National Institute for Health and Care Excellence

Draft for consultation

Osteoporosis: risk assessment

NICE guideline <number>

Methods

January 2026

Draft for Consultation

1

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN:

Contents

1

1. Development of the guideline

1.1. Remit

The remit for this guideline is to fully update the guideline on: Osteoporosis: assessing the risk of fragility fracture (CG146). In addition, although the current guideline only covers risk assessment the update will also include treatment to reduce primary and secondary fracture risk.

This guideline is being consulted on it two parts. This document relates to part one covering risk assessment up to and including determination of clinical appropriateness for treatment.

The methods outlined in this document relate to the following reviews:

- electronic health and social care records (including GP practice lists) to identify adults who should be assessed for fragility fracture risk
- risk assessment tools to predict risk of fragility fractures
- bone assessment methods to predict fragility fractures
- effectiveness of risk prediction tools and bone assessment methods
- diagnostic accuracy of vertebral fracture clinical decision tool (Vfrac) to identify who needs imaging to identify people with a suspected vertebral fracture
- effectiveness of Vfrac to identify people with a suspected vertebral fracture
- diagnostic accuracy of dual-energy X-ray absorptiometry (DXA) with vertebral fracture assessment (VFA) scan to identify vertebral fractures
- effectiveness of DXA with VFA scan to identify vertebral fractures
- automated imaging algorithms and computer-based diagnostics to identify vertebral fragility fractures
- monitoring for people at risk of fragility fracture who are not being treated pharmacologically

The methods outlined in the NICE guideline on osteoporosis (CG146 published 2012) in Appendix A relate to the following review:

- indications for identifying adults who should be assessed for fragility fracture risk

To see what the full guideline will cover and what this guideline does not cover, please see the guideline scope for the Osteoporosis: risk assessment, treatment and prevention of fragility fractures (update).

2. Methods

This guideline was developed using the methods described in the 2014 NICE guidelines manual, updated May 2024.

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

2.1. Developing the review questions and outcomes

The 11 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline development team and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- Population, Intervention, Comparator and Outcome (PICO) for reviews of interventions (including test and treat)
- Population, index test(s), reference standard and outcome for reviews of diagnostic and predictive accuracy
- Population, tests, and target conditions for reviews of risk prediction test accuracy.

Full literature searches, critical appraisals and evidence reviews were completed for all review questions except for the following:

- Risk factors for fragility fractures – consensus approach used to update recommendations from NICE guideline on osteoporosis (published 2012)
- Using artificial intelligence for identification of vertebral fractures – NICE published early value assessment guidance on artificial intelligence (AI) technologies to aid opportunistic detection of vertebral fragility fractures (NICE Health technology evaluation)

2.2. Reviewing research evidence

2.2.1. 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.

2.2.2. Searching for evidence

Evidence was searched for each review question using the methods specified in the 2014 NICE guidelines manual, updated May 2024.

2.2.3. Selecting studies for inclusion

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

with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

The full text of potentially eligible studies was retrieved and assessed according to the criteria specified in the review protocol. A standardised form was used to extract data from included studies.

2.3. Methods of combining evidence

2.3.1. Data synthesis for intervention studies

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome. Network meta-analyses were considered in situations where there were at least 3 treatment alternatives and sufficient studies to make this possible. When there were 2 treatment alternatives, pairwise meta-analysis was used to compare interventions.

2.3.1.1. Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.4. A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event. 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).

A pooled mean difference was calculated for continuous outcomes (using the inverse variance method) when the same scale was used to measure an outcome across different studies.

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

Fixed-effects models were the preferred choice to report, but in situations where the assumption of a shared mean for fixed-effects model were clearly not met, even after appropriate pre-specified subgroup analyses were conducted, random-effects results are presented. Fixed-effects models were deemed to be inappropriate if there was significant statistical heterogeneity in the meta-analysis, defined as I^2 more than or equal to 50%.

2.3.2. Data synthesis for diagnostic accuracy data

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

of true positives and false negatives (in people who, according to the reference standard, truly have the condition) and false positives and true negatives (in people who, according to the reference standard, do not).

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

- **Positive likelihood ratios** describe how many times more likely positive 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)
- **Positive predictive values** describe the probability that a person with a positive feature has the disease.
  - PPV = TP/ (TP+FP)
- **Negative predictive values** describe the probability that a person with a negative feature does not have the disease.
  - NPV = TN/(TN+FN)

Meta-analysis of diagnostic accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.1 (Deeks et al. 2022). Where three or more studies were available for all included strata, a bivariate model was fitted using WinBugs 14, which accounts for the correlations between positive and negative likelihood ratios, and between sensitivities and specificities. Where sufficient data were not available (<3 studies), sensitivity and specificity values were extracted from trial reports or calculated from the raw data and reported separately.

Commented [ES1]: Is this edit correct? I was a bit unsure what the end of the sentence meant

Commented [CJ2R1]: Yes thank you

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

2.3.3. Data synthesis for predictive accuracy data

For the purpose of this guideline predictive accuracy data are classified as any data in which people who have or don't have an index feature are observed to see who develops a condition or outcome of interest after a specified time. Prediction of a condition can consist in classification of individuals into those who will develop the condition and those who will not (as with traditional diagnostic test studies) or in estimation of an individual's risk of developing the condition. AUC/c-statistic and O:E

ratio data was extracted from the included studies or estimated (if not reported) in line with Debray 2017 and Debray 2018 respectively.

