National Institute for Health and Care Excellence

Final

Spinal metastases and metastatic spinal cord compression

NICE guideline: methods

NICE guideline number NG234

Methods

September 2023

FINAL

These supplements were developed by NICE

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN: 978-1-4731-5325-7

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to update and replace the NICE guideline on metastatic spinal cord compression in adults: risk assessment, diagnosis and management (CG75).

To see “What this guideline covers” and “What this guideline does not cover” please see the guideline scope.

Methods

This guideline was developed using the methods described in the 2018 NICE guidelines manual.

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

Developing the review questions and outcomes

The review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NGA technical 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
- diagnostic reviews and reviews of prediction model accuracy – using population, index test, reference standard, target condition and outcome (PIRTO)
- qualitative reviews – using population, phenomenon of interest and context (PICo)

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

The review questions and evidence reviews corresponding to each question (or group of questions) are summarised below.

Table 1: Summary of review questions and index to evidence reviews

Evidence review	Review question	Type of review
[A] Service configuration & delivery (investigations)	What service configuration and delivery arrangements are effective for the investigation and referral of adults with suspected or confirmed spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention
[B] Service configuration & delivery (management and rehabilitation)	What service configuration and delivery arrangements are effective in the management and early rehabilitation of adults with suspected or confirmed spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention[1]
[C] Information and support	What information and support is valued by adults with suspected or confirmed spinal metastases, direct malignant infiltration of the spine, or associated spinal cord compression, and by their families and carers?	Qualitative
[D] Recognition - spinal metastases	What symptoms or signs, individually or in combination, or validated clinical tools, suggest the presence of spinal metastatic malignant disease or direct malignant infiltration of the spine?	Diagnostic
[E] Recognition - MSCC	What symptoms or signs, individually or in combination, or validated clinical tools, suggest spinal cord compression because of spinal	Diagnostic

Evidence review	Review question	Type of review
	metastatic malignancy or direct malignant infiltration of the spine?	
[F] Investigations - diagnosis	How effective are radiological imaging techniques in the diagnosis of spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention / Diagnostic
[G] Investigations - management	How effective are radiological imaging techniques in guiding the management of spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention
[H] Immobilisation	How effective are techniques or methods of immobilisation in managing spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention
[I] Analgesic interventions	How effective are analgesic interventions in managing pain related to spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention
[J] Corticosteroids	How effective is corticosteroid therapy in managing the neurological consequences of metastatic spinal cord compression?	Intervention
[K] Prognostic tools - spinal instability	What is the prognostic value of validated scoring systems in evaluating spinal metastases or direct malignant infiltration of the spine with or without spinal cord compression in relation to spinal instability?	Prediction model
[L] Prognostic tools - overall survival	What is the prognostic value of validated scoring systems for determining survival in people with spinal cord compression caused by spinal metastases or direct malignant infiltration of the spine?	Prediction model
[M] Radiotherapy	How effective is radiotherapy, including both fractionated and unfractionated radiotherapy, for the management of spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention
[N] Invasive interventions	How effective are invasive interventions, such as vertebroplasty, kyphoplasty, ablation and surgery, in managing spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?	Intervention[1]

[1] *Original health economic analysis conducted*

The COMET database was searched for core outcome sets relevant to this guideline. No core outcome sets were identified and therefore the outcomes were chosen based on committee discussions.

Additional information related to development of the guideline is contained in:

- Supplement 2 (Economics)
- Supplement 3 (NGA staff list).

Searching for evidence

Scoping search

During the scoping phase, searches were conducted for previous guidelines, economic evaluations, health technology assessments, systematic reviews and randomised controlled trials.

Systematic literature search

Systematic literature searches were undertaken to identify published evidence relevant to each review question.

Databases were searched using subject headings, free-text terms and, where appropriate, study type filters. Where possible, searches were limited to retrieve studies published in English. Limits to exclude animal studies, letters, editorials, news, conferences were applied where possible. All the searches were conducted in the following databases: Medline, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Epistemonikos, International Network of Agencies for Health Technology Assessment (INAHTA) and Embase. For review question C, Database of Abstracts of Reviews of Effects (DARE) and CINAHL were also searched. For review questions related to D, E, F, G, K, L, H, M and N CINAHL was also searched. For review questions F, G and J Emcare was also searched. For the review questions A and B, searches for grey literature were conducted in:

- British Spine Registry
- Canadian Agency for Drugs and Technologies in Health (CADTH)
- Getting It Right First Time (GIRFT)
- Google Scholar
- Metastatic Spinal Cord Compression, Acute Oncology Services and Cancer of Unknown Primary EAG, NHS England Midlands and East Cancer Expert Advisory Groups
- Neurosurgical National Audit Programme (NNAP)
- Somerset Cancer Register
- The Christie NHS Foundation Trust

Searches were run once for all reviews during development. Searches for the following review were updated in October 2022, 6 weeks in advance of the final committee meeting:

- H. How effective are techniques or methods of immobilisation in managing spinal metastases, direct malignant infiltration of the spine or associated spinal cord compression?

Details of the search strategies, including the study-design filters used and databases searched, are provided in Appendix B of each evidence review.

Economic systematic literature search

Systematic literature searches were also undertaken to identify published economic evidence. Databases were searched using subject headings, free-text terms and, where appropriate, an economic evaluations search filter.

