National Institute for Health and Care Excellence

Final

Adrenal insufficiency: identification and management

NICE guideline: methods

NICE guideline NG243

Methods

August 2024

Final

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN: 978-1-4731-6476-5

Contents

1 Development of the guideline

1.1 Remit

NICE received the remit for this guideline from NHS England.

The remit for this guideline is: Adrenal insufficiency: acute and long-term management.

To see what this guideline covers and what this guideline does not cover, please see the guideline scope Adrenal Insufficiency.

2 Methods

This guideline was developed using the methods described in the NICE guidelines manual[4] updated 2020.

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

Sections 2.1 to 2.3 describe the process used to identify and review evidence. Sections 2.1.1 and 2.7 describe the process used to identify and review the health economic evidence.

2.1 Developing the review questions and outcomes

The review questions developed for this guideline were based on the key areas and draft review questions identified in the guideline scope. They were drafted by the technical team, refined and validated by the committee and signed off by NICE. A total of 19 review questions were developed in this guideline and outlined in Table 2.

The review questions were based on the following frameworks:

- population, intervention, comparator, and outcome (PICO) for reviews of interventions (including test and treat)
- population, index tests, reference standard and target condition for reviews of diagnostic test accuracy
- population, exposure, and outcomes for prognostic reviews
- population, setting and context for qualitative reviews.

This use of a framework informed a more detailed protocol that guided the literature searching process, critical appraisal, and synthesis of evidence, and facilitated the development of recommendations by the guideline committee. Full literature searches, critical appraisals and evidence reviews were completed for all the specified review questions.

Table 1: Review questions

Evidence report	Type of review	Review questions	Outcomes
1.1 (A)	Qualitative	What information and support do people with suspected or diagnosed adrenal insufficiency (and their families and carers) need to routinely manage their health (including how to ensure an adequate supply of medicines, advice on what to do in certain situations such as when exercising, travelling, working non-standard hours, or taking part in religious observances such as fasting)?	Themes will be derived from the evidence identified for this review and are not pre-specified. They may include: • how to ensure an adequate supply of medicines • advice on what to do in certain situations such as: – when exercising, for example, competitive recreational and activities with risk of injury – travelling for example, across time zones, extremes of climates – working non-standard hours – taking part in religious observances such as fasting • how to prevent an adrenal crisis

Evidence report	Type of review	Review questions	Outcomes
			• emergency care of adrenal crisis
1.2 (A)	Qualitative	What information and support do people diagnosed with adrenal insufficiency need for the prevention and emergency care of an adrenal crisis?	• content in particular for transition to adult services • pregnancy • educational information for schools including medicines administration and injections. • illness and physiological stress (e.g., sick-day rules).
2.1 (B)	Diagnostic Prognostic	When should adrenal insufficiency be suspected (for example, based on risk factors or symptoms)?	**For signs and symptoms review:** • Diagnostic accuracy data – sensitivity (prioritised) – specificity If no sensitivity or specificity, LR- and LR+ if raw data unavailable and unable to calculate from 2 x 2 table. Diagnostic association of signs and symptoms with a confirmed diagnosis of adrenal insufficiency. Measured by: • Association data – Adjusted hazard ratios, odds ratios, or risk ratios. • Discrimination – For example, C statistic, area under ROC curve • Calibration – For example, calibration slope **For risk factors review:** • Diagnosis of adrenal insufficiency as defined by authors and reported as adjusted hazard ratios, odds ratios, or risk ratios. • For risk prediction tools: sensitivity, specificity and statistical measures of discrimination and calibration including Area Under the Curve (AUC) for risk tools.
2.2 (C)	Diagnostic	When should a person who is having exogenous corticosteroids withdrawn be referred for investigation and management of adrenal insufficiency related to HPA-axis suppression?	Diagnostic accuracy data • Sensitivity (prioritised) [fewer false • negatives i.e., very few people with the condition will be missed] • Specificity

Evidence report	Type of review	Review questions	Outcomes
			The GC has prioritised sensitivity and specificity as the most important outcomes for their interpretation of the evidence. The following thresholds will be used for imprecision for DTA measures and for deciding on the usefulness of the tests in detecting adrenal insufficiency: • Sensitivity – Upper 0.9 – Lower 0.6 • Specificity – Upper 0.7 – Lower 0.5 Likelihood ratios or other measures such as C statistic or area under ROC curve will only be reported if they are the only measures available, sensitivity and specificity are not reported and cannot be calculated from raw data. Should this be the case, cut-offs for summarising the performance of diagnostic tests or prediction models will be agreed with the guideline committee before the analysis of the evidence is conducted and the protocol will be updated accordingly.
2.3 (D)	Diagnostic	What initial investigations should be done by the non-specialist for people with suspected adrenal insufficiency?	Diagnostic accuracy data • Sensitivity (prioritised) [fewer false negatives i.e., very few people with the condition will be missed] • Specificity The GC has prioritised sensitivity and specificity as the most important outcomes for their interpretation of the evidence. The following thresholds will be used for imprecision for DTA measures and for deciding on the usefulness of the tests in detecting adrenal insufficiency: Sensitivity – Upper 0.9 – Lower 0.6 • Specificity – Upper 0.7 – Lower 0.5

Evidence report	Type of review	Review questions	Outcomes
			Likelihood ratios or other measures such as C statistic or area under ROC curve will only be reported if they are the only measures available, sensitivity and specificity are not reported and cannot be calculated from raw data. Should this be the case, cut-offs for summarising the performance of diagnostic tests or prediction models will be agreed with the guideline committee before the analysis of the evidence is conducted and the protocol will be updated accordingly.
2.4 (D)	Intervention	When should people with suspected adrenal insufficiency be referred to specialists for further investigation?	Diagnostic accuracy data • Sensitivity (prioritised) [fewer false • negatives i.e., very few people with • the condition will be missed] • Specificity The GC has prioritised sensitivity and specificity as the most important outcomes for their interpretation of the evidence. The following thresholds will be used for imprecision for DTA measures and for deciding on the usefulness of the tests in detecting adrenal insufficiency: • Sensitivity – Upper 0.9 – Lower 0.6 • Specificity – Upper 0.7 – Lower 0.5 Likelihood ratios or other measures such as C statistic or area under ROC curve will only be reported if they are the only measures available, sensitivity and specificity are not reported and cannot be calculated from raw data. Should this be the case, cut-offs for summarising the performance of diagnostic tests or prediction models will be agreed with the guideline committee before the analysis of the evidence is conducted and the protocol will be updated accordingly.
3.1 (E)	Intervention	In people at risk of adrenal insufficiency because of prolonged	All outcomes are considered equally important for decision making and therefore have all been rated as critical:

Evidence report	Type of review	Review questions	Outcomes
		corticosteroid use, what is the best way to manage. corticosteroid withdrawal when corticosteroids are no longer needed to control disease activity?	• Health related quality of life for example EQ-5D, SF-36 • Incidence of adrenal insufficiency • Incidence adrenal crisis • Hospital admission • Successful cessation of steroids as indicated by, for example, rate of relapse. • Adverse events – as reported. Due to the sparsity of the evidence, we will include any follow up time.
4.1 (F)	Intervention	What is the clinical and cost effectiveness of pharmacological treatments for the routine management of primary adrenal insufficiency?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36 • Complications of adrenal insufficiency – For all causes of PAI – growth related issues in children – Low blood sugar/ hypoglycaemia − Early satiety – Complications specifically related to mineralocorticoid deficiencies: o Salt wasting / hyponatraemia o Salt cravings o Dizziness o Muscle cramps o Low blood pressure o Muscle weakness o Nocturia Additional Complications of PAI for patients with CAH – delayed/ precocious puberty in children. – lack of periods /virilisation / fertility issues – Hirsutism. • **Fatigue** as measured using specific fatigue scales such as National Fatigue Index (NFI), fatigue Severity Scale (FSS) • **Incidence of adrenal crisis** (as defined by authors) • **Complications of adrenal crisis** - for example neurological complications, psychological, hypoglycaemia, shock,

Evidence report	Type of review	Review questions	Outcomes
			acute kidney injury may be as part of shock and related to hypovolaemia. • **Androgen normalisation** (specific to CAH) determined by biochemical parameters such as 17 OHP, androstenedione, testosterone and DHEAS) • **Admission to hospital and/or ITU** • **Readmission to hospital** • **Length of stay at hospital or ITU.** • **Treatment-related adverse events**: – Hypertension – Obesity/weight gain – Osteoporosis – Fracture – Heart disease/CVS – Cushingoid features: for example, stretch marks. – Diabetes – Impact on sleep- poor sleep due to overnight high cortisol levels – stunted growth in children – Hb1ac – Psychological effects (depression, anxiety) – Fluid retention – Increased risk of glaucoma/high pressure in the eyes – Effects on concentration – Specific to subcutaneous routes: sites reactions, infections, pumps breaking. – Stomach ulcers • Activities of daily living – Social participation – Participation in education (School/university) Participation in physical activity (measured by any validated scale such as Barthel Index, the Katz Index, or the Functional Independence Measure). Note: there is some overlap between outcomes. For example, hypoglycaemia may be due to either complications of AI or be a complication of adrenal crisis. We will note which outcome these relate to. **Follow up:** Any time point as this will be different for different variables. Most will be short term

Evidence report	Type of review	Review questions	Outcomes
			(within 30 days) except for weight or growth-related outcomes, QoL and activities of daily living. We will prioritise data from similar timepoints in order to increase the possibility of conducting a meta-analysis (if appropriate). For QoL and activities of daily living we will also include longer term data where available.
4.2 (G)	Intervention	What is the clinical and cost effectiveness of pharmacological treatments for the routine management of secondary and tertiary adrenal insufficiency?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36 • Complications of adrenal insufficiency – growth related issues in children – Low blood sugar/ hypoglycaemia – Early satiety • Fatigue as measured using specific fatigue scales such as National Fatigue Index (NFI), fatigue Severity Scale (FSS) • Incidence of adrenal crisis (as defined by authors) • Complications of adrenal crisis- for example neurological complications, psychological, hypoglycaemia, shock, acute kidney injury may be as part of shock and related to hypovolaemia Error! No text of specified style in document. [Guideline short title]: [Review title] protocol © National Institute for Health and Care Excellence, 2018 11 • Admission to hospital and/or ITU • Readmission to hospital • Length of hospital stay. • Treatment-related adverse events: – Hypertension – Obesity/weight gain – Osteoporosis – Fracture – Heart disease/CVS – Cushingoid features: for example, stretch marks. – Diabetes

Evidence report	Type of review	Review questions	Outcomes
			– Impact on sleep- poor sleep due to overnight high cortisol levels – stunted growth in children – Hb1ac – Psychological effects (depression, anxiety) – Fluid retention – Increased risk of glaucoma/high pressure in the eyes – Effects on concentration – Specific to subcutaneous routes: sites reactions, infections, pumps breaking. – Stomach ulcers • Activities of daily living – Social participation – Participation in education (School /university) – Participation in physical activity (measured by any validated scale such as Barthel Index, the Katz Index, or the Functional Independence Measure). Note: there is some overlap between outcomes. For example, hypoglycaemia may be due to either complications of AI or be a complication of adrenal crisis. We will note which outcome these relate to. **Follow up:** Any time point as this will be different for different variables. Most will be short term (within 30 days) except for weight or growth-related outcomes, QoL and activities of daily living. We will prioritise data from similar timepoints in order to increase the possibility of conducting a meta-analysis (if appropriate).
4.3 (H)	Intervention	When should adrenal crisis be suspected?	Association data Adjusted hazard ratios, odds ratios, or risk ratios <u>Discrimination data</u> • for example, C statistic, area under ROC curve <u>Calibration data</u> • for example, calibration slope

