National Institute for Health and Care Excellence

Guideline version (Consultation)

Kidney cancer: diagnosis and management

NICE guideline: methods

NICE guideline <number>

Methods

September 2025

Draft for Consultation

Evidence reviews were developed by NICE Professional team

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN:

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence commissioned the Guideline Development Team to develop a new guideline on kidney cancer.

The remit for this new development is to provide NICE guidance on the diagnosis and management of kidney cancer.

To see “What this guideline covers” and “What this guideline does not cover” please see the guideline scope Kidney cancer: diagnosis and management.

Methods

This guideline was developed using the methods described in the 2024 NICE guidelines manual.

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

Developing the review questions and outcomes

The 14 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline development team and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- Population, Intervention, Comparator and Outcome [and Study type] (PICO[S]) for reviews of interventions
- Population, index test(s), comparator (reference standard), target condition and outcome for reviews of diagnostic accuracy
- Population, predictive models, outcomes to be predicted for predictive model accuracy
- Sample, Phenomenon of Interest, Design, Evaluation, Research (SPIDER) for reviews of qualitative evidence

Full literature searches, critical appraisals and evidence reviews were completed for all review questions except for the review questions on pharmacological interventions which were part of the incorporation of technology appraisals in the guideline.

Reviewing research evidence

Review protocols

Review protocols were developed with the guideline committee to outline the inclusion and exclusion criteria used to select studies for each evidence review. No review protocols were registered on PROSPERO.

Searching for evidence

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

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example studies identified by committee members) were uploaded into EPPI reviewer software (version 5) and de-duplicated. Titles and abstracts were assessed for possible inclusion using the criteria specified in the review protocol. At least 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

Priority screening was not used for any reviews in this guideline and therefore the full database was screened for each review.

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. There were no instances that it was thought necessary to contact study investigators for missing data.

Excluding non-randomised studies which did not adjust for the full list of pre-specified confounders (see protocols) was considered. However, it was anticipated that the majority of studies would not adjust for confounders and so studies were included regardless of adjustments. When assessing studies for risk of bias using ROBINS-I, due to the risk of residual confounding, studies were assessed as at moderate risk of bias due to confounding (domain 1) if all pre-specified confounders were adjusted for. Studies were assessed as at high risk of bias for domain 1 if only some or no confounders were adjusted for.

Incorporating published evidence syntheses

If published evidence syntheses were identified sufficiently early in the review process (for example, from the surveillance review or early in the database search), they were considered for use as the primary source of data, rather than extracting information from primary studies. Syntheses considered for inclusion in this way were quality assessed to assess their suitability using the appropriate checklist, as outlined in

Table 1. Note that this quality assessment was solely used to assess the quality of the synthesis in order to decide whether it could be used as a source of data, as outlined in Table 2, not the quality of evidence contained within it, which was assessed in the usual way as outlined in the section on 'Appraising the quality of evidence'.

Table 1: Checklists for published evidence syntheses

Type of synthesis	Checklist for quality appraisal
Systematic review of quantitative evidence	ROBIS
Network meta-analysis	Modified version of the PRISMA NMA tool (see appendix K of 'Developing NICE guidelines, the manual')
Qualitative evidence synthesis	ENTREQ reporting standard for published evidence synthesis (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-181) is the generic reporting standard for QES, however specific reporting standards exist for meta-ethnography (eMERGe [https://emergeproject.org/]) and for realist synthesis (RAMESES II [https://www.ramesesproject.org/]). If these reporting standards are not appropriate to the QES then an adapted PRISMA framework is used (see Flemming K, Booth A, Hannes K, Cargo M, Noyes J. Cochrane Qualitative and Implementation Methods Group guidance series-paper 6: reporting guidelines for qualitative, implementation, and process evaluation evidence syntheses. Journal of Clinical Epidemiology 2018; 97: 79-85).

Type of synthesis	Checklist for quality appraisal
Individual patient data meta-analysis	Checklist based on Tierney, Jayne F., et al. "Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use." PLoS Med 12.7 (2015): e1001855.

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

- Low risk of bias – It is unlikely that additional relevant and important data would be identified from primary studies compared to that reported in the review, and unlikely that any relevant and important studies have been missed by the review.
- Moderate risk of bias – It is possible that additional relevant and important data would be identified from primary studies compared to that reported in the review, but unlikely that any relevant and important studies have been missed by the review.
- High risk of bias – It is possible that relevant and important studies have been missed by the review.