When deciding whether data should be synthesised or presented separately, heterogeneity in the population, index feature and outcome to be predicted were considered to determine whether data could be meaningfully combined. Meta-analysis of AUC/c-statistic and O:E ratio data was conducted when there were three or more studies that reported these measures using the package *metamisc* version 0.1.8 in R and the valmeta function, which performs a random-effects meta-analysis on studies (see Section 4.4, Appendix, Debray 2018). Meta-analysis was conducted using both a frequentist (restricted maximum likelihood ratio) and a Bayesian (rjags) model.

Commented [LF3]: Seems out of place as applies to review D

Commented [CJ4R3]: Deleted as reported in evidence review

2.4. Appraising the quality of evidence

2.4.1. Intervention studies (relative effect estimates)

RCTs were quality assessed using the Cochrane Risk of Bias Tool. Evidence on each outcome for each individual study was classified into one of the following groups:

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

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

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

2.4.1.1.1. Minimally important differences (MIDs) and clinical decision thresholds

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal clinically important difference thresholds relevant to this guideline that might aid the committee in identifying clinical decision thresholds for the purpose of GRADE.

For evidence review E on effectiveness of fragility fracture risk prediction tools and bone assessment methods, a published MID was identified and applied for the quality of life outcome assessed using the EQ-5D score.

For all other reviews, no appropriate published MIDs were identified. The guideline committee did not identify a consensus clinical decision threshold from their experience so default MIDs were used.

For relative risks and hazard ratios, where no other clinical decision threshold was available, a default clinical decision threshold for dichotomous outcomes of 0.8 to 1.25 was used. For the one continuous outcome included, an established MID was used (EQ-5D score in evidence review E).

2.4.1.1.2. GRADE for intervention studies analysed using pairwise analysis

GRADE was used to assess the quality of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials, 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. The quality of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 1. These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 1: Rationale for downgrading quality of evidence for intervention studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	• Not serious (no downgrade): less than 50% overall weighting some concerns/high risk of bias • Serious (downgrade 1 level): more than 50% some concerns/high risk of bias • Very serious (downgrade 2 levels): more than 50% high risk of bias.
Indirectness	• Not serious (no downgrade): less than 50% of overall weighting partially direct or indirect. • Serious (downgrade 1 level): more than 50% of overall weighting partially direct or indirect. • Very serious (downgrade 2 levels): more than 50% of overall weighting indirect
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. • Not serious (no downgrade): I^2 = less than 50%; • Serious (downgrade 1 level): I^2 = 50-75%; • Very serious (downgrade 2 levels): I^2 = more than 75%.
Imprecision	Where established MIDs were available these were used. Where there were no established MIDs, imprecision was assessed using the default values. The outcome was downgraded once if the 95% confidence interval for the effect size crossed one line of the MID, and twice if it crosses both lines of the MID.
Publication bias	Where 5 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

Commented [ES5]: I think the protocols say 5 studies

Commented [CJ6R5]: I have amended. Sorry, it has 10 in the new methods template but I think we used to do it for 5.

GRADE criteria	Reasons for downgrading quality
	publication bias was found for any outcomes in a review this domain was excluded from GRADE profiles to improve readability.

2.4.2. Diagnostic accuracy studies

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

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

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, index test 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.

Commented [CJ7]: Updated to QUADAS terminology - to be consistent with comment below.

2.4.2.1.1. GRADE for diagnostic accuracy evidence

Evidence from diagnostic accuracy studies was initially rated as high quality and then downgraded according to the standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 2 below.

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

In all cases, the downstream effects of diagnostic accuracy on patient-important outcomes were considered. This was done explicitly during committee deliberations and reported as part of the discussion section of the review detailing the likely consequences of true positive, true negative, false positive and false negative test results.

2.4.2.2. Using sensitivity and specificity as the primary outcomes

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

The committee were consulted to set 2 clinical decision thresholds for each measure: the value above which a test would be recommended, and a second below which a test would be considered of no clinical use. These values were used to judge imprecision (see below). If studies could not be pooled in a meta-analysis, GRADE

assessments were undertaken for each study individually and reported as separate lines in the GRADE profile.

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

Table 2: Rationale for downgrading quality of evidence for diagnostic accuracy data

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious (no downgrade): less than 50% overall weighting some concerns/high risk of bias Serious (downgrade 1 level): more than 50% some concerns/high risk of bias Very serious (downgrade 2 levels): more than 50% high risk of bias.
Indirectness	Not serious (no downgrade): less than 50% of overall weighting partially direct or indirect. Serious (downgrade 1 level): more than 50% of overall weighting partially direct or indirect. Very serious (downgrade 2 levels): more than 50% of overall weighting indirect.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. Where data was pooled it was checked visually to identify inconsistency. Where there are apparent differences in effect size due consideration was given to the appropriateness of pooling studies.
Imprecision	The most appropriate primary pair of measures (sensitivity/specificity) were used as described in the review protocols. Appropriate thresholds were discussed with the guideline committee and described within the evidence reviews.
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, this domain was excluded from GRADE profiles to improve readability.

2.4.3. Predictive accuracy studies

Studies that assessed the predictive accuracy of bone assessment methods to predict fragility fracture were quality assessed using an adapted version of the QUADAS-2 checklist**.** Studies that developed or assessed a prediction model were assessed using the PROBAST checklist.

Each individual study was classified into one of the following three groups:

- Low risk of bias – The true effect size for the study is likely to be close to the estimated effect size.