A single search, using the population search terms used in the evidence reviews, was conducted to identify economic evidence in the NHS Economic Evaluation Database (NHS EED) and International Network of Agencies for Health Technology Assessments (INAHTA). Another single search, using the population search terms used in the evidence reviews combined with an economic evaluations search filter, was conducted in Medline, Cochrane Central Register of Controlled Trials (CENTRAL) and Embase. Where possible, searches were limited to studies published in English and to a date limit of 2005.

As with the general literature searches, the economic literature searches were updated in October 2022, 6 weeks in advance of the final committee meeting before consultation on the draft guideline.

Details of the search strategies, including the study-design filter used and databases searched, are provided in Supplement 2 (Health economics).

Quality assurance

Search strategies were quality assured by cross-checking reference lists of relevant studies, analysing search strategies from published systematic reviews and asking members of the committee to highlight key studies. The principal search strategies for each search were also quality assured by a second information scientist using an adaptation of the PRESS 2015 Guideline Evidence-Based Checklist (McGowan 2016).

Reviewing research evidence

Systematic review process

The evidence was reviewed in accordance with the following approach.

- Potentially relevant articles were identified from the search results for each review question by screening titles and abstracts. Full-text copies of the articles were then obtained.
- Full-text articles were reviewed against pre-specified inclusion and exclusion criteria in the review protocol (see Appendix A of each evidence review).
- Key information was extracted from each article on study methods and results, in accordance with factors specified in the review protocol. The information was presented in a summary table in the corresponding evidence review and in a more detailed evidence table (see Appendix D of each evidence review).
- Included studies were critically appraised using an appropriate checklist as specified in Developing NICE guidelines: the manual. Further detail on appraisal of the evidence is provided below.
- Summaries of quantitative evidence by outcome and qualitative evidence by theme were presented in the corresponding evidence review and discussed by the committee.

Any discrepancies were resolved by discussion between the first and second reviewers or by reference to a third (senior) reviewer. For the remaining review questions, internal (NGA) quality assurance processes included consideration of the outcomes of screening, study selection and data extraction and the committee reviewed the results of study selection and data extraction. The review protocol for each question specifies whether dual screening and study selection was undertaken

for that particular question. Drafts of all evidence reviews were quality assured by a senior reviewer.

Type of studies and inclusion/exclusion criteria

Inclusion and exclusion of studies was based on criteria specified in the corresponding review protocol.

Systematic reviews with meta-analyses or meta-syntheses were considered to be the highest quality evidence that could be selected for inclusion.

For intervention reviews, randomised controlled trials (RCTs) were prioritised for inclusion because they are considered to be the most robust type of study design that could produce an unbiased estimate of intervention effects. Where there was insufficient evidence from RCTs to inform guideline decision making, non-randomised studies (NRS) were considered for inclusion. Sufficiency was judged taking into account the number, quality and sample size of RCTs, as well as outcomes reported and availability of data from subgroups of interest. When NRS were considered for inclusion, priority was given to controlled studies, with separate control groups that were not allocated on the basis of the outcome, that adjusted for relevant confounders or matched participants on important confounding domains.

For diagnostic or prediction rule reviews, test-and-treat RCTs were prioritised for inclusion. In the absence of such studies, test accuracy studies were considered for inclusion.

For prognostic reviews, prospective and retrospective cohort and case–control studies and case series were considered for inclusion. Studies that included multivariable analysis were prioritised.

For qualitative reviews, studies using focus groups, structured interviews or semi-structured interviews were considered for inclusion. Where qualitative evidence was sought, data from surveys or other types of questionnaire were considered for inclusion only if they provided data from open-ended questions, but not if they reported only quantitative data.

The committee was consulted about any uncertainty regarding inclusion or exclusion of studies. A list of excluded studies for each review question, including reasons for exclusion is presented in Appendix J of the corresponding evidence review.

Narrative reviews, posters, letters, editorials, comment articles, unpublished studies and studies published in languages other than English were excluded. Conference abstracts were not considered for inclusion because conference abstracts typically do not have sufficient information to allow for full critical appraisal.

Methods of combining evidence

When planning reviews (through preparation of protocols), the following approaches for data synthesis were discussed and agreed with the committee.

Data synthesis for intervention studies

Pairwise meta-analysis

Meta-analysis to pool results from comparative intervention studies was conducted where possible using Cochrane Review Manager (RevMan5) software.

For dichotomous outcomes, such as mortality, the Mantel–Haenszel method with a fixed effect model was used to calculate risk ratios (RRs). For all outcomes with zero events in both arms the risk difference was presented. For outcomes in which the majority of studies had low event rates (<1%), Peto odds ratios (ORs) were calculated as this method performs well when events are rare (Bradburn 2007).

For continuous outcomes, measures of central tendency (mean) and variation (standard deviation; SD) are required for meta-analysis. Data for continuous outcomes, such as quality of life, were meta-analysed using an inverse-variance method for pooling weighted mean differences (WMDs). Where SDs were not reported for each intervention group, the SD of the mean difference was calculated from other reported statistics (p values or 95% confidence intervals; CIs) and then meta-analysis was conducted as described above.

If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into RevMan5. If the control event rate was reported this was used to generate the absolute risk difference in GRADEpro. If multivariable analysis was used to derive the summary statistic but no adjusted control event rate was reported, no absolute risk difference was calculated. Where a study reported multiple adjusted estimates for the same outcome, the one that minimised the risk of bias due to confounding was chosen.

When evidence was based on studies that reported descriptive data or medians with interquartile ranges or p values, this information was included in the corresponding GRADE tables (see below) without calculating relative or absolute effects. Consequently, certain aspects of quality assessment such as imprecision of the effect estimate could not be assessed as per standard methods for this type of evidence and subjective ratings or ratings based on sample size cut-offs were considered instead.