Each published evidence synthesis was also classified into one of three groups for its applicability as a source of data, based on how closely the review matches the specified review protocol in the guideline. Studies were rated as follows:

- Fully applicable – The identified review fully covers the review protocol in the guideline.
- Partially applicable – The identified review fully covers a discrete subsection of the review protocol in the guideline (for example, some of the factors in the protocol only).
- Not applicable – The identified review, despite including studies relevant to the review question, does not fully cover any discrete subsection of the review protocol in the guideline.

The way that a published evidence synthesis was used in the evidence review depended on its quality and applicability, as defined in Table 2. When published evidence syntheses were used as a source of primary data, data from these evidence syntheses were quality assessed and presented in GRADE/CERQual tables in the same way as if data had been extracted from primary studies. In questions where data was extracted from both systematic reviews and primary studies, these were checked to ensure none of the data had been double counted through this process.

Table 2: Criteria for using published evidence syntheses as a source of data

Risk of bias	Applicability	Use of published evidence synthesis
Low	Fully applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search or data analysis. Searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted.

Risk of bias	Applicability	Use of published evidence synthesis
Low	Partially applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search and data analysis for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted. For other sections not covered by the evidence synthesis, searches were undertaken as normal.
Moderate	Fully applicable	Moderate: Details of included studies were used instead of undertaking a new literature search. Full-text papers of included studies were still retrieved for the purposes of data analysis. Searches were only done to cover the period of time since the search date of the review.
Moderate	Partially applicable	Details of included studies were used instead of undertaking a new literature search for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. For other sections not covered by the evidence synthesis, searches were undertaken as normal.

Methods of combining evidence

Data synthesis for intervention studies

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

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane RevMan Web, with the exception of incidence rate ratio analyses which were carried out in R version 4.1.0. using the package ‘metafor’. A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event, and a pooled incidence rate ratio was calculated for dichotomous outcomes reporting total numbers of events. Both relative and absolute risks were presented where sufficient information was reported in the included papers. Absolute risks were calculated by applying the relative risk to the risk in the comparator arm of the meta-analysis (calculated as the total number events in the comparator arms of studies in the meta-analysis divided by the total number of participants in the comparator arms of studies in the meta-analysis).

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

Where outcomes measured the same underlying construct but used different instruments/metrics, data were analysed using standardised mean differences (SMDs, Hedges' g), as implemented in RevMan Web Alternative approaches to standardisation described in the NICE technical support unit guideline methodology document on meta-analysis of continuous outcomes were used when this was needed for consistency with a network meta-analysis.

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. For continuous outcomes analysed as standardised mean differences this was not possible. In this case, if all studies reported final timepoint data, this was used in the analysis. If some studies only reported data as a change from baseline, analysis was done on these data, and for studies where only baseline and final time point values were available, change from baseline standard deviations were estimated, assuming a correlation coefficient derived from studies reporting both baseline and endpoint data, or if no such studies were available, assuming a correlation of 0.5 as a conservative estimate (Follman et al., 1992; Fu et al., 2013).. In cases where SMDs were used they were back converted to a single scale to aid interpretation by the committee where possible, and when it was considered useful for decision making.

Random effects models were fitted when significant between-study heterogeneity in methodology, population, intervention or comparator was identified by the reviewer in advance of data analysis. This decision was made and recorded before any data analysis was undertaken. For all other syntheses, fixed- and random-effects models were fitted, with the presented analysis dependent on the degree of heterogeneity in the assembled evidence. Fixed-effects models were the preferred choice to report, but in situations where the assumption of a shared mean for fixed-effects model was clearly not met, 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 $I^2 \geq 50\%$.

However, in cases where the results from individual pre-specified subgroup analyses were less heterogeneous (with $I^2 < 50\%$) the results from these subgroups were reported using fixed effects models. This may have led to situations where pooled results were reported from random-effects models and subgroup results were reported from fixed-effects models.

Where sufficient studies were available, meta-regression was considered to explore the effect of study level covariates.

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).

Where multiple observer interpretations of the same images were reported in a study, all interpretations were included.

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)

Likelihood ratios provide information on how much more likely a given test result (positive or negative) is in someone with the condition (in this case, renal cell carcinoma) compared to someone without it. If the positive likelihood ratio (LR+) or negative likelihood ratio (LR-) is 1, then the probability of the disease being present after the test (post-test probability) is the same as the probability of the disease being present before the test (pre-test probability). A test with a high LR+ or low LR- is more informative. A high LR+ means that where there is a positive test result, post-test probability of disease increases. A low LR- means that where there is a negative test result, the post-test probability of disease decreases.

Meta-analysis of diagnostic accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.0 (Deeks et al. 2022).