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

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

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

Commented [ES8]: Test and target condition for risk tools?

Commented [CJ9R8]: Updated according to QUADAS criteria

2.4.3.1. Modified GRADE for predictive accuracy data

GRADE has not been developed for use with predictive accuracy data, therefore a modified approach was applied using the GRADE framework. Evidence from cohort, cross sectional or case-control studies was initially rated as high quality and then assessed according to the same criteria as described in the section on standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 4 below.

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

GRADE assessments were only undertaken for AUC statistics for the prognostic accuracy reviews.

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

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

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

The following schema (Table 3) was used to interpret the AUC/c-statistic findings from both predictive accuracy reviews (Safari 2016).

Table 3: Interpretation of AUC/c-statistic findings

Value of AUC/c-statistic	Interpretation
≤0.50	Worse than chance
0.51–0.60	Very poor

Value of AUC/c-statistic	Interpretation
0.61–0.70	Poor
0.71–0.80	Moderate
0.81-0.90	Good
0.91-1.00	Excellent or perfect test

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

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious (no downgrade): less than 50% overall weighting some concerns/high risk of bias Serious (downgrade 1 level): more than 50% some concerns/high risk of bias Very serious (downgrade 2 levels): more than 50% high risk of bias.
Indirectness	Not serious (no downgrade): less than 50% of overall weighting partially direct or indirect. Serious (downgrade 1 level): more than 50% of overall weighting partially direct or indirect. Very serious (downgrade 2 levels): more than 50% of overall weighting indirect.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. **Fragility fracture risk prediction tools review (evidence review C)** When results from the internal validation study and external validation studies were available for a risk prediction tool, inconsistency was assessed by comparison of the point estimates and 95% confidence intervals. Inconsistency was assessed as very serious if the point estimates were on different sides of the clinical decision thresholds (0.5 and 0.7) and the 95% CIs did not overlap; when the point estimates were on the same side of the clinical decision thresholds (for example, all above 0.7) but the 95% CIs did not overlap, inconsistency was assessed as serious. Single studies were not downgraded for inconsistency. **Bone assessment methods review (evidence review D)** This was assessed using a combination of visual inspection of forest plots, and consideration of the I^2 and τ^2 statistics, as well as the width of the 95% prediction intervals. Outcomes were classified as having very serious inconsistency when visual inspection of forest plots indicated wide variation in point estimates and non-overlapping 95% CIs, high I^2 and high τ^2, as well as wide 95% prediction intervals. Inconsistency was assessed as serious for the same reasons but where τ^2 was low due to small standard error. This is because when studies are precise (that is, the standard errors of the AUC are small and the within-study variance is therefore small), a small τ^2 can lead to a high I^2 because the within-study variance is smaller than the between-study variance. Assessment of heterogeneity when I^2 is high but τ^2 is small can be tempered by consideration of the 95% prediction intervals, which estimates the range (that is, uncertainty)

GRADE criteria	Reasons for downgrading quality
	within which future studies may fall. Single studies were not downgraded for inconsistency.
Imprecision	Clinical decision thresholds were agreed with the committee in the context of the topic and type of outcome measure. Imprecision was assessed based on the 95% CI in relation to the clinical decision thresholds: • Not serious (no downgrade): CI does not cross either threshold • Serious (downgrade 1 level): CI crosses 1 threshold • Very serious (downgrade 2 levels): CI crosses 2 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.

2.4.3.2. Modified GRADE for risk prediction models

GRADE has not been developed for use with data from risk prediction models and a modified approach was therefore applied to assess the confidence in the overall estimates for the discriminatory power of the risk tools. GRADE was not conducted for discrimination at specific fracture risk thresholds. There is currently no guidance on how to assess the confidence in the evidence for other performance measures such as O:E ratios. Imprecision was assessed in the same way as for predictive accuracy data outlined above.

Heterogeneity was not assessed when there was only one study for a risk prediction tool. If there was more than one identified study, heterogeneity was assessed by consideration of the AUC 95% confidence intervals as specified.

2.5. Identifying and analysing evidence of cost effectiveness

The committee is required to make decisions based on the best available evidence of effectiveness and cost effectiveness. Guideline recommendations should be based on the expected costs of the different options in relation to their expected benefits (that is, their 'cost effectiveness') rather than the total implementation cost. However, as the cost of implementation increases, the committee needs to be increasingly confident in the cost effectiveness of a recommendation. Recommendations that are expected to have a significant impact on resources (as defined in the NICE Assessing resource impact process manual) need to be supported by robust evidence on effectiveness and cost effectiveness; any uncertainties must be offset by a compelling argument in favour of the recommendation. However, the cost impact or savings potential of a recommendation should not be the sole reason for the committee's decision (Developing NICE Guidelines: the manual).

Health economic evidence was gathered by:

- Undertaking systematics reviews of published economic literature.
- Conducting new analysis in priority areas.

2.6. Reviewing economic evidence

2.6.1. Identifying economic evidence

Systematic reviews of economic literature were conducted in all areas relevant for economic evaluation covered in the guideline. Titles and abstracts of articles identified through the systematic economic literature searches were assessed for inclusion using predefined eligibility criteria reported in the economic review protocol (provided in appendix A of each evidence review).

Once the screening of titles and abstracts was completed, full-text copies of potentially relevant articles were acquired for detailed assessment, applying the economic review protocol inclusion and exclusion criteria.