Evidence report	Type of review	Review questions	Outcomes
			Diagnostic accuracy data • Sensitivity (prioritised) • specificity If no sensitivity or specificity, we will report LR- and LR+ if raw data unavailable and unable to calculate from 2 x 2 table.
4.4 (I)	Intervention	What is the clinical and cost effectiveness of pharmacological treatments for the emergency management of adrenal crisis?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36 • Complications of adrenal crisis for example neurological complications, psychological, hypotension, hypoglycaemia, shock, acute kidney injury may be as part of shock and related to hypovolaemia. • Acute adverse events of drugs: (up to 2 weeks- if none at this FU include shortest FU time reported in paper) – Mania – mood disturbance – blood glucose disturbance – sleep disruption/ insomnia • Admission to hospital and/or ITU • Hospital readmission • Length of hospital stay. • Electrolyte abnormalities such as incidence of hyponatraemia • Adverse effects of hyponatraemia **Follow up:** Up to 4 weeks (all short-term outcomes within hours or days that should all be captured by 4 weeks) If studies report several timepoints – shorter ones are preferable.
4.5 (J)	Intervention	What is the clinical and cost effectiveness of pharmacological treatments for managing periods of physiological stress in people with	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36

Evidence report	Type of review	Review questions	Outcomes
		adrenal insufficiency including: a) planned and emergency invasive procedures b) pregnancy and intrapartum care c) intercurrent illness and periods of physiological stress including minor (for example, colds) and major illnesses (for example, severe infection, cardiac events)?	• Incidence of adrenal crisis • Acute adverse events of drugs: (up to 2 weeks- if none at this FU include shortest FU time reported in paper) − Mania − mood disturbance − blood glucose disturbance − sleep disruption/ insomnia • Long term cumulative adverse effects: − impact on weight − impact on growth − Hypertension − Obesity/weight gain − Osteoporosis − Fracture − Heart disease/CVS − Cushingoid features: for example, stretch marks. − Diabetes (newly diagnosed or exacerbated) − Impact on sleep- poor sleep due to overnight high cortisol levels − stunted growth in children − Hb1ac − Psychological effects (depression, anxiety) − Fluid retention − Increased risk of glaucoma/high pressure in the eyes − Effects on concentration − Stomach ulcers • Admission to hospital • Admission to ITU • Length of hospital stay. • Readmission to hospital • Psychological morbidities e.g., Incidence of stress or PTSD • Adverse effects of hypoglycaemia for example, neurological damage, seizures, • Adverse effects of hyponatraemia for example, neurological damage, seizures, **Follow up:** >12 months but will report other time points if 12 months not available.
4.6 (K)	Intervention	What is the clinical and cost effectiveness of pharmacological	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • **Mortality**

Evidence report	Type of review	Review questions	Outcomes
		treatments for managing periods of psychological stress in people with adrenal insufficiency?	• **Health-related quality of life, for example EQ-5D, SF-36** • **Incidence of adrenal crisis** • **Acute adverse events of drugs:** (up to 2 weeks - if none at this FU include shortest FU time reported in paper) – Mania – mood disturbance – blood glucose disturbance – sleep disruption/ insomnia • **Long term cumulative adverse effects**: – impact on weight – impact on growth – Hypertension – Obesity/weight gain – Osteoporosis – Fracture – Heart disease/CVS – Cushingoid features: for example, stretch marks. – Diabetes (newly diagnosed or exacerbated) – Impact on sleep (may be poor sleep due to overnight high cortisol levels) – stunted growth in children – Hb1ac – Psychological effects (depression, anxiety) – Fluid retention – Increased risk of glaucoma/high pressure in the eye – Effects on concentration – Stomach ulcers • Admission to hospital • Admission to ITU • Length of hospital stay. • Readmission to hospital • Psychological morbidities e.g., Incidence of stress or PTSD • Adverse effects of hypoglycaemia for example, neurological damage, seizures • Adverse effects of hyponatraemia for example, neurological damage, seizures, Follow up >12 months but will report other time points if 12 months not available.

Evidence report	Type of review	Review questions	Outcomes
4.7 (L)	Intervention	What is the clinical and cost effectiveness of non-pharmacological strategies to prevent adrenal crisis during periods of intercurrent illness and periods of physiological stress?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36 • Incidence of adrenal crisis • Admission to hospital • Admission to ITU • Length of hospital stay. • Readmission to hospital • Psychological morbidities for example, Incidence of stress or PTSD **Follow up:** Medium 6 months to a year. If evidence only available for less than 6 months this will be included and downgraded for indirectness.
4.8 (M)	Intervention	What is the clinical and cost effectiveness of non-pharmacological strategies to prevent adrenal crisis during periods of psychological stress?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality • Health-related quality of life, for example EQ-5D, SF-36 • Incidence of adrenal crisis • Admission to hospital • Admission to ITU • Length of hospital stay. • Readmission to hospital • Psychological morbidities e.g., Incidence of stress or PTSD • Mental health admission **Follow up:** Medium 6 months to a year. If evidence only available for less than 6 months this will be included and downgraded for indirectness.
5.1 (N)	Intervention	What ongoing care and monitoring should be offered to people with adrenal insufficiency?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Mortality
5.2 (N)	Intervention	What ongoing care and monitoring should be	

Evidence report	Type of review	Review questions	Outcomes
		offered to people with adrenal insufficiency who are receiving end of life care?	• Health-related quality of life, for example EQ-5D, SF-36 • Complications of adrenal insufficiency For example, in primary AI: – Growth related issues in children – Low blood sugar/ hypoglycaemia – Early satiety – Complications specifically related to mineralocorticoid deficiencies: o Salt wasting / hyponatraemia o Salt cravings o Dizziness o Muscle cramps o Low blood pressure o Muscle weakness o Nocturia • Incidence of adrenal crisis (as defined by authors) • Incidence Vascular events • Incidence of fractures • Incidence of diabetes • Activities of daily living – Social participation – Participation in education (School/ University) Participation in physical activity (measured by any validated scale such as Barthel Index, the Katz Index, or the Functional Independence Measure). **Follow up:** Longest follow up reported. Where different follow up periods are reported in an individual study, we will choose the one most appropriate or most commonly reported to be able to conduct a meta-analysis.

2.1.1 Stratification

The following stratifications were considered in the analyses:

In review 3.1 (E):

- Adults (aged ≥16 years).
- Children aged > 5 to 16 years.
  Infants aged 1-5 years.

- Infants aged <1 year including neonates.

In review 4.1 (F):

- Adults (aged ≥16 years) – All adults with primary adrenal insufficiency including Addison's disease.
- Children aged ≥1 up to 16 years with CAH.
- Children aged ≥1 to <16 years with no CAH.
- Infants aged <1 year with CAH (including neonates up to 28 days)
- Infants aged <1 year with no CAH (including neonates up to 28 days)

In reviews 4.2 (G), 4.4 (I), 4.5 (J), 4.6 (K):

- Adults (aged ≥16 years).
- Children aged ≥5 up to 16 years.
- Infants aged 1-5 years.
- Infants aged <1 year including neonates.

In reviews 4.7 (L) and 4.8 (M):

- Adults (aged ≥16 years)
- Children aged ≥ 5 up to 16 years.
- Children aged < 5.

In reviews 5.1 (N) and 5.2 (N):

- Adults (aged ≥16 years).
- Children aged ≥5 up to 16 years.
- Infants aged 1-5 years (because of more frequent dosing)
- Infants aged <1 year including neonates.
- Adults or children receiving end of life care.

Where studies reported a mix of populations across strata, a threshold of 80% was agreed with the committee as a cut off for what would be acceptable to constitute a predominant group.

2.2 Searching for evidence

2.2.1 Clinical and health economics literature searches

The full strategy including population terms, intervention terms, study types applied, the databases searched, and the years covered can be found in Appendix B of the evidence review.

Systematic literature searches were undertaken to identify published clinical and health economic evidence relevant to the review questions. These were run according to the parameters as stipulated within the NICE guideline's manual, https://www.nice.org.uk/process/pmg20/chapter/identifying-the-evidence-literature-searching-and-evidence-submission.

Databases were searched using relevant medical subject headings, free-text terms and where appropriate study-type filters. Studies published in languages other than English were not reviewed, and where possible, searches were restricted to English language. Searches were updated on 26 September 2023. Where original searches generated no evidence searches were not updated. Papers published or added to databases after this date were not considered. Where new evidence was identified,

for example in consultation comments received from stakeholders, the impact on the guideline was considered, and the action agreed between the technical team and NICE staff with a quality assurance role.

Searches were quality assured using different approaches prior to being run. Medline search strategies were peer reviewed by a second information specialist using a QA process based on the PRESS checklist.[3] Key (seed) papers if provided, were checked if retrieved by the search.

Searching for unpublished literature was not undertaken. NICE do not have access to drug manufacturers' unpublished clinical trial results, so the clinical evidence considered by the committee for pharmaceutical interventions may be different from that considered by the MHRA and European Medicines Agency for the purposes of licensing and safety regulation.

Additional studies were added to the evidence base these consisted of references included in relevant systematic reviews, and those highlighted by committee members.

During the scoping stage, a search was conducted for guidelines and reports on the websites including:

- Guidelines International Network database (www.g-i-n.net)
- ECRI (www.ecri.org)
- TRIP Medical Database (www.tripdatabase.com)
- Society for Endocrinology (www.endocrinology.org)
- Endocrine Society (www.endocrine.org)
- European Society of Endocrinology (www.ese-hormones.org)

2.2.2 Call for evidence

This was initiated where the committee believed that there was relevant evidence in addition to that identified by the searches. This process is outlined in section 5.5 of Developing NICE guidelines: the manual 2014.[4] The committee decided to initiate a 'call for evidence' for the evidence review on emergency pharmaceutical management of adrenal insufficiency, specifically on current practice in the UK.

2.3 Reviewing evidence

The evidence for each review question was reviewed using the following process:

- Potentially relevant studies were identified from the search results by reviewing titles and abstracts. The full papers were then obtained.
- Full papers were evaluated against the pre-specified inclusion and exclusion criteria set out in the protocol to identify studies that addressed the review question. The review protocols are included in an appendix to each of the evidence reports.
- Relevant studies were critically appraised using the preferred study design checklist as specified in the NICE guidelines manual.[4] The checklist used is included in the individual review protocols in each of the evidence reports.

- Key information was extracted about interventional study methods and results into EPPI reviewer version 5. Summary evidence tables were produced from data entered into EPPI Reviewer, including critical appraisal ratings.
- Summaries of the evidence were generated by outcome. Outcome data were combined, analysed and reported according to study design:
  - Randomised data were meta-analysed where appropriate and reported in GRADE evidence profiles.
  - Data from non-randomised studies were meta-analysed where appropriate and reported in GRADE evidence profiles.
  - Prognostic data were meta-analysed where appropriate and reported in adapted GRADE evidence profiles.
  - Diagnostic data were meta-analysed where appropriate or presented as a range of values in GRADE evidence profiles.
  - Qualitative data were synthesised across studies using thematic analysis and presented as summary statements in GRADE CERQual tables.
- A minimum of 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.
- All of the evidence reviews were quality assured by a senior systematic reviewer. This included checking:
  - papers were included or excluded appropriately.
  - a sample of the data extractions
  - a sample of the risk of bias assessments
  - correct methods were used to synthesise data.

  Discrepancies will be identified and resolved through discussion (with a third reviewer where necessary).