For some reviews, evidence was either stratified from the outset or separated into subgroups when heterogeneity was encountered. The stratifications and potential subgroups were pre-defined at the protocol stage (see the protocols for each review for further detail). Where evidence was stratified or subgrouped the committee considered on a case by case basis if separate recommendations should be made for distinct groups. Separate recommendations may be made where there is evidence of a differential effect of interventions in distinct groups. If there is a lack of evidence in one group, the committee considered, based on their experience, whether it was reasonable to extrapolate and assume the interventions will have similar effects in that group compared with others.

Data from RCTs and NRS, or from NRS with substantially different designs (i.e., cohort studies and case-control studies), that were theoretically possible to pool were entered into RevMan5 as subgroups based on study design. This was to take into account the likelihood of increased heterogeneity from studies with different design features and different approaches to appraising the quality of evidence based on study design (see appraising the quality of evidence: intervention studies below).

When meta-analysis was undertaken, the results were presented visually using forest plots generated using RevMan5 (see Appendix E of relevant evidence reviews).

Data synthesis for diagnostic test accuracy reviews

When diagnostic test accuracy was measured dichotomously, sensitivity, specificity, positive and negative predictive values and positive and negative likelihood ratios were used as outcomes. When diagnostic test accuracy was measured over multiple thresholds or when clinical prediction tools provided continuous estimates of risk, the area under the receiver-operating characteristic (ROC) curve (AUC) was used. These diagnostic test accuracy parameters were obtained directly from results reported in the source articles or calculated by the NGA technical team using data reported in the articles. Where possible, 95% CIs for diagnostic test accuracy parameters were reported; alternatively, median values and corresponding ranges were used if CIs were not reported and could not be calculated by the NGA technical team.

Meta-analysis of diagnostic test accuracy parameters was conducted if there was data from two or more studies that could be pooled and heterogeneity was not very serious.

When meta-analysis was undertaken, the study data were also presented in tables (see Appendix L of relevant evidence reviews).

Data synthesis for qualitative reviews

Where possible, a meta-synthesis was conducted to combine evidence from more than one study into a theme or sub-theme. Whenever studies identified a qualitative theme relevant to the protocol, this was extracted and the main characteristics were summarised. When all themes had been extracted from studies, common concepts were categorised and tabulated. This included information on how many studies had contributed to each theme identified by the NGA technical team.

The technical team were guided in their data extraction, synthesis and formulation of review findings, or themes, by a framework of phenomena developed by the guideline committee. This framework consisted of the themes that the committee anticipated would be covered by the included studies and these were set out a priori in the corresponding review protocol. The themes extracted from the data, however, were not limited to those set out in the review protocol.

Themes from individual studies were integrated into a wider context and, when possible, overarching categories of themes with sub-themes were identified. Themes were derived from data presented in individual studies. When themes were extracted from 1 primary study only, theme names used in the guideline mirrored those in the source study. However, when themes were based on evidence from multiple studies, the theme names were assigned by the NGA technical team. The names of overarching categories of themes were also assigned by the NGA technical team.

Emerging themes were placed into a thematic map representing the relationship between themes and overarching categories. The purpose of such a map is to show relationships between overarching categories and associated themes.

Appraising the quality of evidence

Intervention studies

GRADE methodology for intervention reviews

For intervention reviews, the evidence for outcomes from included RCTs and comparative non-randomised studies was evaluated and presented using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology developed by the international GRADE working group.

When GRADE was applied, software developed by the GRADE working group (GRADEpro) was used to assess the quality of each outcome, taking account of individual study quality factors and any meta-analysis results. Results were presented in GRADE profiles (GRADE tables).

The selection of outcomes for each review question was agreed during development of the associated review protocol in discussion with the committee. The evidence for each outcome was examined separately for the quality elements summarised in Table 3. Criteria considered in the rating of these elements are discussed below. Each element was graded using the quality ratings summarised in Table 4. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a 'serious' or 'very serious' quality issue. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 5.

The initial quality rating was based on the study design: RCTs and NRS assessed by ROBINS-I start as 'high' quality evidence, other non-randomised studies start as 'low' quality evidence. The rating was then modified according to the assessment of each quality element (Table 2). Each quality element considered to have a 'serious' or 'very serious' quality issue was downgraded by 1 or 2 levels respectively (for example, evidence starting as 'high' quality was downgraded to 'moderate' or 'low' quality). In addition, there was a possibility to upgrade evidence from non-randomised studies (provided the evidence for that outcome had not previously been downgraded) if there was a large magnitude of effect, a dose–response gradient, or if all plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results showed no effect.

Table 2: Summary of quality elements in GRADE for intervention reviews

Quality element	Description
Risk of bias ('Study limitations')	This refers to limitations in study design or implementation that reduce the internal validity of the evidence
Inconsistency	This refers to unexplained heterogeneity in the results
Indirectness	This refers to differences in study populations, interventions, comparators or outcomes between the available evidence and inclusion criteria specified in the review protocol
Imprecision	This occurs when a study has few participants or few events of interest, resulting in wide confidence intervals that cross minimally important thresholds

Quality element	Description
Publication bias	This refers to systematic under- or over-estimation of the underlying benefit or harm resulting from selective publication of study results

Table 3: GRADE quality ratings (by quality element)

Quality issues	Description
None or not serious	No serious issues with the evidence for the quality element under consideration
Serious	Issues with the evidence sufficient to downgrade by 1 level for the quality element under consideration
Very serious	Issues with the evidence sufficient to downgrade by 2 levels for the quality element under consideration

Table 4: Overall quality of the evidence in GRADE (by outcome)

Overall quality grading	Description
High	Further research is very unlikely to change the level of confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on the level of confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on the level of confidence in the estimate of effect and is likely to change the estimate
Very low	The estimate of effect is very uncertain

Assessing risk of bias in intervention reviews

Bias is a systematic error, or consistent deviation from the truth in results obtained. When a risk of bias is present the true effect can be either under- or over-estimated.