2.6.2. Appraising the quality of economic evidence

The applicability and methodological quality of economic evidence derived either from published studies meeting the inclusion criteria or from new economic analysis conducted for the guideline was assessed using the economic evaluations checklist specified in Developing NICE guidelines: the manual, Appendix H. This process led to applicability and quality statements for each included study, made by the health economist, following the criteria shown in Table 5.

Table 5: Criteria for developing applicability and quality statements of economic evidence

Appraised element	Statement and criteria
Applicability	• Directly applicable – the study meets all applicability criteria, or fails to meet 1 or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness. • Partially applicable – the study fails to meet 1 or more applicability criteria, and this could change the conclusions about cost effectiveness. • Not applicable – the study fails to meet 1 or more of the applicability criteria, and this is likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review.
Quality	• Minor limitations – the study meets all quality criteria, or fails to meet 1 or more quality criteria, but this is unlikely to change the conclusions about cost effectiveness. • Potentially serious limitations – the study fails to meet 1 or more quality criteria, and this could change the conclusions about cost effectiveness. • Very serious limitations – the study fails to meet 1 or more quality criteria, and this is highly likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review.

Health economic studies were prioritised for inclusion based on their relative applicability to the development of this guideline and the study limitations. For example, if a high quality, directly applicable UK analysis was available, then other less relevant studies may not have been included. For more detail about prioritisation see the health economics review protocol, which can be found in each of the

evidence reports. If exclusions occurred on this basis, this is noted in the relevant evidence report with reasons.

Details on methods and results of included economic studies are shown in economic evidence study extraction tables, provided in respective appendices of the economic reviews.

Characteristics and results (cost-effectiveness estimates) of economic studies used in decision making, including applicability and quality statements, have been summarised in economic evidence profile tables in the economic sections of the evidence report, as relevant.

2.7. New economic analysis

New economic analysis was undertaken by the guideline health economist in topic areas prioritised by the committee. The rationale for prioritising topic areas and/or specific review questions for economic modelling was set out in an economic plan agreed between members of the NICE technical team developing the guideline, the committee, and members of the NICE team quality assuring the guideline. New economic analysis was prioritised in areas with likely major resource implications, where the current extent of uncertainty over cost effectiveness was significant and economic analysis was expected to reduce this uncertainty. The process was completed for the full guideline. Analyses related to part one of the guideline only are covered below.

The criteria for BMD assessment with DXA was identified as a key health economic issue and a partial analysis was undertaken assessing potential differences in DXA resource use and numbers and characteristics of people identified for treatment with alternative strategies.

The following general principles were adhered to in developing the guideline cost-effectiveness analysis/es:

- Methods were consistent with the NICE reference case for interventions with health outcomes in NHS settings except where specified.
- The committee was involved in the design of the analysis and related assumptions, selection of inputs, discussion of limitations and interpretation of the results.
- Model inputs and assumptions were reported fully and transparently.
- The analysis was peer-reviewed by another health economist who was independent of the guideline development process.

Full methods and results of the analysis are described in Evidence report E.

Commented [LC10]: No longer a supplement

Commented [KL11R10]: revised

2.8. Cost effectiveness criteria

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

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of overall resource use and more effective compared with all other relevant alternative strategies)

- the intervention cost less than £20,000 per QALY gained compared with the next best strategy.

If the committee recommended an intervention that was estimated to cost more than £20,000 per QALY gained, or did not recommend one that was estimated to cost less than £20,000 per QALY gained, then the reasons for this decision were provided and recorded, with reference to issues around the plausibility of the estimate or to other factors, for example the degree of uncertainty around the ICER, aspects that relate to uncaptured benefits and non-health factors, or aspects that relate to health inequalities, as set out in the NICE health technology evaluations manual.

When new economic evidence was not available and new economic analysis was not prioritised, the committee made a qualitative judgement about cost effectiveness by considering expected differences in resource use and/or related UK NHS unit costs between options, alongside respective effectiveness evidence. Where possible, relevant UK NHS unit costs related to the compared interventions were presented to the committee (and listed under a 'Unit costs' section in the respective evidence review) to inform the possible economic implications of the recommendations.

The committee's considerations of cost effectiveness are discussed explicitly in the section 'Committee discussion and interpretation of the evidence' under the subheading 'Cost-effectiveness and resource use', in each evidence review.

2.9. References

1. Debray, Thomas PA; Damen, Johanna AAG; Riley, Richard D et al. (2018) A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Statistical Methods in Medical Research 28(9): 2768-2786.
2. Debray, Thomas PA; Damen, Johanna AAG; Snell, Kym IE et al. (2017) A guide to systematic review and meta-analysis of prediction model performance. BMJ 356: i6460.
3. Deeks 2010 Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy
4. Deeks 2022 Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.1
5. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3), 177-188.
6. NICE guidelines manual (2014) Developing-nice-guidelines-the-manual-2014.
7. Norman, G.; Sloan, JA; Wyrwich, KW. (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41(5):582-92.
8. Safari, S; Baratloo, A; Elfil, M et al. (2016) Evidence Based Emergency Medicine; Part 5 Receiver Operating Curve and Area under the Curve. Emerg (Tehran). Spring;4(2):111-3.

Commented [LF12]: Added debray 2017 and 2019 and amended some formatting

Commented [LF13]: Not sure these are right, quick search yields:
Deeks, J. J., Bossuyt, P. M., Leeflang, M. M., & Takwoingi, Y. (Eds.). (2023). *Cochrane handbook for systematic reviews of diagnostic test accuracy*. John Wiley & Sons.