MetaDTA was used to synthesise diagnostic test accuracy data and obtain sensitivity and specificity outputs. MetaDTA is based on the glmer function in the lme4 package in R v 3.4.0 and runs a bivariate meta-analysis using a generalised linear mixed model (GLMM). This accounts for the correlations between positive and negative likelihood ratios, and between sensitivity and specificity. The mada package in R was used to produce plots of positive and negative likelihood ratios by inputting the output data from MetaDTA.

MetaDTA requires sufficient data in order to accurately estimate the parameters required. Insufficient data results in the model failing to converge. When data would not converge in MetaDTA, meta-analysis was not undertaken. Sensitivity and specificity were transformed to the logit scale and the correlation between the two were exampled. If correlation is considered negligible (less than 0.3), univariate

analysis for sensitivity and specificity would be performed. If there is correlation then meta-analysis would not be performed and results would be analysed individually.

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

Methods for combining c-statistics for prognostic models

Meta-analyses were carried out using the metamisc package in R v4.5.0, which confines the analysis results to between 0 and 1 matching the limited range of values that c-statistics can take. Random effects meta-analysis was used when the I^2 was 50% or greater.

Appraising the certainty of evidence

Intervention studies (relative effect estimates)

RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias Tool. Non-randomised controlled trials and cohort studies were quality assessed using the ROBINS-I tool. Other study types (for example controlled before and after studies) were assessed using the preferred option specified in the NICE guidelines manual 2018 (appendix H). 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.
- Some concerns – There is a possibility the true effect size for the study is substantially different to the estimated effect size.
- High risk of bias – It is likely the true effect size for the study is substantially different to the estimated effect size.
- Critical risk of bias (ROBINS-I only) - It is very likely the true effect size for the study is substantially different to the estimated effect size.

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

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

Minimally important differences (MIDs) and clinical decision thresholds

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

had been developed and validated in a methodologically rigorous way, and were applicable to the populations, interventions and outcomes specified in this guideline. In addition, the Guideline Committee were asked to prospectively specify any outcomes where they felt a consensus clinical decision threshold could be defined from their experience.

Clinical decision thresholds were used to assess imprecision using GRADE and aid interpretation of the size of effects for different outcomes. Clinical decision threshold that were used in the guideline are given in Table 3 and also reported in the relevant evidence reviews.

Table 3: Identified Clinical decision thresholds

Outcome	Clinical decision threshold	Source
EQ-5D	0.08 for UK-based scores and 0.07 for VAS scores	Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes. 2007 Dec 21;5:70.

For continuous outcomes expressed as a mean difference where no other clinical decision threshold was available, a clinical decision threshold of 0.5 of the median standard deviations of the comparison group arms was used (Norman et al. 2003). For continuous outcomes expressed as a standardised mean difference where no other clinical decision threshold was available, a clinical decision threshold of 0.5 standard deviations was used. For SMDs that were back converted to one of the original scales to aid interpretation, rating of imprecision was carried out before back calculation. 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. Where possible, odds ratios were converted to risk ratios before presentation to the committee to aid interpretation.

GRADE for intervention studies analysed using pairwise analysis

GRADE was used to assess the certainty of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials, non-randomised controlled trials and cohort studies (which were quality assessed using the Cochrane risk of bias tool or ROBINS-I) were initially rated as high certainty while data from other study types were initially rated as low certainty. The certainty of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 4. 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 4: Rationale for downgrading certainty of evidence for intervention studies

GRADE criteria	Reasons for downgrading certainty
Risk of bias	Not serious: If less than 50% of the weight in a meta-analysis came from studies with some concerns or high risk of bias, the overall outcome was not downgraded.

GRADE criteria	Reasons for downgrading certainty
	Serious: If greater than 50% of the weight in a meta-analysis came from studies with some concerns or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 50% of the weight in a meta-analysis came from studies at high risk of bias (combine serious and critical as high risk for ROBINS-I), the outcome was downgraded two levels.
Indirectness	Not serious: If less than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 50% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. Not serious: If the I^2 was less than 40%, the outcome was not downgraded. Serious: If the I^2 was between 40% and 60%, the outcome was downgraded one level or if data on the outcome was only available from one study. Very serious: If the I^2 was greater than 60%, the outcome was downgraded two levels. Where I^2 is 80% or above, data may be too heterogeneous to meaningfully pool. This will be considered on a case by case basis.
Imprecision	If an MID other than the line of no effect was defined for the outcome, the outcome was downgraded once if the 95% confidence interval for the effect size crossed one line of the MID, and twice if it crosses both lines of the MID. If the line of no effect was defined as an MID for the outcome, it was downgraded once if the 95% confidence interval for the effect size crossed the line of no effect (i.e. the outcome was not statistically significant), and twice if the sample size of the study was sufficiently small that it is not plausible any realistic effect size could have been detected. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to clinically equivalent scenarios.
Publication bias	Where 10 or more studies were included as part of a single meta-analysis, a funnel plot was produced to graphically assess the potential for publication bias. When a funnel plot showed convincing evidence of publication bias, or the review team became aware of other evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