Risk of bias in RCTs was assessed using the Cochrane risk of bias tool (RoB 2; see Appendix H in Developing NICE guidelines: the manual).

The Cochrane risk of bias tool assesses the following possible sources of bias:

- risk of bias arising from the randomization process
- risk of bias due to deviations from the intended interventions
- risk of bias due to missing outcome data
- risk of bias due to measurement of the outcome
- risk of bias in selection of the reported result.

A study with a poor methodological design does not automatically imply high risk of bias; the bias is considered individually for each outcome and it is assessed whether the chosen design and methodology will impact on the estimation of the intervention effect.

More details about the Cochrane risk of bias tool can be found in Section 8 of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins 2022).

For systematic reviews the ROBIS checklist was used (see Appendix H in Developing NICE guidelines: the manual).

For non-randomised controlled studies, cohort studies or historical controlled studies the ROBINS-I checklist was used and for before-after studies the Effective Practice and Organisation of Care (EPOC) RoB Tool was used (see Appendix H in Developing NICE guidelines: the manual).

Assessing inconsistency in intervention reviews

Inconsistency refers to unexplained heterogeneity in results of meta-analysis. When estimates of treatment effect vary widely across studies (that is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only truly applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled). When outcomes were derived from a single study the rating 'no serious inconsistency' was used when assessing this domain, as per GRADE methodology (Santesso 2016).

Inconsistency was assessed visually by inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis (for example if the point estimates of the individual studies consistently showed benefits or harms). This was supported by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating serious heterogeneity, and more than 80% indicating very serious heterogeneity. When serious or very serious heterogeneity was observed, possible reasons were explored and subgroup analyses were performed as pre-specified in the review protocol where possible. In the case of unexplained heterogeneity, sensitivity analyses were planned based on the quality of studies, eliminating studies at high risk of bias (in relation to randomisation, allocation concealment and blinding, and/or missing outcome data).

When no plausible explanation for the serious or very serious heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency and the meta-analysis was re-run using the Der-Simonian and Laird method with a random effects model and this was used for the final analysis. For outcomes with unexplained very serious heterogeneity the range of effects was reported rather than the pooled value.

Assessing indirectness in intervention reviews

Directness refers to the extent to which populations, interventions, comparisons and outcomes reported in the evidence are similar to those defined in the inclusion criteria for the review and was assessed by comparing the PICO elements in the studies to the PICO defined in the review protocol. Indirectness is important when such differences are expected to contribute to a difference in effect size or may affect the balance of benefits and harms considered for an intervention.

Assessing imprecision and importance in intervention reviews

Imprecision in GRADE methodology refers to uncertainty around the effect estimate and whether or not there is an important difference between interventions (that is, whether the evidence clearly supports a particular recommendation or appears to be consistent with several candidate recommendations). Therefore, imprecision differs from other aspects of evidence quality because it is not concerned with whether the point estimate is accurate or correct (has internal or external validity). Instead, it is concerned with uncertainty about what the point estimate actually represents. This uncertainty is reflected in the width of the CI.

The 95% CI is defined as the range of values within which the population value will fall on 95% of repeated samples, were the procedure to be repeated. The larger the study, the smaller the 95% CI will be and the more certain the effect estimate.

Imprecision was assessed in the guideline evidence reviews by considering whether the width of the 95% CI of the effect estimate was relevant to decision making, considering each outcome independently. This is illustrated in Figure 1, which considers a positive outcome for the comparison of two treatments. Three decision-making zones can be differentiated, bounded by the thresholds for minimal importance (minimally important differences; MIDs) for benefit and harm.

When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

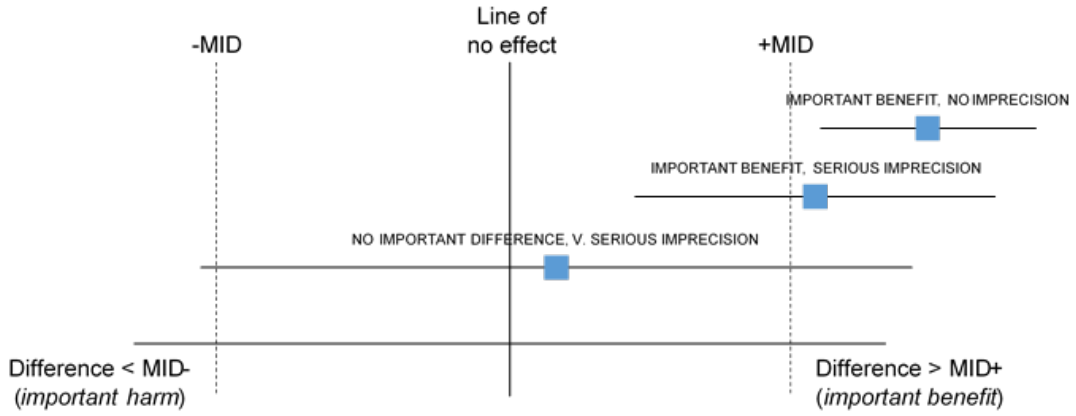
When the CI crosses 2 zones, it is uncertain in which zone the true value of the effect estimate lies and therefore there is uncertainty over which decision to make. The CI is consistent with 2 possible decisions, therefore, the effect estimate is considered to be imprecise in the GRADE analysis and the evidence is downgraded by 1 level ('serious imprecision').