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., & Takwoingi, Y. (2010). Cochrane handbook for systematic reviews of diagnostic test accuracy.

2.10. General terms

Term	Definition
Abstract	Summary of a study, which may be published alone or as an introduction to a full scientific paper.
Algorithm (in guidelines)	A flow chart of the clinical decision pathway described in the guideline, where decision points are represented with boxes, linked with arrows.
Allocation concealment	The process used to prevent advance knowledge of group assignment in an RCT. The allocation process should be impervious to any influence by the individual making the allocation, by being administered by someone who is not responsible for recruiting participants.
Applicability	How well the results of a study or NICE evidence review can answer a clinical question or be applied to the population being considered.
Arm (of a clinical study)	Subsection of individuals within a study who receive one particular intervention, for example placebo arm.
Association	Statistical relationship between 2 or more events, characteristics or other variables. The relationship may or may not be causal.
Base case analysis	In an economic evaluation, this is the main analysis based on the most plausible estimate of each input. In contrast, see Sensitivity analysis.
Baseline	The initial set of measurements at the beginning of a study (after runin period where applicable), with which subsequent results are compared.

Term	Definition
Bayesian analysis	A method of statistics, where a statistic is estimated by combining established information or belief (the 'prior') with new evidence (the 'likelihood') to give a revised estimate (the 'posterior').
Bias	Influences on a study that can make the results look better or worse than they really are. (Bias can even make it look as if a treatment works when it does not.) Bias can occur by chance, deliberately or as a result of systematic errors in the design and execution of a study. It can also occur at different stages in the research process, for example, during the collection, analysis, interpretation, publication or review of research data. For examples see selection bias, performance bias, information bias, confounding factor, and publication bias.
Blinding	A way to prevent researchers, doctors and patients in a clinical trial from knowing which study group each patient is in so they cannot influence the results. The best way to do this is by sorting patients into study groups randomly. The purpose of 'blinding' or 'masking' is to protect against bias. A single-blinded study is one in which patients do not know which study group they are in (for example whether they are taking the experimental drug or a placebo). A double-blinded study is one in which neither patients nor the researchers and doctors know which study group the patients are in. A triple blind study is one in which neither the patients, clinicians or the people carrying out the statistical analysis know which treatment patients received.
Carer (caregiver)	Someone who looks after family, partners or friends in need of help because they are ill, frail or have a disability.
Clinical efficacy	The extent to which an intervention is active when studied under controlled research conditions.
Clinical effectiveness	How well a specific test or treatment works when used in the 'real world' (for example, when used by a doctor with a patient at home), rather than in a carefully controlled clinical trial. Trials that assess clinical effectiveness are sometimes called management trials. Clinical effectiveness is not the same as efficacy.
Clinician	A healthcare professional who provides patient care. For example, a doctor, nurse or physiotherapist.
Cochrane Review	The Cochrane Library consists of a regularly updated collection of evidence-based medicine databases including the Cochrane Database of Systematic Reviews (reviews of randomised controlled trials prepared by the Cochrane Collaboration).
Comorbidity	A disease or condition that someone has in addition to the health problem being studied or treated.
Comparability	Similarity of the groups in characteristics likely to affect the study results (such as health status or age).
Concordance	This is a recent term whose meaning has changed. It was initially applied to the consultation process in which doctor and patient agree therapeutic decisions that incorporate their respective views, but now includes patient support in medicine taking as well as prescribing communication. Concordance reflects social values but does not address medicine-taking and may not lead to improved adherence.

Term	Definition
Confidence interval (CI)	A range of values for an unknown population parameter with a stated 'confidence' (conventionally 95%) that it contains the true value. The interval is calculated from sample data, and generally straddles the sample estimate. The 'confidence' value means that if the method used to calculate the interval is repeated many times, then that proportion of intervals will actually contain the true value.
Confounding factor	Something that influences a study and can result in misleading findings if it is not understood or appropriately dealt with. For example, a study of heart disease may look at a group of people that exercises regularly and a group that does not exercise. If the ages of the people in the 2 groups are different, then any difference in heart disease rates between the 2 groups could be because of age rather than exercise. Therefore, age is a confounding factor.
Consensus methods	Techniques used to reach agreement on a particular issue. Consensus methods may be used to develop NICE guidance if there is not enough good quality research evidence to give a clear answer to a question. Formal consensus methods include Delphi and nominal group techniques.
Control group	A group of people in a study who do not receive the treatment or test being studied. Instead, they may receive the standard treatment (sometimes called 'usual care') or a dummy treatment (placebo). The results for the control group are compared with those for a group receiving the treatment being tested. The aim is to check for any differences. Ideally, the people in the control group should be as similar as possible to those in the treatment group, to make it as easy as possible to detect any effects due to the treatment.
Cost–benefit analysis (CBA)	Cost–benefit analysis is one of the tools used to carry out an economic evaluation. The costs and benefits are measured using the same monetary units (for example, pounds sterling) to see whether the benefits exceed the costs.
Cost–consequences analysis (CCA)	Cost–consequences analysis is one of the tools used to carry out an economic evaluation. This compares the costs (such as treatment and hospital care) and the consequences (such as health outcomes) of a test or treatment with a suitable alternative. Unlike cost–benefit analysis or cost-effectiveness analysis, it does not attempt to summarise outcomes in a single measure (like the quality-adjusted life year) or in financial terms. Instead, outcomes are shown in their natural units (some of which may be monetary) and it is left to decision-makers to determine whether, overall, the treatment is worth carrying out.
Cost-effectiveness analysis (CEA)	Cost-effectiveness analysis is one of the tools used to carry out an economic evaluation. The benefits are expressed in non-monetary terms related to health, such as symptom-free days, heart attacks avoided, deaths avoided or life years gained (that is, the number of years by which life is extended as a result of the intervention).
Cost-effectiveness model	An explicit mathematical framework, which is used to represent clinical decision problems and incorporate evidence from a variety of sources in order to estimate the costs and health outcomes.