2.3.1 Types of studies and inclusion and exclusion criteria

The inclusion and exclusion of studies was based on the criteria defined in the review protocols, which can be found in an appendix to each of the evidence reports. Excluded studies (with the reasons for their exclusion) are listed in an appendix to each of the evidence reports. The committee was consulted about any uncertainty regarding inclusion or exclusion.

Conference abstracts were not considered for inclusion. If abstracts were included the authors were contacted for further information. Literature reviews, posters, letters, editorials, comment articles, unpublished studies, and studies not in published in English language were also excluded.

2.3.1.1 Type of studies

Randomised controlled trials, non-randomised intervention studies, and other observational studies (including diagnostic or prognostic studies) were included in the evidence reviews as appropriate.

For intervention reviews, randomised controlled trials (RCTs) were included where identified as because they are considered the most robust type of study design that can produce an unbiased estimate of the intervention effects. Non-randomised intervention studies were considered appropriate for inclusion if there was insufficient

randomised evidence for the committee to make a decision. In this case the committee stated a priori in the protocol that either certain identified variables must be equivalent at baseline or else the analysis had to adjust for any baseline differences. If the study did not fulfil either criterion it was excluded. Refer to the review protocols in each evidence report for full details on the study design of studies that were appropriate for each review question.

For diagnostic review questions, cross-sectional studies and prospective studies were included. For prognostic review questions, prospective studies were included. Retrospective studies were only included if evidence from cross sectional or prospective studies was insufficient to inform decision making. Case–control studies were not included.

Systematic reviews and meta-analyses conducted to the same methodological standards as the NICE reviews were included within the evidence reviews in preference to primary studies, where they were available and applicable to the review questions and updated or added to where appropriate to the guideline review question. Individual patient data (IPD) meta-analyses were preferentially included if meeting the protocol and methodological criteria.

2.3.1.1.1 Qualitative studies

In the qualitative reviews, studies using focus groups, or structured or semi-structured interviews were considered for inclusion. Survey data or other types of questionnaires were only included if they provided analysis from open-ended questions, but not if they reported descriptive quantitative data only.

2.4 Methods of combining evidence

2.4.1 Data synthesis for intervention reviews

Meta-analyses were conducted using Cochrane Review Manager (RevMan5)[9] software

2.4.1.1 Analysis of different types of data

Dichotomous outcomes

Fixed-effects (Mantel–Haenszel) techniques were used to calculate risk ratios (relative risk, RR) for the binary outcomes. The absolute risk difference was also calculated using GRADEpro[1] software, using the median event rate in the control arm of the pooled results.

For binary variables where there were zero events in either arm or a less than 1% event rate, Peto odds ratios, rather than risk ratios, were calculated as they are more appropriate for data with a low number of events. Where there are zero events in both arms, the risk difference was calculated and reported instead.

Continuous outcomes

Continuous outcomes were analysed using an inverse variance method for pooling weighted mean differences.

Where the studies within a single meta-analysis had different scales of measurement for the same outcomes, standardised mean differences were used (providing all studies reported either change from baseline or final values rather than a mixture of both); each different measure in each study was 'normalised' to the standard deviation value pooled between the intervention and comparator groups in that same study.

The means and standard deviations of continuous outcomes are required for meta-analysis. However, in cases where standard deviations were not reported, the standard error was calculated if the p values or 95% confidence intervals (95% CI) were reported, and meta-analysis was undertaken with the mean and standard error using the generic inverse variance method in RevMan5.[9]

Generic inverse variance

If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into RevMan5.[9] If the control event rate was reported this was used to generate the absolute risk difference in GRADEpro.[1] If multivariate analysis was used to derive the summary statistic but no adjusted control event rate was reported no absolute risk difference was calculated.

Complex analysis

Where studies had used a crossover design, paired continuous data were extracted where possible, and forest plots were generated in RevMan5[9] with the generic inverse variance function. When a crossover study had categorical data and the number of subjects with an event in both interventions was known, the standard error (of the log of the risk ratio) was calculated using the simplified Mantel–Haenszel method for paired outcomes. Forest plots were also generated in RevMan5[9] with the generic inverse variance function. If paired continuous or categorical data were not available from the crossover studies, the separate group data were analysed in the same way as data from parallel groups, on the basis that this approach would overestimate the confidence intervals and thus artificially reduce study weighting resulting in a conservative effect. Where a meta-analysis included a mixture of studies using both paired and parallel group approaches, all data were entered into RevMan5[9] using the generic inverse variance function.

2.4.1.2 Diagnostic accuracy studies

For diagnostic test accuracy studies, a positive result on the index test was found if the person had values of the measured quantity above or below a threshold value, and different thresholds could be used. The thresholds were pre-specified by the committee including whether or not data could be pooled across a range of thresholds. The threshold of a diagnostic test is defined as the value at which the test can best differentiate between those with and without the target condition. In practice this usually varies across studies. If a test has a high sensitivity, then very few people with the condition will be missed (few false negatives). For example, a test with a sensitivity of 97% will only miss 3% of people with the condition. Conversely, if a test has a high specificity, then few people without the condition would be incorrectly diagnosed (few false positives).

Coupled forest plots of the agreed primary paired outcome measure for decision making (sensitivity and specificity) with their 95% CIs across studies (at various thresholds) were produced for each test, using RevMan5.[9] In order to do this, 2 by 2

tables (the number of true positives, false positives, true negatives and false negatives) were directly taken from the study if given, or else were derived from raw data or calculated from the set of test accuracy statistics.

Diagnostic meta-analysis was conducted where appropriate, that is, when 3 or more studies were available per threshold. Test accuracy for the studies was pooled using the bivariate method for the direct estimation of summary sensitivity and specificity using a random-effects approach in WinBUGS software.[10] The advantage of this approach is that it produces summary estimates of sensitivity and specificity that account for the correlation between the 2 statistics. The bivariate method uses logistic regression on the true positives, true negatives, false positives, and false negatives reported in the studies. Overall sensitivity and specificity and confidence regions were plotted (using methods outlined by Novielli 2010.[7]) The pooled median sensitivity and specificity and their 95% CIs were reported in the clinical evidence summary tables. For analyses with fewer than 3 studies included, the results of the study with the lower sensitivity value were reported when there were 2 studies or reported individually for a single study.

If appropriate, to allow comparison between tests, summary ROC curves were generated for each diagnostic test from the pairs of sensitivity and specificity calculated from the 2by 2 tables, selecting 1 threshold per study. A ROC plot shows true positive rate (sensitivity) as a function of false positive rate (1 minus specificity). Data were entered into RevMan5[9] and ROC curves were fitted using the Moses-Littenberg approach. In order to compare diagnostic tests, 2 or more tests were plotted on the same graph. The performance of the different diagnostic tests was then assessed by examining the summary ROC curves visually: the test that had a curve lying closest to the upper left corner (100% sensitivity and 100% specificity) was interpreted as the best test.

A second analysis was conducted by restricting the set of studies to those with the same clinically relevant threshold as agreed by the committee, to ensure the data were comparable. They were presented as forest plots and ROC curves and heterogeneity was investigated.

Area under the ROC curve (AUC) data for each study were also plotted on a graph, for each diagnostic test. The AUC describes the overall diagnostic accuracy across the full range of thresholds. The following criteria were used for evaluating AUCs:

- ≤0.50: worse than chance
- 0.50–0.60: very poor
- 0.61–0.70: poor
- 0.71–0.80: moderate
- 0.81–0.90: good
- 0.91–1.00: excellent or perfect test.

Heterogeneity or inconsistency amongst studies was visually inspected.

2.4.2 Data synthesis for qualitative reviews

The main findings for each included paper were identified and thematic analysis methods were used to synthesise this information into broad overarching themes which were summarised into the main review findings. The evidence was presented in the form of a narrative summary detailing the evidence from the relevant papers

and how this informed the overall review finding plus a statement on the level of confidence for that review finding. Considerable limitations and issues around relevance were listed. A summary evidence table with the succinct summary statements for each review finding was produced including the associated quality assessment.

2.5 Appraising the quality of evidence by outcomes

2.5.1 Intervention reviews

The evidence for outcomes from the included RCTs and, where appropriate, non-randomised intervention studies, were evaluated and presented using the 'Grading of Recommendations Assessment, Development and Evaluation (GRADE) toolbox' developed by the international GRADE working group (http://www.gradeworkinggroup.org/). The software (GRADEpro[1]) developed by the GRADE working group was used to assess the quality of each outcome, taking into account individual study quality and the meta-analysis results.

Each outcome was first examined for each of the quality elements listed and defined in Table 3.

Table 2: Description of quality elements in GRADE for intervention studies

Quality element	Description
Risk of bias	Limitations in the study design and implementation may bias the estimates of the treatment effect. Major limitations in studies decrease the confidence in the estimate of the effect. Examples of such limitations are selection bias (often due to poor allocation concealment), performance and detection bias (often due to a lack of blinding of the patient, healthcare professional or assessor) and attrition bias (due to missing data causing systematic bias in the analysis).
Indirectness	Indirectness refers to differences in study population, intervention, comparator and outcomes between the available evidence and the review question.
Inconsistency	Inconsistency refers to an unexplained heterogeneity of effect estimates between studies in the same meta-analysis.
Imprecision	Results are imprecise when studies include relatively few patients and few events (or highly variable measures) and thus have wide confidence intervals around the estimate of the effect relative to clinically important thresholds. 95% confidence intervals denote the possible range of locations of the true population effect at a 95% probability, and so wide confidence intervals may denote a result that is consistent with conflicting interpretations (for example a result may be consistent with both clinical benefit AND clinical harm) and thus be imprecise.
Publication bias	Publication bias is a systematic underestimate or overestimate of the underlying beneficial or harmful effect due to the selective publication of studies. A closely related phenomenon is where some papers fail to report an outcome that is inconclusive, thus leading to an overestimate of the effectiveness of that outcome.
Other issues	Sometimes randomisation may not adequately lead to group equivalence of confounders, and if so this may lead to bias, which should be taken into account. Potential conflicts of interest, often caused by excessive pharmaceutical company involvement in the publication of a study, should also be noted.

Details of how the 4 main quality elements (risk of bias, indirectness, inconsistency and imprecision) were appraised for each outcome are given below. Publication bias

was considered with the committee. If there was reason to suspect it was present, it was explored with funnel plots. Funnel plots were constructed using RevMan5 software to assess against potential publication bias for outcomes containing more than 5 studies. This was taken into consideration when assessing the quality of the evidence.

2.5.1.1 Risk of bias

The main domains of bias for RCTs are listed in Table 4. Each outcome had its risk of bias assessed within each study first using the appropriate checklist for the study design (Cochrane RoB 2 for RCTs, or ROBINS-I for non-randomised studies or ROBIS for systematic reviews). For each study, if there was no risk of bias in any domain, the risk of bias was given a rating of 'low risk of bias'. An overall judgment of 'some concerns' was made if some concerns were present in at least one domain and the domain was judged to be at high risk of bias. An overall judgment of 'high risk of bias' was made if high risk domains in a way that substantially lowers confidence in the result. An overall rating is of: not serious, serious or very serious, is applied in GRADEpro across all studies combined in a meta-analysis by taking into account the weighting of studies according to study precision.