When the CI crosses all 3 zones, the effect estimate is considered to be very imprecise because the CI is consistent with 3 possible decisions and there is therefore a considerable lack of confidence in the results. The evidence is therefore downgraded by 2 levels in the GRADE analysis ('very serious imprecision').

If a single-sided CI was reported then the evidence could only be downgraded by one-level.

Implicitly, assessing whether a CI is in, or partially in, an important zone, requires the guideline committee to estimate an MID or to say whether they would make different decisions for the 2 confidence limits.

Figure 1: Assessment of imprecision and importance in intervention reviews using GRADE



MID, minimally important difference

Defining minimally important differences for intervention reviews

The committee was asked whether there were any recognised or acceptable MIDs in the published literature and community relevant to the review questions under

consideration. The committee was not aware of any MIDs that could be used for the guideline.

In the absence of published or accepted MIDs, the committee agreed to use the GRADE default MIDs to assess imprecision. For dichotomous outcomes minimally important thresholds for a RR of 0.8 and 1.25 respectively were used as default MIDs in the guideline. The committee also chose to use 0.8 and 1.25 as the MIDs for ORs & HRs in the absence of published or accepted MIDs. ORs were predominantly used in the guideline when Peto OR were indicated due to low event rates, at low event rates OR are mathematically similar to RR making the extrapolation appropriate. While no default MIDs exist for HR, the committee agreed for consistency to continue to use 0.8 and 1.25 for these outcomes.

If risk difference was used for meta-analysis, for example if the majority of studies had zero events in either arm, imprecision was assessed based on sample size using 100 and 300 as cut-offs for very serious and serious imprecision respectively. The committee used these numbers based on commonly used optimal information size thresholds.

The same thresholds were used as default MIDs in the guideline for all dichotomous outcomes considered in intervention evidence reviews. For continuous outcomes default MIDs are equal to half the median SD of the control groups at baseline (or at follow-up if the SD is not available a baseline). Where results were reported as medians imprecision was assessed based on sample size using 100 and 300 as cut-offs for very serious and serious imprecision respectively.

Assessing publication bias in intervention reviews

As there were fewer than 10 studies were included for any of the single meta-analysis done for the guideline, the committee subjectively assessed the likelihood of publication bias based on factors such as the proportion of trials funded by industry and the propensity for publication bias in the topic area.

Diagnostic studies

Adapted GRADE methodology for diagnostic and prognostic prediction model reviews

For diagnostic reviews and prediction models, an adapted GRADE approach was used. GRADE methodology is designed for intervention reviews, but the quality assessment elements and outcome presentation were adapted by the guideline developers for diagnostic test accuracy reviews and prediction models. For example, GRADE tables were modified to include diagnostic test accuracy measures (sensitivity, specificity, likelihood ratios and predictive values) depending on the review question.

Although the modified GRADE tables included multiple diagnostic parameters, the quality rating was based on a single pair of parameters prioritised by the guideline committee. The evidence for the chosen parameters was examined separately for the quality elements listed and defined in Table 5. The quality rating was based on the positive and negative likelihood ratios for evidence reviews F and K, but for evidence review D the quality rating was based on positive predictive value only. This was because evidence review D was concerned with the predictive value of signs and symptoms and the main evidence of interest was therefore for people with the sign or symptom of interest.

Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a 'serious' or 'very serious' quality issue. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 5.

The initial quality rating was based on the study design: cross-sectional or cohort studies start as 'high' quality and case–control studies start as 'low' quality.

Table 5: Adaptation of GRADE quality elements for diagnostic reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in study design and implementation may bias estimates of diagnostic accuracy. High risk of bias for the majority of the evidence reduces confidence in the estimated effect. Diagnostic accuracy studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Inconsistency	This refers to unexplained heterogeneity in test accuracy measures (such as sensitivity and specificity) between studies
Indirectness	This refers to differences in study populations, index tests, reference standards or outcomes between the available evidence and inclusion criteria specified in the review protocol
Imprecision	This occurs when a study has relatively few participants and the probability of a correct diagnosis is low. Accuracy measures would therefore have wide confidence intervals around the estimated effect

Assessing risk of bias in diagnostic reviews

Risk of bias in diagnostic reviews was assessed using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklist (see Appendix H in Developing NICE guidelines: the manual).

Risk of bias in primary diagnostic accuracy reviews in QUADAS-2 consists of 4 domains:

- participant selection
- index test
- reference standard
- flow and timing.

More details about the QUADAS-2 tool can be found on the developer's website.

Assessing risk of bias in prognostic prediction model reviews

Risk of bias in reviews of prediction models was assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) checklist (see Appendix H in Developing NICE guidelines: the manual).

Risk of bias in prediction model reviews in PROBAST consists of 4 domains:

- participant selection
- predictors or their assessment
- outcome or its determination
- analysis.

For details about the PROBAST tool see Wolff (2019).

Assessing inconsistency in diagnostic and prognostic prediction model reviews

Inconsistency refers to the unexplained heterogeneity of the results in meta-analysis. When estimates of diagnostic accuracy and prediction model parameters vary widely across studies (that is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only truly applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled).