Term	Definition
Cost–utility analysis (CUA)	Cost–utility analysis is one of the tools used to carry out an economic evaluation. The benefits are assessed in terms of both quality and duration of life, and expressed as quality-adjusted life years (QALYs). See also utility.
Credible interval (CrI)	The Bayesian equivalent of a confidence interval.
Decision analysis	An explicit quantitative approach to decision-making under uncertainty, based on evidence from research. This evidence is translated into probabilities, and then into diagrams or decision trees which direct the clinician through a succession of possible scenarios, actions and outcomes.
Deterministic analysis	In economic evaluation, this is an analysis that uses a point estimate for each input. In contrast, see Probabilistic analysis
Discounting	Costs and perhaps benefits incurred today have a higher value than costs and benefits occurring in the future. Discounting health benefits reflects individual preference for benefits to be experienced in the present rather than the future. Discounting costs reflects individual preference for costs to be experienced in the future rather than the present.
Disutility	The loss of quality of life associated with having a disease or condition. See Utility
Dominance	A health economics term. When comparing tests or treatments, an option that is both less effective and costs more is said to be 'dominated' by the alternative.
Drop-out	A participant who withdraws from a trial before the end.
Economic evaluation	An economic evaluation is used to assess the cost effectiveness of healthcare interventions (that is, to compare the costs and benefits of a healthcare intervention to assess whether it is worth doing). The aim of an economic evaluation is to maximise the level of benefits – health effects – relative to the resources available. It should be used to inform and support the decision-making process; it is not supposed to replace the judgement of healthcare professionals. There are several types of economic evaluation: cost–benefit analysis, cost–consequences analysis, cost-effectiveness analysis, costminimisation analysis and cost–utility analysis. They use similar methods to define and evaluate costs, but differ in the way they estimate the benefits of a particular drug, programme or intervention.
Effect (as in effect measure, treatment effect, estimate of effect, effect size)	A measure that shows the magnitude of the outcome in one group compared with that in a control group. For example, if the absolute risk reduction is shown to be 5% and it is the outcome of interest, the effect size is 5%. The effect size is usually tested, using statistics, to find out how likely it is that the effect is a result of the treatment and has not just happened by chance (that is, to see if it is statistically significant).
Effectiveness	How beneficial a test or treatment is under usual or everyday conditions, compared with doing nothing or opting for another type of care.
Efficacy	How beneficial a test, treatment or public health intervention is under ideal conditions (for example, in a laboratory), compared with doing nothing or opting for another type of care.

Term	Definition
EQ-5D (EuroQol 5 dimensions)	A standardised instrument used to measure health-related quality of life. It provides a single index value for health status.
Evidence	Information on which a decision or guidance is based. Evidence is obtained from a range of sources including randomised controlled trials, observational studies, expert opinion (of clinical professionals or patients).
Exclusion criteria (literature review)	Explicit standards used to decide which studies should be excluded from consideration as potential sources of evidence.
Exclusion criteria (clinical study)	Criteria that define who is not eligible to participate in a clinical study.
Extended dominance	If Option A is both more clinically effective than Option B and has a lower cost per unit of effect, when both are compared with a do nothing alternative then Option A is said to have extended dominance over Option B. Option A is therefore cost effective and should be preferred, other things remaining equal.
Extrapolation	An assumption that the results of studies of a specific population will also hold true for another population with similar characteristics.
Follow-up	Observation over a period of time of an individual, group or initially defined population whose appropriate characteristics have been assessed in order to observe changes in health status or health related variables.
Generalisability	The extent to which the results of a study hold true for groups that did not participate in the research. See also external validity.
GRADE, GRADE evidence profile	A system developed by the GRADE Working Group to address the shortcomings of present grading systems in healthcare. The GRADE system uses a common, sensible and transparent approach to grading the quality of evidence. The results of applying the GRADE system to clinical trial data are displayed in a table known as a GRADE evidence profile.
Harms	Adverse effects of an intervention.
Hazard Ratio	The hazard or chance of an event occurring in the treatment arm of a study as a ratio of the chance of an event occurring in the control arm over time.
Health economics	Study or analysis of the cost of using and distributing healthcare resources.
Health-related quality of life (HRQoL)	A measure of the effects of an illness to see how it affects someone's day-to-day life.
Heterogeneity or Lack of homogeneity	The term is used in meta-analyses and systematic reviews to describe when the results of a test or treatment (or estimates of its effect) differ significantly in different studies. Such differences may occur as a result of differences in the populations studied, the outcome measures used or because of different definitions of the variables involved. It is the opposite of homogeneity.
Imprecision	Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of effect.
Inclusion criteria (literature review)	Explicit criteria used to decide which studies should be considered as potential sources of evidence.
Incremental analysis	The analysis of additional costs and additional clinical outcomes with different interventions.