Table 3: Principle domains of bias in randomised controlled trials

Limitation	Explanation
Selection bias (sequence generation and allocation concealment)	If those enrolling participants are aware of the group to which the next enrolled patient will be allocated, either because of a non-random sequence that is predictable, or because a truly random sequence was not concealed from the researcher, this may translate into systematic selection bias. This may occur if the researcher chooses not to recruit a participant into that specific group because of: • knowledge of that participant's likely prognostic characteristics, and • a desire for one group to do better than the other.
Performance and detection bias (lack of blinding)	Patients, caregivers, those adjudicating or recording outcomes, and data analysts should not be aware of the arm to which the participants are allocated. Knowledge of the group can influence: • the experience of the placebo effect • performance in outcome measures • the level of care and attention received, and • the methods of measurement or analysis all of which can contribute to systematic bias.
Attrition bias	Attrition bias results from an unaccounted-for loss of data beyond a certain level (a differential of at least 10% between groups). Loss of data can occur when participants are compulsorily withdrawn from a group by the researchers (for example, when a per-protocol approach is used) or when participants do not attend assessment sessions. If the missing data are likely to be different from the data of those remaining in the groups, and there is a differential rate of such missing data from groups, systematic attrition bias may result.
Selective outcome reporting	Reporting of some outcomes and not others on the basis of the results can also lead to bias, as this may distort the overall impression of efficacy.
Other limitations	For example: • Stopping early for benefit observed in randomised trials, in particular in the absence of adequate stopping rules. • Use of unvalidated patient-reported outcome measures.

Limitation	Explanation
	• Lack of washout periods to avoid carry-over effects in crossover trials. • Recruitment bias in cluster-randomised trials.

The assessment of risk of bias differs for non-randomised intervention studies, due to the possibility of confounding and the greater risk of selection bias. The assessment of risk of bias therefore requires a different checklist (ROBINS-I) and involves consideration of more domains and varies by study type. **Table 5** shows the domains considered for most types of non-randomised studies.

Table 4 Principal domains of bias in non-randomised studies

Bias	Explanation
Pre-intervention	
Confounding bias	Baseline confounding occurs when one or more prognostic variables (factors that predict the outcome of interest) also predicts the intervention received at baseline. ROBINS-I can also address time-varying confounding, which occurs when post-baseline prognostic factors affect the intervention received after baseline.
Selection bias	When exclusion of some eligible participants, or the initial follow-up time of some participants, or some outcome events, is related to both intervention and outcome, there will be an association between interventions and outcome even if the effect of interest is truly null. This type of bias is distinct from confounding. A specific example is bias due to the inclusion of prevalent users, rather than new users, of an intervention.
At intervention	
Information bias	Bias introduced by either differential or non-differential misclassification of intervention status. Non-differential misclassification is unrelated to the outcome and will usually bias the estimated effect of intervention towards the null. Differential misclassification occurs when misclassification of intervention status is related to the outcome or the risk of the outcome.
Post-intervention	
Confounding bias	Bias that arises when there are systematic differences between experimental intervention and comparator groups in the care provided, which represent a deviation from the intended intervention(s). Assessment of bias in this domain will depend on the effect of interest (either the effect of assignment to intervention or the effect of adhering to intervention).
Selection bias	Bias that arises when later follow-up is missing for individuals initially included and followed (e.g., differential loss to follow-up that is affected by prognostic factors); bias due to exclusion of individuals with missing information about intervention status or other variables such as confounders.
Information bias	Bias introduced by either differential or non-differential errors in measurement of outcome data. Such bias can arise when outcome assessors are aware of intervention status, if different methods are used to assess outcomes in different intervention groups, or if measurement errors are related to intervention status or effects.
Reporting bias	Selective reporting of results from among multiple measurements of the outcome, analyses or subgroups in a way that depends on the findings.

2.5.1.2 Indirectness

Indirectness refers to the extent to which the populations, interventions, comparisons, and outcome measures are dissimilar to those defined in the inclusion criteria for the

reviews. Indirectness is important when these differences are expected to contribute to a difference in effect size or may affect the balance of harms and benefits considered for an intervention. As for the risk of bias, each outcome had its indirectness assessed within each study first. For each study, if there were no sources of indirectness, indirectness was given a rating of 'directly applicable'. If there was indirectness in just 1 source (for example in terms of population), indirectness was given a rating of 'partially applicable', but if there was indirectness in 2 or more sources (for example, in terms of population and treatment) the indirectness was given an 'indirectly applicable' rating. An overall rating of; not serious, serious, or very serious, was applied GRADEpro across all studies by taking into account the weighting of studies according to study precision.

2.5.1.3 Inconsistency

Inconsistency refers to an unexplained heterogeneity of results for an outcome across different studies. When estimates of the treatment effect across studies differ widely, this suggests true differences in the underlying treatment effect, which may be due to differences in populations, settings, or doses. Statistical heterogeneity was assessed for each meta-analysis estimate by an I-squared (I^2) inconsistency statistic.

Heterogeneity or inconsistency amongst studies was also visually inspected. Where statistical heterogeneity as defined above was present or there was clear visual heterogeneity not captured in the I^2 value predefined subgrouping of studies was carried out according to the protocol. See the review protocols for the subgrouping strategy.

When heterogeneity existed within an outcome (I^2>50%), but no plausible explanation could be found, the quality of evidence for that outcome was downgraded. Inconsistency for that outcome was given a 'serious' rating if the I^2 was 50–74%, and a 'very serious' rating if the I^2 was 75% or more.

If inconsistency could be explained based on pre-specified subgroup analysis (that is, each subgroup had an I^2<50%) then each of the derived subgroups were presented separately for that forest plot and GRADE profile (providing at least 2 studies remained in each subgroup). The committee took this into account and considered whether to make separate recommendations based on the variation in effect across subgroups within the same outcome. In such a situation the quality of evidence was not downgraded.

If all predefined strategies of subgrouping were unable to explain statistical heterogeneity, then a random effects (DerSimonian and Laird) model was employed to the entire group of studies in the meta-analysis. A random-effects model assumes a distribution of populations, rather than a single population. This leads to a widening of the confidence interval around the overall estimate. If, however, the committee considered the heterogeneity was so large that meta-analysis was inappropriate, then the results were not pooled and were described narratively.

2.5.1.4 Imprecision

The criteria applied for imprecision were based on the 95% CIs for the pooled estimate of effect, and the minimal important differences (MID) for the outcome. The MIDs are the threshold for appreciable benefits and harms, separated by a zone either side of the line of no effect where there is assumed to be no clinically important

effect. If either end of the 95% CI of the overall estimate of effect crossed 1 of the MID lines, imprecision was regarded as serious in the GRADEpro rating. This was because the overall result, as represented by the span of the confidence interval, was consistent with 2 interpretations as defined by the MID (for example, both no clinically important effect and clinical benefit were possible interpretations). If both MID lines were crossed by either or both ends of the 95% CI, then imprecision was regarded as very serious. This was because the overall result was consistent with all 3 interpretations defined by the MID (no clinically important effect, clinical benefit, and clinical harm). This is illustrated in Figure 1.

The value / position of the MID lines is ideally determined by values reported in the literature. 'Anchor-based' methods aim to establish clinically meaningful changes in a continuous outcome variable by relating or 'anchoring' them to patient-centred measures of clinical effectiveness that could be regarded as gold standards with a high level of face validity. For example, a MID for an outcome could be defined by the minimum amount of change in that outcome necessary to make patients feel their quality of life had 'significantly improved'. MIDs in the literature may also be based on expert clinician or consensus opinion concerning the minimum amount of change in a variable deemed to affect quality of life or health.

In the absence of values identified in the literature, the alternative approach to deciding on MID levels is to use the modified GRADE 'default' values, as follows:

- For dichotomous outcomes the MIDs were taken to be RRs of 0.8* and 1.25. For 'positive' outcomes such as 'patient satisfaction', the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit. For 'negative' outcomes such as 'bleeding', the opposite occurs, so the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm. There aren't established default values for ORs, and the same values (0.8 and 1.25) are applied here but are acknowledged as arbitrary thresholds agreed by the committee.
    - In cases where there are zero events in one arm of a single study, or some or all of the studies in one arm of a meta-analysis, the same process is followed as for dichotomous outcomes. However, if there are no events in either arm in a meta-analysis (or in a single un-pooled study) the sample size is used to determine imprecision using the following rule of thumb:
        - No imprecision: sample size ≥350
        - Serious imprecision: sample size ≥70 but <350
        - Very serious imprecision: sample size <70.
    - When there was more than one study in an analysis and zero events occurred in both groups for some but not all of the studies across both arms, the optimum information size was used to determine imprecision using the following guide:
        - No imprecision: >90% power
        - Serious imprecision: 80-90% power
        - Very serious imprecision: <80% power.
- For mortality any change was considered to be clinically important, and the imprecision was assessed on the basis of the whether the confidence intervals

crossed the line of no effect, that is whether the result was consistent with both benefit and harm.

- For continuous outcome variables the MID was taken as half the median baseline standard deviation of that variable, across all studies in the meta-analysis. Hence the MID denoting the minimum clinically important benefit was positive for a 'positive' outcome (for example, a quality-of-life measure where a higher score denotes better health), and negative for a 'negative' outcome (for example, a visual analogue scale [VAS] pain score). Clinically important harms will be the converse of these. If baseline values are unavailable, then half the median comparator group standard deviation of that variable will be taken as the MID. As these vary for each outcome per review, details of the values used are reported in the footnotes of the relevant GRADE summary table.

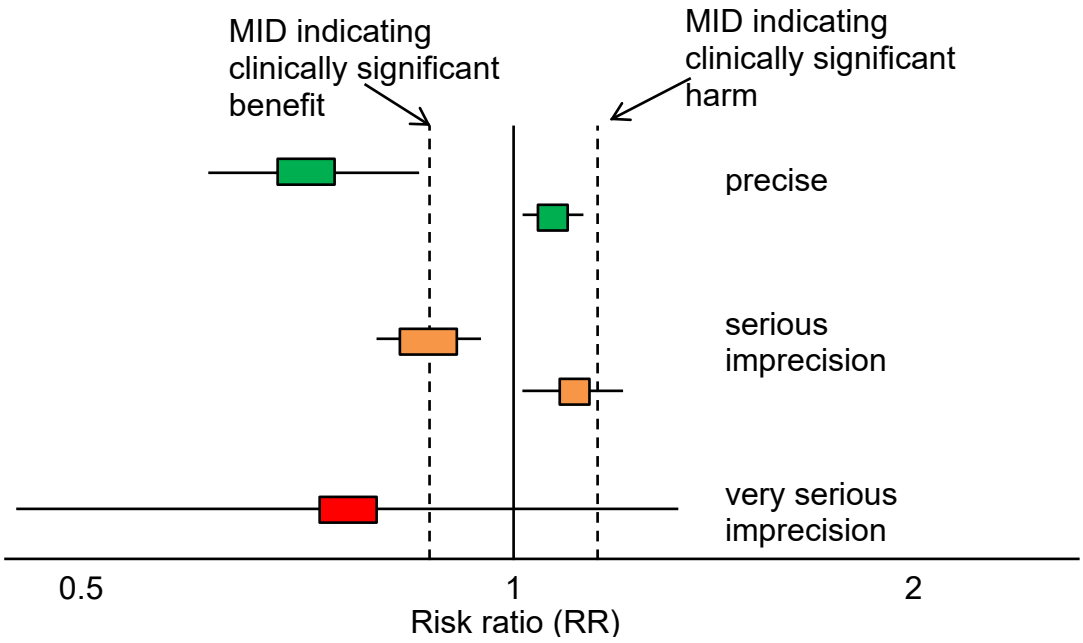
*NB GRADE report the default values as 0.75 and 1.25. These are consensus values. This guideline follows NICE process to use modified values of 0.8 and 1.25 as they are symmetrical on a relative risk scale.