Inconsistency for diagnostic reviews and prediction models was assessed based on visual inspection of the point estimates and confidence intervals of the included studies. If these varied widely (for example, point estimates for some studies lying outside the CIs of other studies) the evidence was downgraded for inconsistency.

Assessing indirectness in diagnostic reviews

Indirectness in diagnostic reviews was assessed using the QUADAS-2 checklist by assessing the applicability of the studies in relation to the review question in the following domains:

- participant selection
- index test
- reference standard.

More details about the QUADAS-2 tool can be found on the developer's website.

Assessing indirectness in prognostic prediction model reviews

Indirectness in prognostic prediction model reviews was assessed using the PROBAST checklist by assessing the applicability of the studies in relation to the review question in the following domains:

- participant selection
- predictors or their assessment
- outcome or its determination

For details about the PROBAST tool see Wolff (2019).

Assessing imprecision and importance in diagnostic and prognostic prediction model reviews

The judgement of precision for diagnostic and prediction model evidence was based on the CIs of the single pair of parameters prioritised by the guideline committee. The judgement of precision was based on the positive and negative likelihood ratios for evidence reviews F and K, but for evidence review D precision was based on positive predictive value (PPV) only. As the committee agreed a single threshold value for PPV of 3% this meant that the evidence for review D would be imprecise if the 95% CI of PPV included 3% or was otherwise precise.

The following thresholds were used when summarising the performance of diagnostic tests or prediction models in terms of likelihood ratios:

- useful test: LR+ ≥ 5.0, LR- ≤0.2
- not a useful test: LR+ between 1.0 and 2.0, LR- between 0.5 and 1.0.

The above thresholds were also used to assess imprecision of likelihood ratios.

The committee agreed 2 decision thresholds for sensitivity and specificity, a value above which the test could be recommended and a value below which the test would

be considered of no use. These thresholds were based on the committee's experience and consensus.

The thresholds were:

- sensitivity: low threshold 50%, high threshold 80%
- specificity: low threshold 50%, high threshold 80%.

The above thresholds were also used to assess imprecision of sensitivity and specificity.

The following cut-offs were used when summarising the performance of diagnostic tests or prediction models in terms of AUC:

- excellent: 0.90
- moderate: 0.70
- the test performs at chance: 0.50.

The above thresholds were also used to assess imprecision of sensitivity and specificity.

The following cut-off was used when summarising the performance of symptoms or signs in terms of positive predictive values (PPVs) for primary care studies where only people with the symptom or sign were included:

- important symptom or sign: PPV 3% or greater.

This threshold was used to asses the imprecision of PPV and as a consequence PPV outcomes could only be downgraded one level for imprecision.

Outcomes were downgraded for imprecision when their 95% CI crossed at least 1 threshold. If the CI crossed 1 threshold, the outcome was downgraded one level for imprecision. If the CI crossed 2 thresholds, the outcome was downgraded two levels for imprecision. These assessments were made on the meta-analysed outcomes where applicable or if outcomes were not meta-analysed, on the individual study results themselves.

In evidence review L some studies used prediction models for overall survival to divide patients into groups according to their predicted length of survival. For each survival group the calibration accuracy (proportion of patients whose observed survival matched their predicted survival) was summarised according to the following cut-offs:

- excellent: 90%
- moderate: 70%.

Assessment of imprecision was based on sample size criteria for this outcome, downgraded one level for sample size < 100 and by 2 levels for sample size < 300.

Qualitative studies

GRADE-CERQual methodology for qualitative reviews

For qualitative reviews an adapted GRADE Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) approach (Lewin 2018) was used. In this approach the quality of evidence is considered according to themes in the evidence. The themes may have been identified in the primary studies or they may have been identified by considering the reports of a number of studies. Quality

elements assessed using GRADE-CERQual are listed and defined in Table 6. Each element was graded using the levels of concern summarised in Table 7.

The ratings for each component were combined (as with other types of evidence) to obtain an overall assessment of quality for each theme as described in Table 8. 'Confidence' in this context refers to the extent to which the review finding is a reasonable representation of the phenomenon of interest set out in the protocol. Similar to other types of evidence all review findings start off with 'high confidence' and are rated down by one or more levels if there are concerns about any of the individual CERQual components. In line with advice from the CERQual developers, the overall assessment does not involve numerical scoring for each component but in order to ensure consistency across and between guidelines, the NGA established some guiding principles for overall ratings. For example, a review finding would not be downgraded (and therefore would be assessed with 'high' confidence) if at least 2 of the individual components were rated as 'no or very minor; and none of the components were rated as having moderate or serious concerns.

At the other extreme, a review finding would be downgraded 3 times (to 'very low') if at least 2 components had serious concerns or 3 had moderate concerns (as long as the 4th component was rated 'serious') or if all components had moderate concerns . A basic principle was that if any components had any serious concerns then overall confidence in the review finding would be downgraded at least twice, to low. Transparency about overall judgements is provided in the CERQual tables, with explanations for downgrading given in the individual domain cells.

