Review Manager v5.3.

1.6.3 Minimal clinically important differences (MIDs)

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. In particular, any questions looking to evaluate non-inferiority (that one treatment is not meaningfully worse than another) required an MID to be defined to act as a non-inferiority margin.

No MIDs were found through this process and used to assess imprecision in the guideline.

When decisions were made in situations where MIDs were not available, the 'Evidence to Recommendations' section of that review should make explicit the committee's view of the expected clinical importance and relevance of the findings.

1.6.4 Modified GRADE for association studies

GRADE has not been developed for use with predictive studies; therefore a modified approach was applied using the GRADE framework. Data from cohort studies was initially rated as high quality, and data from case-control studies as low quality, with the quality of the evidence for each outcome then downgraded or not from this initial point. However, this did not occur in practice as no case-control studies were included in any of the reviews.

Adjusted odds ratios from multivariate models could not be pooled because different sets of predictor variables were used across multiple studies and varying thresholds to measure predictors were adopted across studies. In the absence of meta-analyses, the following decision rules were used to assess risk of bias, indirectness, imprecision and inconsistency for each outcome:

Table 7: Rationale for downgrading quality of evidence for association studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	Risk of bias was calculated as normal, but using study weight by population size, rather than weight in the meta-analysis: Not serious: If less than 33.3% of the weight by population size 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 by population size 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 by population size came from studies at high risk of bias, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies at high and low risk of bias. In addition, unadjusted odds ratio outcomes from univariate analyses were downgraded one level, in addition to any downgrading for risk of bias in individual studies. Adjusted odds ratios from multivariate analyses were not similarly downgraded.
Indirectness	Indirectness was calculated as normal, but using study weight by population size, rather than weight in the meta-analysis: Not serious: If less than 33.3% of the weight by population size came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight by population size came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight by population size came from indirect studies, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between direct and indirect studies.

GRADE criteria	Reasons for downgrading quality
Inconsistency	Single study with or without 95% CI: N/A For multiple studies inconsistency was assessed by visually inspecting point estimates and 95% confidence intervals across included studies, taking into consideration sample sizes. When reported findings across studies highlighted inconsistent directions of effect, but there appeared to be an overall trend towards one direction, the evidence was downgraded once (serious). When reported findings across studies highlighted inconsistent directions of effect, but it was not possible to establish a trend towards one direction, the evidence was downgraded twice (very serious).
Imprecision	As no MIDs were identified from the COMET database, and none were specified by the committee, the line of no effect was defined as an MID for the outcome. For individual studies, imprecision 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. The evidence was also downgraded twice if 95% confidence intervals were not reported along with point estimates. When, multiple studies were identified for an outcome, imprecision was determined for each individual study (as above) and the overall imprecision was evaluated considering study taking sample size into consideration: • Not serious: If less than 33.3% of the weight by population size came from studies categorised as having no imprecision, the overall outcome was not downgraded. • Serious: If greater than 33.3% of the weight by population size came from studies categorised as having serious imprecision, the outcome was downgraded one level. • Very serious: If greater than 33.3% of the weight by population size came from studies with categorised as having very serious imprecision, the outcome was downgraded two levels.

1.6.5 Publication bias

Publication bias was not assessed because it was not possible to perform meta-analyses. Adjusted odds ratios from multivariate models could not be pooled because different sets of predictor variables were used across multiple studies and varying thresholds to measure predictors were adopted across studies.

1.6.6 Assessing c-statistics

C-statistics were assessed in a similar manner to likelihood ratios using the categories in **Error! Reference source not found.** below. Thresholds were set in line with those specified by Hosmer 2000.

Table 8 Interpretation of c-statistics

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

Meta-analyses could not be carried out as the data included large numbers of studies without 95% CI. In the absence of meta-analyses, 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. Minimal important difference (MID) levels of 0.7 and 0.8 were chosen for the assessment of imprecision, to be applied to the range of AUC scores reported across contributing studies (or to the 95% confidence interval where a model was evaluated by a single study).
    b. When evidence on the prognostic utility of a risk assessment tool was obtained from a single study, the evidence was downgraded one level (serious) if the 95% CI around an AUC crossed one MID, or two levels (very serious) if the 95% CI crossed both MIDs.
    c. When evidence on the prognostic utility of a risk assessment tool was obtained from more than one study, the evidence was downgraded one level (serious) if the AUC range crossed one MID, or two levels (very serious) if the AUC range crossed both MIDs
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.

1.7 Health economics

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 9.

Table 9 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 10.

Table 10 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.