For this guideline, the following MIDs for continuous or dichotomous outcomes were found in the literature and adopted for use:

Table 5: Published or pre-agreed MIDs

Outcome measure	MID	Source
EQ-5D	0.03	Consensus pragmatic MID used in some previous NICE guidelines
SF36	Physical component summary: 2 Mental component summary: 3 Physical functioning: 3 Role-physical: 3 Bodily pain: 3 General health: 2 Vitality: 2 Social functioning: 3 Role-emotional: 4 Mental health: 3	User's manual for the SF-36v2 Health Survey, Third Edition[2]

Figure 1: Illustration of precise and imprecise outcomes based on the 95% CI of dichotomous outcomes in a forest plot (Note that all 3 results would be pooled estimates, and would not, in practice, be placed on the same forest plot)



2.5.1.5 Overall grading of the quality of clinical evidence

Once an outcome had been appraised for the main quality elements, as above, an overall quality grade was calculated for that outcome from the ratings from each of the main quality elements were summed to give a score that could be anything from high to very low. The evidence for each outcome started at High, and the overall quality (or confidence in the evidence) remained High if there were no reasons for downgrading, or became Moderate, Low or Very Low according to the number of independent reasons for downgrading. The significance of these overall ratings is explained in Table 7. The reasons for downgrading in each case are specified in the footnotes of the GRADE tables.

Table 6: Overall quality of outcome evidence in GRADE

Level	Description
High	Further research is very unlikely to change our confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very low	Any estimate of effect is very uncertain

2.5.2 Diagnostic reviews

2.5.2.1 Diagnostic test accuracy

2.5.2.1.1 Risk of bias

Risk of bias and indirectness of evidence for diagnostic data were evaluated by study using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklists (see appendix H in the NICE guidelines manual 2014[4]). Risk of bias and applicability in primary diagnostic accuracy studies in QUADAS-2 consists of 4 domains (see **Table 8**):

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

Table 7 Summary of QUADAS-2 with list of signalling, risk of bias and applicability questions.

Domain	Patient selection	Index test	Reference standard	Flow and timing
Description	Describe methods of patient selection. Describe included patients (prior testing, presentation, intended use of index test and setting)	Describe the index test and how it was conducted and interpreted	Describe the reference standard and how it was conducted and interpreted	Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2×2 table (refer to flow diagram). Describe the time interval and any interventions between index test(s) and reference standard
Signalling questions (yes/no/ unclear)	Was a consecutive or random sample of patients enrolled?	Were the index test results interpreted without knowledge of the results of the reference standard?	Is the reference standard likely to correctly classify the target condition?	Was there an appropriate interval between index test(s) and reference standard?
	Was a case–control design avoided?	If a threshold was used, was it pre-specified?	Were the reference standard results interpreted without knowledge of the results of the index test?	Did all patients receive a reference standard?
	Did the study avoid inappropriate exclusions?			Did all patients receive the same reference standard?
				Were all patients included in the analysis?
Risk of bias; (high/low/ unclear)	Could the selection of patients have introduced bias?	Could the conduct or interpretation of the index test	Could the reference standard, its conduct or its	Could the patient flow have introduced bias?

Domain	Patient selection	Index test	Reference standard	Flow and timing
		have introduced bias?	interpretation have introduced bias?	
Concerns regarding applicability (high/low/ unclear)	Are there concerns that the included patients do not match the review question?	Are there concerns that the index test, its conduct, or interpretation differ from the review question?	Are there concerns that the target condition as defined by the reference standard does not match the review question?	

2.5.2.1.2 Inconsistency

Inconsistency refers to an unexplained heterogeneity of results for an outcome across different studies. Inconsistency was assessed by visual inspection of the primary outcome measures (sensitivity and specificity) using the point estimates and 95% CIs of the individual studies on the forest plots or the summary value if a diagnostic meta-analysis had been conducted. The evidence was downgraded by 1 increment if there was no overlap of 95% confidence intervals or by 2 increments if there was wide variability. only a single study reports an outcome, inconsistency is rated as 'not detected'.

2.5.2.1.3 Imprecision

The judgement of precision was based on visual inspection of the confidence region around the summary sensitivity and specificity point from the diagnostic meta-analysis, if a diagnostic meta-analysis was conducted. Where a diagnostic meta-analysis was not conducted, imprecision was assessed according to the range of point estimates or, if only one study contributed to the evidence, the 95% CI around the single study. The decision thresholds set by the committee were used to determine whether imprecision is not serious, serious, or very serious depending on whether confidence intervals cross zero, one or two thresholds. In situations where the 95% CI's were not reported by the study and the raw data was not available to calculate them, then the outcome was downgraded twice for imprecision.

2.5.2.1.4 Overall grading

Quality rating started at high for prospective and retrospective cross-sectional studies, and each major limitation (risk of bias, indirectness, inconsistency, and imprecision) brought the rating down by 1 increment to a minimum grade of very low, as explained for intervention reviews. This was presented in a GRADE evidence profile.

2.5.3 Prognostic reviews

An adapted GRADE evidence profile was used for quality assessment per outcome. If data were meta-analysed, the quality for pooled studies was presented. If the data were not pooled, then a quality rating was presented for each study.

2.5.3.1.1 Risk of bias

The risk of bias for prognostic studies was evaluated according to the QUIPS checklist, the main criteria are given in **Table 9**.

Table 8: Description of risk of bias criteria for prognostic studies

Risk of bias	Aim of section
Study participation	To judge selection bias (likelihood that relationship between the prognostic factor and outcome is different for participants and eligible non-participants)
Study attrition	To judge the risk of attrition bias (likelihood that relationship between prognostic factor and outcome are different for completing and non-completing participants).
Prognostic factor measurement	To judge the risk of measurement bias related to how the prognostic factor was measured (differential measurement of prognostic factor related to the baseline level of outcome).
Outcome measurement	To judge the risk of bias related to the measurement of outcome (differential measurement of outcome related to the baseline level of prognostic factor).
Study confounding	To judge the risk of bias due to confounding (i.e. the effect of the prognostic factor is distorted by another factor that is related to the prognostic factor and outcome).
Statistical Analysis and Reporting	To judge the risk of bias related to the statistical analysis and presentation of results.

2.5.3.1.2 Inconsistency

Inconsistency was assessed as for intervention studies.

2.5.3.1.3 Imprecision

In meta-analysed outcomes, or for non-pooled outcomes, the position of the 95% CIs in relation to the null line determined the existence of imprecision. If the 95% CI did not cross the null line, then no serious imprecision was recorded. If the 95% CI crossed the null line, then serious imprecision was recorded.

2.5.3.1.4 Overall grading

Quality rating was assigned by study. However, if there was more than 1 outcome involved in a study, then the quality rating of the evidence statements for each outcome was adjusted accordingly. For example, if one outcome was based on an invalidated measurement method, but another outcome in the same study was not, the second outcome would be graded 1 grade higher than the first outcome.

Quality rating started at high for prospective studies and each major limitation brought the rating down by 1 increment to a minimum grade rating of very low, as explained for interventional reviews. For prognostic reviews prospective cohort studies with a multivariate analysis are regarded as the gold standard because RCTs are usually an inappropriate design to answer the question for these types of review. Furthermore, if the study is looking at more than 1 prognostic factor of interest then randomisation would be inappropriate as it can only be applied to 1 of the prognostic factors.

2.5.4 Qualitative reviews

Review findings from the included qualitative studies were evaluated and presented using the 'Confidence in the Evidence from Reviews of Qualitative Research' (CERQual) Approach developed by the GRADE-CERQual Project Group, a subgroup of the GRADE Working Group.

The CERQual Approach assesses the extent to which a review finding is a reasonable representation of the phenomenon of interest (the focus of the review question). Each review finding was assessed for each of the 4 quality elements listed and defined below in Table 10.

Table 9: Description of quality elements in GRADE-CERQual for qualitative studies

Quality element	Description
Methodological limitations	The extent of problems in the design or conduct of the included studies that could decrease the confidence that the review finding is a reasonable representation of the phenomenon of interest. Assessed at the study level using the CASP checklist.
Coherence	The extent to how clear and cogent the fit is between the data from the primary studies and the review finding.
Relevance	The extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol.
Adequacy	The degree of the confidence that the review finding is being supported by sufficient data. This is an overall determination of the richness (depth of analysis) and quantity of the evidence supporting a review finding or theme.

Details of how the 4 quality elements (methodological limitations, coherence, relevance, and adequacy) were appraised for each review finding are given below.

2.5.4.1 Methodological limitations

Each review finding had its methodological limitations assessed within each study first using the CASP checklist. Based on the degree of methodological limitations, studies were evaluated as having minor, moderate or severe limitations. A summary of the domains and questions covered is given below.

Table 10: Description of limitations assessed in the CASP checklist for qualitative studies.

Domain	Aspects considered
Are the results valid?	• Was there a clear statement of the aims of the research? • Is qualitative methodology appropriate? • Was the research design appropriate to address the aims of the research? • Was the recruitment strategy appropriate to the aims of the research? • Was the data collected in a way that addressed the research issue? • Has the relationship between researcher and participants been adequately considered?
What are the results?	Have ethical issues been taken into consideration? Was the data analysis sufficiently rigorous?

Domain	Aspects considered
	Is there a clear statement of findings?
Will the results help locally?	How valuable is the research?

The overall assessment of the methodological limitations of the evidence was based on the limitations of the primary studies contributing to the review finding. The relative contribution of each study to the overall review finding and of the type of methodological limitation(s) were taken into account when giving an overall rating of concerns for this component.

2.5.4.2 Relevance

Relevance is the extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol. As such, relevance is dependent on the individual review and discussed with the guideline committee.

2.5.4.3 Coherence

Coherence is the extent to which the reviewer is able to identify a clear pattern across the studies included in the review, and if there is variation present (contrasting or disconfirming data) whether this variation is explained by the contributing study authors. For example, if a review finding in 1 study does not support the main finding and there is no plausible explanation for this variation, or if there is ambiguity in the descriptions in the primary data, then the confidence that the main finding reasonably reflects the phenomenon of interest is decreased.

2.5.4.4 Adequacy

The judgement of adequacy is based on the confidence of the finding being supported by sufficient data. This is an overall determination of the richness (and quantity of the evidence supporting a review finding or theme. Rich data provide sufficient detail to gain an understanding of the theme or review finding, whereas thin data do not provide enough detail for an adequate understanding. Quantity of data is the second pillar of the assessment of adequacy. For review findings that are only supported by 1 study or data from only a small number of participants, the confidence that the review finding reasonably represents the phenomenon of interest might be decreased because there is less confidence that studies undertaken in other settings or participants would have reported similar findings. As with richness of data, quantity of data is review dependent. Based on the overall judgement of adequacy, a rating of no concerns, minor concerns, or substantial concerns about adequacy was given.

2.5.4.5 Overall judgement of the level of confidence for a review finding

GRADE-CERQual is used to assess the body of evidence as a whole through a confidence rating representing the extent to which a review finding is a reasonable representation of the phenomenon of interest. For each of the above components, level of concern is categorised as either:

- no or very minor concerns
- minor concerns
- moderate concerns, or

- serious concerns.

The concerns from the 4 components (methodological limitations, coherence, relevance, and adequacy) are used in combination to form an overall judgement of confidence in the finding. GRADE-CERQual uses 4 levels of confidence: high, moderate, low, and very low confidence. The significance of these overall ratings is explained in Table 12. Each review finding starts at a high level of confidence and is downgraded based on the concerns identified in any 1 or more of the 4 components. Quality assessment of qualitative reviews is a subjective judgement by the reviewer based on the concerns that have been noted. An explanation of how such a judgement had been made for each component is included in the footnotes of the summary of evidence tables.