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

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

Diagnostic accuracy studies

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

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

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

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

GRADE for diagnostic accuracy evidence

Evidence from diagnostic accuracy studies was initially rated as high certainty, and then downgraded according to the standard GRADE criteria (risk of bias, inconsistency, imprecision and indirectness) as detailed in Table 6 below.

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

In 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. In reviews where a decision model is being carried (for example, as part of an economic analysis), these consequences were incorporated here in addition.

Using likelihood ratios as the primary outcomes

The following schema (Table 5), adapted from the suggestions of Jaeschke et al. (1994), was used to interpret the likelihood ratio findings from diagnostic test accuracy reviews.

Table 5: Interpretation of likelihood ratios

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

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

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

The committee were consulted to set 2 clinical decision thresholds for each measure: the likelihood ratio above (or below for negative likelihood ratios) which a test would be recommended, and a second below (or above for negative likelihood ratios) which a test would be considered of no clinical use. These were used to judge imprecision (see below). If the committee were unsure which values to pick, then the default values of 2 for LR+ and 0.5 for LR- were used based on Table 5, with the line of no effect (being 1.0) as the second clinical decision line in both cases.

If studies could not be pooled in a meta-analysis, GRADE assessments were undertaken on the body of evidence that was considered for meta-analysis, rather than on individual estimates or median estimates.

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 6: Rationale for downgrading certainty of evidence for diagnostic accuracy data

GRADE criteria	Reasons for downgrading certainty
Risk of bias	Not serious: If less than 50% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 50% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level.

GRADE criteria	Reasons for downgrading certainty
	Very serious: If greater than 50% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 50% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
Inconsistency	Inconsistency was assessed by visual inspection of the point estimates and confidence intervals of the included studies. The evidence was downgraded if these varied widely between studies, for example, point estimates for some studies lying outside the CIs of other studies. Weighted subjective judgement was used to downgrade as follows: No serious: No studies were considered to be meaningfully inconsistent with point estimates outside the CIs of other studies Serious: <50% of studies were inconsistent Very serious: ≥50% of studies were inconsistent.
Imprecision	If the 95% confidence interval for the outcome crossed one of the clinical decision thresholds, the outcome was downgraded one level. If the 95% confidence interval spanned both thresholds, the outcome was downgraded twice. See the section on 'Using likelihood ratios as the primary outcome' for a description of how clinical decision thresholds were agreed.
Publication bias	If the review team became aware of evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

Studies developing or evaluating prediction models

Individual studies developing or validating prediction models were assessed using the PROBAST checklist. Each individual study was classified into one of the following three groups:

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

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, index features and/or reference standard/outcome to be predicted 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/outcome to be predicted.
- Partially indirect – Important deviations from the protocol in one of the population, index feature and/or reference standard/outcome to be predicted.
- Indirect – Important deviations from the protocol in at least two of the population, index feature and/or reference standard/outcome to be predicted.

Using c-statistics for measuring discriminative ability of prediction models

C-statistics were assessed using the categories in Table 7 below

Table 7: Interpretation of c-statistics

C-statistic	Interpretation of discriminative ability
c-statistic <0.6	Very poor classification of accuracy
0.6 ≤ c-statistic <0.7	Poor classification of accuracy
0.7 ≤ c-statistic <0.8	Fair classification of accuracy
0.8 ≤ c-statistic <0.9	Good classification of accuracy
0.9 ≤ c-statistic < 1.0	Excellent classification of accuracy

Modified GRADE for prediction models

GRADE has not been developed for use with data from prediction models, therefore a modified approach was applied using the GRADE framework. The approach taken depended on the outcome data produced by the decision model. C-statistic data was assessed as described in Table 8.

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 8: Rationale for downgrading certainty of evidence for association studies

GRADE criteria	Reasons for downgrading certainty
Risk of bias	Not serious: If less than 50% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 50% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 50% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 50% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 50% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.