Table 6: Adaptation of GRADE quality elements for qualitative reviews

Quality element	Description
Methodological limitations	Limitations in study design and implementation may bias interpretation of qualitative themes identified. High risk of bias for the majority of the evidence reduces our confidence that the review findings reflect the phenomena of interest. Qualitative studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Relevance (or applicability) of evidence	This refers to the extent to which the context of the studies supporting the review findings is applicable to the context specified in the review question
Coherence of findings	This refers to the extent to which review findings are well grounded in data from the contributing primary studies and provide a credible explanation for patterns identified in the evidence. If the data from the underlying studies are ambiguous or contradict the review finding this would reduce our confidence in the finding.
Adequacy of data (theme saturation or sufficiency)	This corresponds to a similar concept in primary qualitative research, that is, whether a theoretical point of theme saturation was achieved, at which point no further citations or observations would provide more insight or suggest a different interpretation of the particular theme. Judgements are not based on the number of studies but do take account of the quantity and also richness of data underpinning a finding. The more complex the finding, the more detailed the supporting data need to be. For simple findings, relatively superficial data would be considered adequate to explain and explore the phenomenon being described.

Table 7: CERQual levels of concern (by quality element)

Level of concern	Definition
None or very minor concerns	Unlikely to reduce confidence in the review finding
Minor concerns	May reduce confidence in the review finding
Moderate concerns	Will probably reduce confidence in the review finding
Serious concerns	Very likely to reduce confidence in the review finding

Table 8: Overall confidence in the evidence in CERQual (by review finding)

Overall confidence level	Definition
High	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest
Moderate	It is likely that the review finding is a reasonable representation of the phenomenon of interest
Low	It is possible that the review finding is a reasonable representation of the phenomenon of interest
Very low	It is unclear whether the review finding is a reasonable representation of the phenomenon of interest

Assessing methodological limitations in qualitative reviews

Methodological limitations in qualitative studies were assessed using the Critical Appraisal Skills Programme (CASP) checklist for qualitative studies (see appendix H in Developing NICE guidelines: the manual). Overall methodological limitations were derived by assessing the methodological limitations across the 6 domains summarised in Table 9.

Table 9: Methodological limitations in qualitative studies

Aim and appropriateness of qualitative evidence	This domain assesses whether the aims and relevance of the study were described clearly and whether qualitative research methods were appropriate for investigating the research question
Rigour in study design or validity of theoretical approach	This domain assesses whether the study approach was documented clearly and whether it was based on a theoretical framework (such as ethnography or grounded theory). This does not necessarily mean that the framework has to be stated explicitly, but a detailed description ensuring transparency and reproducibility should be provided
Sample selection	This domain assesses the background, the procedure and reasons for the method of selecting participants. The assessment should include consideration of any

	relationship between the researcher and the participants, and how this might have influenced the findings
Data collection	This domain assesses the documentation of the method of data collection (in-depth interviews, semi-structured interviews, focus groups or observations). It also assesses who conducted any interviews, how long they lasted and where they took place
Data analysis	This domain assesses whether sufficient detail was documented for the analytical process and whether it was in accordance with the theoretical approach. For example, if a thematic analysis was used, the assessment would focus on the description of the approach used to generate themes. Consideration of data saturation would also form part of this assessment (it could be reported directly or it might be inferred from the citations documented that more themes could be found)
Results	This domain assesses any reasoning accompanying reporting of results (for example, whether a theoretical proposal or framework is provided)

Assessing relevance of evidence in qualitative reviews

Relevance (applicability) of findings in qualitative research is the equivalent of indirectness for quantitative outcomes, and refers to how closely the aims and context of studies contributing to a theme reflect the objectives outlined in the guideline review protocol.

Assessing coherence of findings in qualitative reviews

For qualitative research, a similar concept to inconsistency is coherence, which refers to the way findings within themes are described and whether they make sense. This concept was used in the quality assessment across studies for individual themes. This does not mean that contradictory evidence was automatically downgraded, but that it was highlighted and presented, and that reasoning was provided. Provided the themes, or components of themes, from individual studies fit into a theoretical framework, they do not necessarily have to reflect the same perspective. It should, however, be possible to explain these by differences in context (for example, the views of health or social care professionals might not be the same as those of family members, but they could contribute to the same overarching themes).

Assessing adequacy of data in qualitative reviews

Adequacy of data (theme saturation or sufficiency) corresponds to a similar concept in primary qualitative research in which consideration is made of whether a theoretical point of theme saturation was achieved, meaning that no further citations or observations would provide more insight or suggest a different interpretation of the theme concerned. As noted above, it is not equivalent to the number of studies contributing to a theme, but it does take account of the quantity of data supporting a

review finding (for instance whether sufficient quotations or observations were provided to underpin the findings) and in particular the degree of 'richness' of supporting data. Concerns about richness arise when insufficient details are provided by the data to enable an understanding of the phenomenon being described. Generally, if a review finding is fairly simple then relatively superficial data will be needed to understand it. Data underpinning a more complex finding would need to offer greater detail, allowing for interpretation and exploration of the phenomenon being described. Therefore in assessing adequacy our downgrading involved weighing up the complexity of the review finding against the explanatory contribution of the supporting data.

Reviewing economic evidence

Titles and abstracts of articles identified through the economic literature searches were independently assessed for inclusion using the predefined inclusion and exclusion criteria

Inclusion and exclusion criteria for systematic reviews of economic evaluations

Inclusion and exclusion criteria

- If a study is rated as both 'Directly applicable' and with 'Minor limitations' then it was included in the guideline. A health economic evidence table was completed and it was included in the health economic evidence profile.
- If a study is rated as either 'Not applicable' or with 'Very serious limitations' then it was excluded from the guideline. If it is excluded then a health economic evidence table was not be completed and it was not be included in the health economic evidence profile.
- If a study is rated as 'Partially applicable', with 'Potentially serious limitations' or both then discretion was used over whether it should be included.

Where there is discretion

The health economist made a decision based on the relative applicability and quality of the available evidence for that question, in discussion with the guideline committee if required. The ultimate aim was to include health economic studies that are helpful for decision-making in the context of the guideline and the current NHS setting.