Term	Definition
Incremental cost	The extra cost linked to using one test or treatment rather than another. Or the additional cost of doing a test or providing a treatment more frequently.
Incremental cost effectiveness ratio (ICER)	The difference in the mean costs in the population of interest divided by the differences in the mean outcomes in the population of interest for one treatment compared with another.
Incremental net benefit (INB)	The value (usually in monetary terms) of an intervention net of its cost compared with a comparator intervention. The INB can be calculated for a given cost-effectiveness (willingness to pay) threshold. If the threshold is £20,000 per QALY gained then the INB is calculated as: (£20,000 × QALYs gained) − Incremental cost.
Indirectness	The available evidence is different to the review question being addressed, in terms of PICO (population, intervention, comparison and outcome).
Intention-to-treat analysis (ITT)	An assessment of the people taking part in a clinical trial, based on the group they were initially (and randomly) allocated to. This is regardless of whether or not they dropped out, fully complied with the treatment or switched to an alternative treatment. Intention-to-treat analyses are often used to assess clinical effectiveness because they mirror actual practice: that is, not everyone complies with treatment and the treatment people receive may be changed according to how they respond to it.
Intervention	In medical terms this could be a drug treatment, surgical procedure, diagnostic or psychological therapy. Examples of public health interventions could include action to help someone to be physically active or to eat a healthier diet.
Intraoperative	The period of time during a surgical procedure.
Length of stay	The total number of days a participant stays in hospital.
Licence	See ‘Product licence’.
Life years gained	Mean average years of life gained per person as a result of the intervention compared with an alternative intervention.
Long-term care	Residential care in a home that may include skilled nursing care and help with everyday activities. This includes nursing homes and residential homes.
Logistic regression or Logit model	In statistics, logistic regression is a type of analysis used for predicting the outcome of a binary dependent variable based on one or more predictor variables. It can be used to estimate the log of the odds (known as the ‘logit’).
Loss to follow-up	A patient, or the proportion of patients, actively participating in a clinical trial at the beginning, but whom the researchers were unable to trace or contact by the point of follow-up in the trial
Markov model	A method for estimating long-term costs and effects for recurrent or chronic conditions, based on health states and the probability of transition between them within a given time period (cycle).
Meta-analysis	A method often used in systematic reviews. Results from several studies of the same test or treatment are combined to estimate the overall effect of the treatment.

Term	Definition
Multivariate model	A statistical model for analysis of the relationship between 2 or more predictor (independent) variables and the outcome (dependent) variable.
Net monetary benefit (NMB)	The value in monetary terms of an intervention net of its cost. The NMB can be calculated for a given cost-effectiveness threshold. If the threshold is £20,000 per QALY gained then the NMB for an intervention is calculated as: (£20,000 × mean QALYs) − mean cost. The most preferable option (that is, the most clinically effective option to have an ICER below the threshold selected) will be the treatment with the highest NMB.
Odds ratio	A measure of treatment effectiveness. The odds of an event happening in the treatment group, expressed as a proportion of the odds of it happening in the control group. The 'odds' is the ratio of events to non-events.
Opportunity cost	The loss of other healthcare programmes displaced by investment in or introduction of another intervention. This may be best measured by the health benefits that could have been achieved had the money been spent on the next best alternative healthcare intervention.
Outcome	The impact that a test, treatment, policy, programme or other intervention has on a person, group or population. Outcomes from interventions to improve the public's health could include changes in knowledge and behaviour related to health, societal changes (for example, a reduction in crime rates) and a change in people's health and wellbeing or health status. In clinical terms, outcomes could include the number of patients who fully recover from an illness or the number of hospital admissions, and an improvement or deterioration in someone's health, functional ability, symptoms or situation. Researchers should decide what outcomes to measure before a study begins.
P value	The p value is a statistical measure that indicates whether or not an effect is statistically significant. For example, if a study comparing 2 treatments found that one seems more effective than the other, the p value is the probability of obtaining these, or more extreme results by chance. By convention, if the p value is below 0.05 (that is, there is less than a 5% probability that the results occurred by chance) it is considered that there probably is a real difference between treatments. If the p value is 0.001 or less (less than a 1% probability that the results occurred by chance), the result is seen as highly significant. If the p value shows that there is likely to be a difference between treatments, the confidence interval describes how big the difference in effect might be.
Placebo	A fake (or dummy) treatment given to participants in the control group of a clinical trial. It is indistinguishable from the actual treatment (which is given to participants in the experimental group). The aim is to determine what effect the experimental treatment has had – over and above any placebo effect caused because someone has received (or thinks they have received) care or attention.