GRADE criteria	Reasons for downgrading certainty
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic in combination with visual assessment. Not serious: If the I^2 was less than 40%, the outcome was not downgraded. Serious: If the I^2 was between 40% and 60%, the outcome was downgraded one level. Also if data on the outcome was only available from one study. Very serious: If the I^2 was greater than 60%, the outcome was downgraded two levels. Where I^2 is 80% or above, data may be too heterogeneous to meaningfully pool. This will be considered on a case by case basis.
Imprecision	Not serious: If the 95% confidence interval for the outcome did not cross a decision threshold. Serious: If the 95% confidence interval for the outcome crossed one of the clinical decision thresholds, the outcome was downgraded one level. Very Serious: If the 95% confidence interval crossed more than two clinical decision thresholds, the outcome was downgraded twice. See the sections on 'Using c-statistics for measuring discriminative ability of prediction models' for a description of how clinical decision thresholds were agreed.
Publication bias	If the review team became aware of evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

Qualitative studies

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

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

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

- Highly relevant – No important deviations from the protocol in perspective, population, phenomenon of interest and/or setting.

- Relevant – Important deviations from the protocol in one of the perspective, population, phenomenon of interest and/or setting.
- Partially relevant – Important deviations from the protocol in at least two of the perspective, population, phenomenon of interest and/or setting.

CERQual was used to assess the confidence we have in each of the review findings. Evidence from all qualitative study designs (interviews, focus groups etc.) was initially rated as high confidence and the confidence in the evidence for each theme was then downgraded from this initial point as detailed in Table 9 below.

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 9: Rationale for downgrading confidence in evidence for qualitative questions

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

Reviewing economic evidence

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 de novo cost effectiveness analysis in priority areas.

Inclusion and exclusion of economic studies

Systematic reviews of economic literature were conducted in all areas 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 protocol inclusion and exclusion criteria summarised above.

Details of economic evidence study selection and lists of excluded studies after full-text assessment (together with reasons for exclusion) are presented in respective appendices of the evidence reviews.

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 economic modelling 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 10.

Table 10: 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.

All studies that fully or partially met the applicability and quality criteria described in the methodology checklist were considered by the committee during the guideline development process. However, high-quality studies in line with the NICE reference case (such as recent UK NHS/PSS cost-utility analyses using the QALY as the measure of outcome) were prioritised for use in decision making as these are the most relevant to the NICE guideline development context. Therefore, not all economic studies meeting the inclusion criteria were necessarily used in decision-making. Further criteria for prioritising economic studies for use in decision making are provided in the economic review protocols (under 'Review strategy'), within the respective appendices of the evidence reviews.

Details on methods and results of economic studies that met inclusion criteria and were subsequently used in decision making are shown in economic evidence study extraction tables, provided in respective appendices of the economic reviews. Economic evidence study extraction tables are followed by lists of economic studies that met inclusion criteria but were not used in decision making (with reasons for non-use).

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

Health economic modelling

Cost-effectiveness analysis, based on decision-analytic modelling, 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. Economic modelling was undertaken 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 following economic questions were selected as key issues to be addressed by economic modelling in the guideline:

- Review F: Clinically and cost-effective risk-stratified follow-up strategies (based on method, duration, and frequency), for adults who have had treatment for localised or locally advanced RCC.
- Review J: Diagnostic accuracy and cost effectiveness of core biopsy for diagnosing renal masses. Although this question was originally prioritised for

economic modelling, a new economic study that was relevant to this question was identified during guideline development. The study was directly applicable and of high quality. Therefore, there was no need for de novo economic modelling to be conducted.

- Review A, Review B and Review C: Cost-effectiveness of non-surgical interventions or active surveillance in adults with localised or locally advanced RCC. This was prioritised as a costing analysis only, not a cost-utility analysis.

The following general principles were adhered to in developing the guideline cost-effectiveness analyses:

- Methods were consistent with the NICE reference case for interventions with health outcomes in NHS settings.
- The committee was involved in the design of the model and related assumptions, selection of inputs and interpretation of the results.
- Model inputs were based on the systematic review of the effectiveness literature supplemented with other published data sources where needed.
- When published data were not available for a model input, the committee's expert opinion was used to inform it.
- Model inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.
- The model was peer-reviewed by another health economist who was independent of the guideline development process.
- Full methods and results of the cost-effectiveness analysis for Review F and the costing analysis for Reviews A, B and C are described in their respective economic supplements.

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 'The committee's discussion and interpretation of the evidence' under the subheading 'Cost-effectiveness and resource use', in each evidence review.

References

Follmann D, Elliott P, Suh I, Cutler J (1992) Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology 45:769–73

Fu R, Vandermeer BW, Shamliyan TA, et al. (2013) Handling Continuous Outcomes in Quantitative Synthesis In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK154408/

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA, 271(9), 703–707.

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.