Table 11: Overall level of confidence for a review finding in GRADE-CERQual

Level	Description
High confidence	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest.
Moderate confidence	It is likely that the review finding is a reasonable representation of the phenomenon of interest.
Low confidence	It is possible that the review finding is a reasonable representation of the phenomenon of interest.
Very low confidence	It is not clear whether the review finding is a reasonable representation of the phenomenon of interest.

2.6 Assessing clinical importance

The committee assessed the evidence by outcome in order to determine if there was, or potentially was, a clinically important benefit, a clinically important harm or no clinically important difference between interventions. To facilitate this, binary outcomes were converted into absolute risk differences (ARDs) using GRADEpro[1] software: the median control group risk across studies was used to calculate the ARD and its 95% CI from the pooled risk ratio.

The assessment of clinical benefit, harm, or no benefit or harm was based on the point estimate of absolute effect for intervention studies, which was standardised across the reviews. The committee considered for most of the dichotomous outcomes in the intervention reviews that if at least 100 more participants per 1000 (10%) achieved the outcome of interest in the intervention group compared to the comparison group for a positive outcome then this intervention was considered beneficial. The same point estimate but in the opposite direction applied for a negative outcome. For mortality any reduction represented a clinical benefit. For adverse events 50 events or more per 1000 (5%) represented clinical harm.]

For continuous outcomes if the mean difference was greater than the minimally important difference (MID) then this represented a clinical benefit or harm. For outcomes such as mortality any reduction or increase was considered to be clinically important.

Established MIDs found in the literature and were agreed to be used for Quality-of-Life measures SF-36 and EQ5D.

The published values used for imprecision and clinical importance are provided in **Table 6**. For continuous outcomes where the GRADE default MID has been used,

the values for each outcome are provided in the footnotes of the relevant GRADE tables.

2.7 Appraising the quality of external guidelines

External guidelines were appraised using a two-stage process. The first stage assessed whether the external guideline development process was robust and high quality. This was conducted using the Appraisal of Guidelines for Research and Evaluation (AGREE) II instrument

The second stage of the appraisal process assessed the applicability and acceptability of the external recommendations themselves. It covered areas that are important for NICE such as the quality guideline development process, barriers to implementation, compatibility with cultures and values and health inequalities.

First stage assessment using the AGREE II instrument:

The AGREE II instrument is an internationally validated tool that is used to assess the methodological rigour and transparency of clinical practice guidelines.

The AGREE II tool was used to assess the following domains:

- Domain 1**. Scope and Purpose** are concerned with the overall aim of the guideline, the specific health questions, and the target population.
- Domain 2. **Stakeholder Involvement** focuses on the extent to which the guideline was developed by the appropriate stakeholders and represents the views of its intended users.
- Domain 3. **Rigour of Development** relates to the process used to gather and synthesize the evidence, the methods to formulate the recommendations, and to update them.
- Domain 4. **Clarity of Presentation** deals with the language, structure, and format of the guideline.
- Domain 5. **Applicability** pertains to the likely barriers and facilitators to implementation, strategies to improve uptake, and resource implications of applying the guideline.
- Domain 6. **Editorial Independence** is concerned with the formulation of recommendations not being unduly biased with competing interests.

For further details on each domain see the agree reporting checklist at agreetrust.org

Each of the 23 AGREE II items were rated on a 7-point scale (1 indicating strong disagreement and 7 indicating strong agreement) by 2 reviewers. An overall rating for each of the 6 AGREE II domains was then calculated by summing all the scores of the individual items in a domain from both reviewers and then calculating the total as a percentage of the maximum possible score for that domain, as follows:

$$\frac{\text{Obtained score} - \text{Minimum possible score}}{\text{Maximum possible score} - \text{Minimum possible score}} \times 100$$

Once the assessment of the 6 domains (23 items) is completed, the AGREE II tool suggests two overall assessments. One is a rating of the overall quality of the guideline and the other asks whether the guideline would be recommended for use in practice. However, for the purposes of this guideline, the committee did not assess the guidelines on overall scores as the aim of the review was not to recommend a particular guideline but to obtain an overview of the recommendations in national and international guidelines to inform the committee's recommendations or to cross-refer to specific recommendations if they would add efficiencies to the guideline development process and add value to NICE guidance. The committee acknowledged whilst rigour of development is considered a high priority in evaluating guidelines, they were aware of the difficulty in producing evidence-based guidelines for AI due to the limited research in this area which would inevitably lead to all the guidelines scoring low for rigour of development. They agreed that prioritising specific domains that would help their decision making would be more informative. This would also enable them to distinguish between guidelines because they were aware that most guidelines would follow a similar approach which was consensus due to lack of evidence. The committee agreed, that in the absence of evidence, a high-quality guideline should include a wide range of experience and expert opinion, and for them to consider cross-referring external guideline recommendations, they should be applicable to a UK setting. Therefore, the committee agreed that stakeholder involvement, particularly patient representation, and suitability to UK settings should be given stronger consideration when assessing the quality of the guideline.

Second stage appraisal of guidelines using NICE checklist

For external guidelines assessed to be high quality using the AGREE II instrument, the suitability of specific recommendations for cross-referencing in a NICE guideline was assessed using a NICE checklist. The checklist included discussion points that were developed to provide a structured framework for assessing issues not covered by AGREE II. It was developed by NICE as no checklists were identified that fully covered all relevant. It was informed by the content of other related tools such as ADAPTE and AGREE-REX (see Table 12**)**.

The discussion points were used to assess the recommendations in more detail in terms of applicability and acceptability. For example, they included questions on health inequality considerations, applicability to UK settings, compatibility with cultures and values and consideration of health economics.

The final decision on whether to cross refer to recommendations was made based on the quality of the external guidelines as assessed by the AGREE II instrument and on whether the recommendations within those guidelines satisfied the criteria in the NICE second stage assessment checklist.

Table 13/Box X

NICE second stage assessment discussion points

2.8 Identifying and analysing evidence of cost effectiveness

The committee is required to make decisions based on the best available evidence of both clinical effectiveness and cost effectiveness. Guideline recommendations should be based on the expected costs of the different options in relation to their expected health benefits (that is, their 'cost effectiveness') rather than the total implementation cost. However, the committee will also need to be increasingly confident in the cost effectiveness of a recommendation as the cost of implementation increases. Therefore, the committee may require more robust evidence on the effectiveness and cost effectiveness of any recommendations that are expected to have a substantial impact on resources; any uncertainties must be offset by a compelling argument in favour of the recommendation. The cost impact or savings potential of a recommendation should not be the sole reason for the committee's decision.[4]

Health economic evidence was sought relating to the key clinical issues being addressed in the guideline. Health economists:

- Undertook a systematic review of the published economic literature.

2.8.1 Literature review

The health economists:

- Identified potentially relevant studies for each review question from the health economic search results by reviewing titles and abstracts. Full papers were then obtained.
- Reviewed full papers against prespecified inclusion and exclusion criteria to identify relevant studies (see below for details).
- Critically appraised relevant studies using economic evaluations checklists as specified in the NICE guidelines manual.[4]
- Extracted key information about the studies' methods and results into health economic evidence tables (which can be found in appendices to the relevant evidence reports).
- Generated summaries of the evidence in NICE health economic evidence profile tables (included in the relevant evidence report for each review question) – see below for details.

2.8.1.1 Inclusion and exclusion criteria

Full economic evaluations (studies comparing costs and health consequences of alternative courses of action: cost–utility, cost-effectiveness, cost–benefit and cost–consequences analyses) and comparative costing studies that addressed the review question in the relevant population were considered potentially includable as health economic evidence.

Studies that only reported cost per hospital (not per patient), or only reported average cost effectiveness without disaggregated costs and effects were excluded. Literature reviews, abstracts, posters, letters, editorials, comment articles, unpublished studies and studies not in English were excluded. Studies published before 2007 and studies from non-OECD countries or the USA were also excluded, on the basis that the applicability of such studies to the present UK NHS context is likely to be too low for them to be helpful for decision-making.

Remaining health economic studies were prioritised for inclusion based on their relative applicability to the development of this guideline and the study limitations. For example, if a high quality, directly applicable UK analysis was available, then other less relevant studies may not have been included. However, in this guideline, no economic studies were excluded on the basis that more applicable evidence was available.

For more details about the assessment of applicability and methodological quality see **Table 13** below and the economic evaluation checklist (appendix H of the NICE guidelines manual[4]) and the health economics review protocol, which can be found in each of the evidence reports.

When no relevant health economic studies were found from the economic literature review, relevant UK NHS unit costs related to the compared interventions were presented to the committee to inform the possible economic implications of the recommendations.

2.8.1.2 NICE health economic evidence profiles

NICE health economic evidence profile tables were used to summarise cost and cost-effectiveness estimates for the included health economic studies in each evidence review report. The health economic evidence profile shows an assessment of applicability and methodological quality for each economic study, with footnotes indicating the reasons for the assessment. These assessments were made by the health economist using the economic evaluation checklist from the NICE guidelines manual.[4] It also shows the incremental costs, incremental effects (for example, quality-adjusted life years [QALYs]) and incremental cost-effectiveness ratio (ICER) for the base case analysis in the study, as well as information about the assessment of uncertainty in the analysis. See **Table 13** for more details.

When a non-UK study was included in the profile, the results were converted into pounds sterling using the appropriate purchasing power parity.[8]

Table 14: Content of NICE health economic evidence profile

Item	Description
Study	Surname of first author, date of study publication and country perspective with a reference to full information on the study.
Applicability	An assessment of applicability of the study to this guideline, the current NHS situation and NICE decision-making:[(a)] • 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.
Limitations	An assessment of methodological quality of the study:[(a)] • 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.

Item	Description
	• 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.
Other comments	Information about the design of the study and particular issues that should be considered when interpreting it.
Incremental cost	The mean cost associated with one strategy minus the mean cost of a comparator strategy.
Incremental effects	The mean QALYs (or other selected measure of health outcome) associated with one strategy minus the mean QALYs of a comparator strategy.
Cost effectiveness	Incremental cost-effectiveness ratio (ICER): the incremental cost divided by the incremental effects (usually in £ per QALY gained).
Uncertainty	A summary of the extent of uncertainty about the ICER reflecting the results of deterministic or probabilistic sensitivity analyses, or stochastic analyses of trial data, as appropriate.

(a) Applicability and limitations were assessed using the economic evaluation checklist in appendix H of the NICE guidelines manual[4]

2.8.2 Undertaking new health economic analysis

No health economic modelling was undertaken for this guideline.

2.8.3 Cost-effectiveness criteria

NICE sets out the principles that committees should consider when judging whether an intervention offers good value for money.[4-6] In general, an intervention was considered to be cost effective (given that the estimate was considered plausible) if either of the following criteria applied:

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more clinically effective compared with all the other relevant alternative strategies), or
- 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, the reasons for this decision are discussed explicitly in 'The committee's discussion of the evidence' section of the relevant evidence report, with reference to issues regarding the plausibility of the estimate or to factors set out in NICE methods manuals.[4]

When QALYs or life years gained are not used in the analysis, results are difficult to interpret unless one strategy dominates the others with respect to every relevant health outcome and cost.

2.8.4 In the absence of health economic evidence

When no relevant published health economic studies were found, and a new analysis was not prioritised, the committee made a qualitative judgement about cost effectiveness by considering expected differences in resource use between options and relevant UK NHS unit costs, alongside the results of the review of clinical effectiveness evidence.