The health economist was guided by the following hierarchies.

Setting:

- UK NHS (most applicable).
- OECD countries with predominantly public health insurance systems (for example, France, Germany, Sweden).
- OECD countries with predominantly private health insurance systems (for example, Switzerland).
- Studies set in non-OECD countries or in the USA were excluded before being assessed for applicability and methodological limitations.

Health economic study type:

- Cost–utility analysis (most applicable).

- Other type of full economic evaluation (cost–benefit analysis, cost-effectiveness analysis, cost–consequences analysis).
- Comparative cost analysis.
- Non-comparative cost analyses including cost-of-illness studies were excluded before being assessed for applicability and methodological limitations.

Year of analysis:

- The more recent the study, the more applicable it will be.
- Studies published in 2005 or later (including any such studies included in the previous guideline(s)) but that depend on unit costs and resource data entirely or predominantly from before 2005 was rated as 'Not applicable'.
- Studies published before 2005 (including any such studies included in the previous guideline(s)) was excluded before being assessed for applicability and methodological limitations.

Appraising the quality of economic evidence

The quality of economic evidence was assessed using the economic evaluations checklist specified in Developing NICE guidelines: the manual.

Economic modelling

The aims of the economic input to the guideline were to inform the guideline committee of potential economic issues to ensure that recommendations represented a cost effective use of healthcare resources. Economic evaluations aim to integrate data on healthcare benefits (ideally in terms of quality-adjusted life-years; QALYs) with the costs of different options. In addition, the economic input aimed to identify areas of high resource impact; these are recommendations which (while cost effective) might have a large impact on Clinical Commissioning Group / Integrated Care Board or Trust finances and so need special attention.

The guideline committee prioritised the following review questions for economic modelling where it was thought that economic considerations would be particularly important in formulating recommendations.

- [B] Service configuration & delivery (management and rehabilitation)
- [N] Invasive interventions.

The methods and results of the de novo economic analyses are reported in Appendix I of the relevant evidence reports. When new economic analysis was not prioritised, the committee made a qualitative judgement regarding cost effectiveness by considering expected differences in resource and cost use between options, alongside clinical effectiveness evidence identified from the clinical evidence review.

Cost effectiveness criteria

NICE's Our Principles sets out the 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 resource use and more effective compared with all the other relevant alternative strategies)

- the intervention cost less than £20,000 per QALY gained compared with the next best strategy
- the intervention provided important benefits at an acceptable additional cost when compared with the next best strategy.

The committee's considerations of cost effectiveness are discussed explicitly under the heading 'Consideration of economic benefits and harms' in the relevant evidence reviews.

Developing recommendations

Guideline recommendations

Recommendations were drafted on the basis of the committee's interpretation of the available evidence, taking account of the balance of benefits, harms and costs between different courses of action. When effectiveness, qualitative and economic evidence was of poor quality, conflicting or absent, the committee drafted recommendations based on their expert opinion. The considerations for making consensus-based recommendations include the balance between potential benefits and harms, the economic costs or implications compared with the economic benefits, current practices, recommendations made in other relevant guidelines, person's preferences and equality issues.

The main considerations specific to each recommendation are outlined under the heading 'The committee's discussion of the evidence' within each evidence review.

For further details refer to Developing NICE guidelines: the manual.

Research recommendations

When areas were identified for which evidence was lacking, the committee considered making recommendations for future research. For further details refer to Developing NICE guidelines: the manual and NICE's Research recommendations process and methods guide.

Validation process

This guideline was subject to a 6-week public consultation and feedback process. All comments received from registered stakeholders were responded to in writing and posted on the NICE website at publication. For further details refer to Developing NICE guidelines: the manual.

Updating the guideline

Following publication, NICE will undertake a surveillance review to determine whether the evidence base has progressed sufficiently to consider altering the guideline recommendations and warrant an update. For further details refer to Developing NICE guidelines: the manual.

Funding

The NGA was commissioned by NICE to develop this guideline.

References

Bradburn 2007

Bradburn, M. J., Deeks, J. J., Berlin, J. A., & Localio, A. R. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Statistics in Medicine, 26, 53–77, 2007.

Dixon-Woods 2005

Dixon-Woods M, Agarwal S, Jones D et al. (2005) Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research & Policy. 10(1), 45–53

Hayden 2013

Hayden JA, van der Windt DA, Cartwright JL et al (2013). Assessing Bias in Studies of Prognostic Factors. Annals of Internal Medicine. 158, 280–286

Higgins 2022

Higgins JPT, Thomas J, Chandler J, et al. (editors; updated 2022). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 Cochrane. Available from www.training.cochrane.org/handbook (accessed 4 January 2023)

Lewin 2018

Lewin S, Booth A, Glenton C, Munthe-Kaas H et al. (2018) Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implementation Science. 13 (Supplement 1), 2

McGowan 2016

McGowan J, Sampson M, Salzwedel DM et al. (2016) PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. Journal of Clinical Epidemiology. 75, 40–6

NICE 2018

National Institute for Health and Care Excellence (NICE) (2014) NICE Policy on conflicts of interest (updated 2022). Available from https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/declaration-of-interests-policy.pdf (accessed 7 February 2023)

Santesso 2016

Santesso N, Carrasco-Labra A, Langendam M et al. (2016) Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. Journal of Clinical Epidemiology. 74, 28-39

Wolff 2019

Wolff RF, Moons KGM, Whiting PD et al. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine. 170, 51-60