Term	Definition
Polypharmacy	The use or prescription of multiple medications.
Posterior distribution	In Bayesian statistics this is the probability distribution for a statistic based after combining established information or belief (the prior) with new evidence (the likelihood).
Power (statistical)	The ability to demonstrate an association when one exists. Power is related to sample size; the larger the sample size, the greater the power and the lower the risk that a possible association could be missed.
Preoperative	The period before surgery commences.
Prevalence	See Pre-test probability.
Prior distribution	In Bayesian statistics this is the probability distribution for a statistic based on previous evidence or belief.
Primary care	Healthcare delivered outside hospitals. It includes a range of services provided by GPs, nurses, health visitors, midwives and other healthcare professionals and allied health professionals such as dentists, pharmacists and opticians.
Primary outcome	The outcome of greatest importance, usually the one in a study that the power calculation is based on.
Probabilistic analysis	In economic evaluation, this is an analysis that uses a probability distribution for each input. In contrast, see Deterministic analysis.
Product licence	An authorisation from the MHRA to market a medicinal product.
Publication bias	Publication bias occurs when researchers publish the results of studies showing that a treatment works well and don't publish those showing it did not have any effect. If this happens, analysis of the published results will not give an accurate idea of how well the treatment works. This type of bias can be assessed by a funnel plot.
Quality of life	See 'Health-related quality of life'.
Quality-adjusted life year (QALY)	A measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life. One QALY is equal to 1 year of life in perfect health. QALYS are calculated by estimating the years of life remaining for a patient following a particular treatment or intervention and weighting each year with a quality of life score (on a scale of 0 to 1). It is often measured in terms of the person's ability to perform the activities of daily life, freedom from pain and mental disturbance.
Randomisation	Assigning participants in a research study to different groups without taking any similarities or differences between them into account. For example, it could involve using a random numbers table or a computer-generated random sequence. It means that each individual (or each group in the case of cluster randomisation) has the same chance of receiving each intervention.
Randomised controlled trial (RCT)	A study in which a number of similar people are randomly assigned to 2 (or more) groups to test a specific drug or treatment. One group (the experimental group) receives the treatment being tested, the other (the comparison or control group) receives an alternative treatment, a dummy treatment (placebo) or no treatment at all. The groups are followed up to see how effective the experimental treatment was. Outcomes are

Term	Definition
	measured at specific times and any difference in response between the groups is assessed statistically. This method is also used to reduce bias.
Rate ratio	The ratio of the rate of an event occurring among those exposed to certain conditions compared with the rate for those who are not exposed to the same conditions.
RCT	See 'Randomised controlled trial'.
Reporting bias	See 'Publication bias'.
Resource implication	The likely impact in terms of finance, workforce or other NHS resources.
Retrospective study	A research study that focuses on the past and present. The study examines past exposure to suspected risk factors for the disease or condition. Unlike prospective studies, it does not cover events that occur after the study group is selected.
Review question	In guideline development, this term refers to the questions about treatment and care that are formulated to guide the development of evidence-based recommendations.
Risk ratio (RR)	The ratio of the risk of disease or death among those exposed to certain conditions compared with the risk for those who are not exposed to the same conditions (for example, the risk of people who smoke getting lung cancer compared with the risk for people who do not smoke). If both groups face the same level of risk, the risk ratio is 1. If the first group had a risk ratio of 2, subjects in that group would be twice as likely to have the event happen. A risk ratio of less than 1 means the outcome is less likely in the first group. The risk ratio is sometimes referred to as relative risk.
Secondary outcome	An outcome used to evaluate additional effects of the intervention deemed a priori as being less important than the primary outcomes.
Selection bias	Selection bias occurs if: The characteristics of the people selected for a study differ from the wider population from which they have been drawn, or There are differences between groups of participants in a study in terms of how likely they are to get better.
Sensitivity analysis	A means of representing uncertainty in the results of economic evaluations. Uncertainty may arise from missing data, imprecise estimates or methodological controversy. Sensitivity analysis also allows for exploring the generalisability of results to other settings. The analysis is repeated using different assumptions to examine the effect on the results. One-way simple sensitivity analysis (univariate analysis): each parameter is varied individually in order to isolate the consequences of each parameter on the results of the study. Multi-way simple sensitivity analysis (scenario analysis): 2 or more parameters are varied at the same time and the overall effect on the results is evaluated. Threshold sensitivity analysis: the critical value of parameters above or below which the conclusions of the study will change are identified. Probabilistic sensitivity analysis: probability distributions are assigned to the uncertain parameters and are incorporated into

Term	Definition
	evaluation models based on decision analytical techniques (for example, Monte Carlo simulation).
Significance (statistical)	A result is deemed statistically significant if the probability of the result occurring by chance is less than 1 in 20 ($p<0.05$).
Stakeholder	An organisation with an interest in a topic that NICE is developing a guideline or piece of public health guidance on. Organisations that register as stakeholders can comment on the draft scope and the draft guidance. Stakeholders may be: • manufacturers of drugs or equipment national patient and carer organisations NHS organisations organisations representing healthcare professionals.
State transition model	See Markov model
Stratification	When a different estimate effect is thought to underlie two or more groups based on the PICO characteristics. The groups are therefore kept separate from the outset and are not combined in a metaanalysis, for example; children and adults. Specified a priori in the protocol.
Sub-groups	Planned statistical investigations if heterogeneity is found in the metaanalysis. Specified a priori in the protocol.
Systematic review	A review in which evidence from scientific studies has been identified, appraised and synthesised in a methodical way according to predetermined criteria. It may include a meta-analysis.
Time horizon	The time span over which costs and health outcomes are considered in a decision analysis or economic evaluation.
Transition probability	In a state transition model (Markov model), this is the probability of moving from one health state to another over a specific period of time.
Treatment allocation	Assigning a participant to a particular arm of a trial.
Univariate	Analysis which separately explores each variable in a data set.
Utility	In health economics, a 'utility' is the measure of the preference or value that an individual or society places upon a particular health state. It is generally a number between 0 (representing death) and 1 (perfect health). The most widely used measure of benefit in cost– utility analysis is the quality-adjusted life year, but other measures include disability-adjusted life years (DALYs) and healthy year equivalents (HYEs).

1
2