The UK NHS costs reported in the guideline are those that were presented to the committee and were correct at the time recommendations were drafted. They may have changed subsequently before the time of publication. However, we have no reason to believe they have changed substantially.

2.9 Developing recommendations

Over the course of the guideline development process, the committee was presented with:

- Summaries of clinical and health economic evidence and quality (as presented in evidence reports A - N).
- Evidence tables of the clinical and health economic evidence reviewed from the literature. All evidence tables can be found in appendices to the relevant evidence reports.
- Forest plots (in appendices to the relevant evidence reports).
- A description of the methods and results of the cost-effectiveness analysis undertaken for the guideline (in a separate economic analysis report).

Decisions on whether a recommendation could be made, and if so in which direction, were made on the basis of the committee's interpretation of the available evidence, taking into account the balance of benefits, harms and costs between different courses of action. This was either done formally in an economic model, or informally. The net clinical benefit over harm (clinical effectiveness) was considered, focusing on the magnitude of the effect (or clinical importance), quality of evidence (including the uncertainty) and amount of evidence available. When this was done informally, the committee took into account the clinical benefits and harms when one intervention was compared with another. The assessment of net clinical benefit was moderated by the importance placed on the outcomes (the committee's values and preferences), and the confidence the committee had in the evidence (evidence quality). Secondly, the committee assessed whether the net clinical benefit justified any differences in costs between the alternative interventions. When the clinical harms were judged by the committee to outweigh any clinical benefits, they considered making a recommendation not to offer an intervention. This was dependant on whether the intervention had any reasonable prospect of providing cost-effective benefits to people using services and whether stopping the intervention was likely to cause harm for people already receiving it.

When clinical and health economic evidence was of poor quality, conflicting or absent, the committee decided on whether a recommendation could be made based on its expert opinion. The considerations for making consensus-based recommendations include the balance between potential harms and benefits, the economic costs compared to the economic benefits, current practices, recommendations made in other relevant guidelines, patient preferences and equality issues. The consensus recommendations were agreed through discussions in the committee. The committee also considered whether the uncertainty was sufficient to justify delaying making a recommendation to await further research, taking into account the potential harm of failing to make a clear recommendation (see section 2.8.1 below).

The committee considered the appropriate 'strength' of each recommendation. This takes into account the quality of the evidence but is conceptually different. Some recommendations are 'strong' in that the committee believes that the vast majority of

healthcare and other professionals and patients would choose a particular intervention if they considered the evidence in the same way that the committee has. This is generally the case if the benefits clearly outweigh the harms for most people and the intervention is likely to be cost effective. However, there is often a closer balance between benefits and harms, and some patients would not choose an intervention whereas others would. This may happen, for example, if some patients are particularly averse to some side effect and others are not. In these circumstances the recommendation is generally weaker, although it may be possible to make stronger recommendations about specific groups of patients.

The committee focused on the following factors in agreeing the wording of the recommendations:

- The actions health professionals need to take.
- The information readers need to know.
- The strength of the recommendation (for example the word 'offer' was used for strong recommendations and 'consider' for weaker recommendations).
- The involvement of patients (and their carers if needed) in decisions on treatment and care.
- Consistency with NICE's standard advice on recommendations about drugs, waiting times and ineffective interventions (see section 9.2 in the NICE guidelines manual[4]).

The main considerations specific to each recommendation are outlined in 'The committee's discussion of the evidence' section within each evidence report.

2.9.1 Research recommendations

When areas were identified for which, good evidence was lacking, the committee considered making recommendations for future research. Decisions about the inclusion of a research recommendation were based on factors such as:

- the importance to patients or the population
- national priorities
- potential impact on the NHS and future NICE guidance
- ethical and technical feasibility.

2.9.2 Validation process

This guidance is subject to a 6-week public consultation and feedback as part of the quality assurance and peer review of the document. All comments received from registered stakeholders are responded to in turn and posted on the NICE website.

2.9.3 Updating the guideline

Following publication, and in accordance with the NICE guidelines manual, NICE will undertake a review of whether the evidence base has progressed significantly to alter the guideline recommendations and warrant an update.

2.10 Terms used in the guideline.

2.10.1 Methodology terms

Term	Definition
Abstract	Summary of a study, which may be published alone or as an introduction to a full scientific paper.
Algorithm (in guidelines)	A flow chart of the clinical decision pathway described in the guideline, where decision points are represented with boxes, linked with arrows.
Allocation concealment	The process used to prevent advance knowledge of group assignment in an RCT. The allocation process should be impervious to any influence by the individual making the allocation, by being administered by someone who is not responsible for recruiting participants.
Applicability	How well the results of a study or NICE evidence review can answer a clinical question or be applied to the population being considered.
Arm (of a clinical study)	Subsection of individuals within a study who receive one particular intervention, for example placebo arm.
Association	Statistical relationship between 2 or more events, characteristics, or other variables. The relationship may or may not be causal.
Base case analysis	In an economic evaluation, this is the main analysis based on the most plausible estimate of each input. In contrast, see Sensitivity analysis.
Baseline	The initial set of measurements at the beginning of a study (after run-in period where applicable), with which subsequent results are compared.
Bayesian analysis	A method of statistics, where a statistic is estimated by combining established information or belief (the 'prior') with new evidence (the 'likelihood') to give a revised estimate (the 'posterior').
Before-and-after study	A study that investigates the effects of an intervention by measuring particular characteristics of a population both before and after taking the intervention, and assessing any change that occurs.
Bias	Influences on a study that can make the results look better or worse than they really are. (Bias can even make it look as if a treatment works when it does not.) Bias can occur by chance, deliberately or as a result of systematic errors in the design and execution of a study. It can also occur at different stages in the research process, for example, during the collection, analysis, interpretation, publication or review of research data. For examples see selection bias, performance bias, information bias, confounding factor, and publication bias.
Blinding	A way to prevent researchers, doctors and patients in a clinical trial from knowing which study group each patient is in so they cannot influence the results. The best way to do this is by sorting patients into study groups randomly. The purpose of 'blinding' or 'masking' is to protect against bias. A single-blinded study is one in which patients do not know which study group they are in (for example whether they are taking the experimental drug or a placebo). A double-blinded study is one in which neither patients nor the researchers and doctors know which study group the patients are in. A triple blind study is one in which

Term	Definition
	neither the patients, clinicians or the people carrying out the statistical analysis know which treatment patients received.
Carer (caregiver)	Someone who looks after family, partners or friends in need of help because they are ill, frail or have a disability.
Case–control study	A study to find out the cause(s) of a disease or condition. This is done by comparing a group of patients who have the disease or condition (cases) with a group of people who do not have it (controls) but who are otherwise as similar as possible (in characteristics thought to be unrelated to the causes of the disease or condition). This means the researcher can look for aspects of their lives that differ to see if they may cause the condition. For example, a group of people with lung cancer might be compared with a group of people the same age that do not have lung cancer. The researcher could compare how long both groups had been exposed to tobacco smoke. Such studies are retrospective because they look back in time from the outcome to the possible causes of a disease or condition.
Case series	Report of a number of cases of a given disease, usually covering the course of the disease and the response to treatment. There is no comparison (control) group of patients.
Clinical efficacy	The extent to which an intervention is active when studied under controlled research conditions.
Clinical effectiveness	How well a specific test or treatment works when used in the 'real world' (for example, when used by a doctor with a patient at home), rather than in a carefully controlled clinical trial. Trials that assess clinical effectiveness are sometimes called management trials. Clinical effectiveness is not the same as efficacy.
Clinician	A healthcare professional who provides patient care. For example, a doctor, nurse, or physiotherapist.
Cochrane Review	The Cochrane Library consists of a regularly updated collection of evidence-based medicine databases including the Cochrane Database of Systematic Reviews (reviews of randomised controlled trials prepared by the Cochrane Collaboration).
Cohort study	A study with 2 or more groups of people – cohorts – with similar characteristics. One group receives a treatment, is exposed to a risk factor, or has a particular symptom and the other group does not. The study follows their progress over time and records what happens. See also observational study.
Comorbidity	A disease or condition that someone has in addition to the health problem being studied or treated.
Comparability	Similarity of the groups in characteristics likely to affect the study results (such as health status or age).
Concordance	This is a recent term whose meaning has changed. It was initially applied to the consultation process in which doctor and patient agree therapeutic decisions that incorporate their respective views, but now includes patient support in medicine taking as well as prescribing communication. Concordance reflects social values but does not address medicine-taking and may not lead to improved adherence.
Confidence interval (CI)	A range of values for an unknown population parameter with a stated 'confidence' (conventionally 95%) that it contains the true value. The interval is calculated from sample data, and generally

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

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

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

Term	Definition
Inclusion criteria (literature review)	Explicit criteria used to decide which studies should be considered as potential sources of evidence.
Incremental analysis	The analysis of additional costs and additional clinical outcomes with different interventions.
Incremental cost	The extra cost linked to using one test or treatment rather than another. Or the additional cost of doing a test or providing a treatment more frequently.
Incremental cost-effectiveness ratio (ICER)	The difference in the mean costs in the population of interest divided by the differences in the mean outcomes in the population of interest for one treatment compared with another.
Incremental net benefit (INB)	The value (usually in monetary terms) of an intervention net of its cost compared with a comparator intervention. The INB can be calculated for a given cost-effectiveness (willingness to pay) threshold. If the threshold is £20,000 per QALY gained, then the INB is calculated as: (£20,000 × QALYs gained) − Incremental cost.
Indirectness	The available evidence is different to the review question being addressed, in terms of PICO (population, intervention, comparison and outcome).
Intention-to-treat analysis (ITT)	An assessment of the people taking part in a clinical trial, based on the group they were initially (and randomly) allocated to. This is regardless of whether or not they dropped out, fully complied with the treatment or switched to an alternative treatment. Intention-to-treat analyses are often used to assess clinical effectiveness because they mirror actual practice: that is, not everyone complies with treatment and the treatment people receive may be changed according to how they respond to it.
Intervention	In medical terms this could be a drug treatment, surgical procedure, diagnostic or psychological therapy. Examples of public health interventions could include action to help someone to be physically active or to eat a healthier diet.
Intraoperative	The period of time during a surgical procedure.
Kappa statistic	A statistical measure of inter-rater agreement that takes into account the agreement occurring by chance.
Length of stay	The total number of days a participant stays in hospital.
Licence	See ‘Product licence’.
Life years gained	Mean average years of life gained per person as a result of the intervention compared with an alternative intervention.
Likelihood ratio	The likelihood ratio combines information about the sensitivity and specificity. It tells you how much a positive or negative result changes the likelihood that a patient would have the disease. The likelihood ratio of a positive test result (LR+) is sensitivity divided by (1 minus specificity).
Long-term care	Residential care in a home that may include skilled nursing care and help with everyday activities. This includes nursing homes and residential homes.
Logistic regression or Logit model	In statistics, logistic regression is a type of analysis used for predicting the outcome of a binary dependent variable based on one or more predictor variables. It can be used to estimate the log of the odds (known as the ‘logit’).

Term	Definition
Loss to follow-up	A patient, or the proportion of patients, actively participating in a clinical trial at the beginning, but whom the researchers were unable to trace or contact by the point of follow-up in the trial
Markov model	A method for estimating long-term costs and effects for recurrent or chronic conditions, based on health states and the probability of transition between them within a given time period (cycle).
Meta-analysis	A method often used in systematic reviews. Results from several studies of the same test or treatment are combined to estimate the overall effect of the treatment.
Multivariate model	A statistical model for analysis of the relationship between 2 or more predictor (independent) variables and the outcome (dependent) variable.
Negative predictive value (NPV)	In screening or diagnostic tests: A measure of the usefulness of a screening or diagnostic test. It is the proportion of those with a negative test result who do not have the disease and can be interpreted as the probability that a negative test result is correct. It is calculated as follows: TN/(TN+FN)
Net monetary benefit (NMB)	The value in monetary terms of an intervention net of its cost. The NMB can be calculated for a given cost-effectiveness threshold. If the threshold is £20,000 per QALY gained, then the NMB for an intervention is calculated as: (£20,000 × mean QALYs) − mean cost. The most preferable option (that is, the most clinically effective option to have an ICER below the threshold selected) will be the treatment with the highest NMB.
Non-randomised intervention study	A quantitative study investigating the effectiveness of an intervention that does not use randomisation to allocate patients (or units) to treatment groups. Non-randomised studies include observational studies, where allocation to groups occurs through usual treatment decisions or people's preferences. Non-randomised studies can also be experimental, where the investigator has some degree of control over the allocation of treatments. Non-randomised intervention studies can use a number of different study designs, and include cohort studies, case–control studies, controlled before-and-after studies, interrupted-time-series studies, and quasi-randomised controlled trials.
Number needed to treat (NNT)	The average number of patients who need to be treated to get a positive outcome. For example, if the NNT is 4, then 4 patients would have to be treated to ensure 1 of them gets better. The closer the NNT is to 1, the better the treatment. For example, if you give a stroke prevention drug to 20 people before 1 stroke is prevented, the number needed to treat is 20. See also number needed to harm, absolute risk reduction.
Observational study	Individuals or groups are observed, or certain factors are measured. No attempt is made to affect the outcome. For example, an observational study of a disease or treatment would allow 'nature' or usual medical care to take its course. Changes or differences in one characteristic (for example, whether or not people received a specific treatment or intervention) are studied without intervening. There is a greater risk of selection bias than in experimental studies.

Term	Definition
Odds ratio	A measure of treatment effectiveness. The odds of an event happening in the treatment group, expressed as a proportion of the odds of it happening in the control group. The 'odds' is the ratio of events to non-events.
Opportunity cost	The loss of other healthcare programmes displaced by investment in or introduction of another intervention. This may be best measured by the health benefits that could have been achieved had the money been spent on the next best alternative healthcare intervention.
Outcome	The impact that a test, treatment, policy, programme, or other intervention has on a person, group, or population. Outcomes from interventions to improve the public's health could include changes in knowledge and behaviour related to health, societal changes (for example, a reduction in crime rates) and a change in people's health and wellbeing or health status. In clinical terms, outcomes could include the number of patients who fully recover from an illness or the number of hospital admissions, and an improvement or deterioration in someone's health, functional ability, symptoms, or situation. Researchers should decide what outcomes to measure before a study begins.
P value	The p value is a statistical measure that indicates whether or not an effect is statistically significant. For example, if a study comparing 2 treatments found that one seems more effective than the other, the p value is the probability of obtaining these, or more extreme results by chance. By convention, if the p value is below 0.05 (that is, there is less than a 5% probability that the results occurred by chance) it is considered that there probably is a real difference between treatments. If the p value is 0.001 or less (less than a 1% probability that the results occurred by chance), the result is seen as highly significant. If the p value shows that there is likely to be a difference between treatments, the confidence interval describes how big the difference in effect might be.
Perioperative	The period from admission through surgery until discharge, encompassing the preoperative and postoperative periods.
Placebo	A fake (or dummy) treatment given to participants in the control group of a clinical trial. It is indistinguishable from the actual treatment (which is given to participants in the experimental group). The aim is to determine what effect the experimental treatment has had – over and above any placebo effect caused because someone has received (or thinks they have received) care or attention.
Posterior distribution	In Bayesian statistics this is the probability distribution for a statistic based after combining established information or belief (the prior) with new evidence (the likelihood).
Positive predictive value (PPV)	In screening or diagnostic tests: A measure of the usefulness of a screening or diagnostic test. It is the proportion of those with a positive test result who have the disease and can be interpreted as the probability that a positive test result is correct. It is calculated as follows: TP/(TP+FP)
Postoperative	Pertaining to the period after patients leave the operating theatre, following surgery.

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

Term	Definition
	groups are followed up to see how effective the experimental treatment was. Outcomes are measured at specific times and any difference in response between the groups is assessed statistically. This method is also used to reduce bias.
RCT	See 'Randomised controlled trial'.
Receiver operated characteristic (ROC) curve	A graphical method of assessing the accuracy of a diagnostic test. Sensitivity is plotted against 1 minus specificity. A perfect test will have a positive, vertical linear slope starting at the origin. A good test will be somewhere close to this ideal.
Reference standard	The test that is considered to be the best available method to establish the presence or absence of the outcome – this may not be the one that is routinely used in practice.
Reporting bias	See 'Publication bias'.
Resource implication	The likely impact in terms of finance, workforce, or other NHS resources.
Retrospective study	A research study that focuses on the past and present. The study examines past exposure to suspected risk factors for the disease or condition. Unlike prospective studies, it does not cover events that occur after the study group is selected.
Review question	In guideline development, this term refers to the questions about treatment and care that are formulated to guide the development of evidence-based recommendations.
Risk ratio (RR)	The ratio of the risk of disease or death among those exposed to certain conditions compared with the risk for those who are not exposed to the same conditions (for example, the risk of people who smoke getting lung cancer compared with the risk for people who do not smoke). If both groups face the same level of risk, the risk ratio is 1. If the first group had a risk ratio of 2, subjects in that group would be twice as likely to have the event happen. A risk ratio of less than 1 means the outcome is less likely in the first group. The risk ratio is sometimes referred to as relative risk.
Secondary outcome	An outcome used to evaluate additional effects of the intervention deemed a priori as being less important than the primary outcomes.
Selection bias	Selection bias occurs if: a) The characteristics of the people selected for a study differ from the wider population from which they have been drawn, or b) There are differences between groups of participants in a study in terms of how likely they are to get better.
Sensitivity	How well a test detects the thing it is testing for. If a diagnostic test for a disease has high sensitivity, it is likely to pick up all cases of the disease in people who have it (that is, give a 'true positive' result). But if a test is too sensitive it will sometimes also give a positive result in people who don't have the disease (that is, give a 'false positive'). For example, if a test were developed to detect if a woman is 6 months pregnant, a very sensitive test would detect everyone who was 6 months pregnant but would probably also include those who are 5 and 7 months pregnant. If the same test were more specific (sometimes referred to as having higher specificity), it would detect only those who are 6 months pregnant, and someone who was 5 months pregnant would

Term	Definition
	get a negative result (a 'true negative'). But it would probably also miss some people who were 6 months pregnant (that is, give a 'false negative'). Breast screening is a 'real-life' example. The number of women who are recalled for a second breast screening test is relatively high because the test is very sensitive. If it were made more specific, people who don't have the disease would be less likely to be called back for a second test but more women who have the disease would be missed.
Sensitivity analysis	A means of representing uncertainty in the results of economic evaluations. Uncertainty may arise from missing data, imprecise estimates, or methodological controversy. Sensitivity analysis also allows for exploring the generalisability of results to other settings. The analysis is repeated using different assumptions to examine the effect on the results. One-way simple sensitivity analysis (univariate analysis): each parameter is varied individually in order to isolate the consequences of each parameter on the results of the study. Multi-way simple sensitivity analysis (scenario analysis): 2 or more parameters are varied at the same time and the overall effect on the results is evaluated. Threshold sensitivity analysis: the critical value of parameters above or below which the conclusions of the study will change are identified. Probabilistic sensitivity analysis: probability distributions are assigned to the uncertain parameters and are incorporated into evaluation models based on decision analytical techniques (for example, Monte Carlo simulation).
Significance (statistical)	A result is deemed statistically significant if the probability of the result occurring by chance is less than 1 in 20 ($p<0.05$).
Specificity	The proportion of true negatives that are correctly identified as such. For example, in diagnostic testing the specificity is the proportion of non-cases correctly diagnosed as non-cases. See related term 'Sensitivity'. In terms of literature searching a highly specific search is generally narrow and aimed at picking up the key papers in a field and avoiding a wide range of papers.
Stakeholder	An organisation with an interest in a topic that NICE is developing a guideline or piece of public health guidance on. Organisations that register as stakeholders can comment on the draft scope and the draft guidance. Stakeholders may be: • manufacturers of drugs or equipment • national patient and carer organisations • NHS organisations • organisations representing healthcare professionals.
State transition model	See Markov model.
Stratification	When a different estimate effect is thought to underlie two or more groups based on the PICO characteristics. The groups are therefore kept separate from the outset and are not combined in a meta-analysis, for example, children and adults. Specified a priori in the protocol.

Term	Definition
Sub-groups	Planned statistical investigations if heterogeneity is found in the meta-analysis. Specified a priori in the protocol.
Systematic review	A review in which evidence from scientific studies has been identified, appraised, and synthesised in a methodical way according to predetermined criteria. It may include a meta-analysis.
Time horizon	The time span over which costs and health outcomes are considered in a decision analysis or economic evaluation.
Transition probability	In a state transition model (Markov model), this is the probability of moving from one health state to another over a specific period of time.
Treatment allocation	Assigning a participant to a particular arm of a trial.
Univariate	Analysis which separately explores each variable in a data set.
Utility	In health economics, a 'utility' is the measure of the preference or value that an individual or society places upon a particular health state. It is generally a number between 0 (representing death) and 1 (perfect health). The most widely used measure of benefit in cost–utility analysis is the quality-adjusted life year, but other measures include disability-adjusted life years (DALYs) and healthy year equivalents (HYEs).

2.10.2 Clinical terms

Term	Definition
Adrenal crisis	A critical condition that arises in individuals with adrenal insufficiency, marked by a profound shortage of cortisol, a hormone made by the adrenal glands. Signs may encompass severe weakness, dehydration, hypotension, and cognitive disorientation. Urgent medical intervention with glucocorticoids such as hydrocortisone and fluid is imperative.
Adrenocorticotropic hormone (ACTH) stimulation test	A medical diagnostic test used to assess the function of the adrenal glands. The test evaluates how well the adrenal glands respond to adrenocorticotropic hormone (ACTH); a hormone produced by the pituitary gland that stimulates the adrenal glands to release cortisol.
Ambiguous genitalia	A condition in which an individual's external genitalia do not appear distinctly male or female at birth.
Autoimmune Addison’s Disease	A rare autoimmune disorder in which the body's immune system mistakenly attacks and damages the adrenal glands, leading to insufficient production of essential hormones, such as cortisol and aldosterone. This results in symptoms like fatigue, weakness, weight loss, and low blood pressure. Treatment typically involves hormone replacement therapy to manage the hormonal deficiencies.
BD	Related to medical prescriptions, ‘bis in die’ is a Latin term meaning ‘twice a day’.
Chronic Obstructive Pulmonary Disease	A progressive lung disease characterized by airflow limitation and difficulty breathing. It includes conditions such as chronic bronchitis and emphysema and is often caused by smoking or exposure to harmful substances.

Term	Definition
	Symptoms include coughing, wheezing, shortness of breath, and increased susceptibility to respiratory infections. Management involves lifestyle changes, medications, and pulmonary rehabilitation.
Congenital adrenal hyperplasia	A group of genetic disorders causing enzyme defects affecting the adrenal glands' hormone production from birth. It leads to abnormal hormone levels, including cortisol, aldosterone, testosterone and 17 hydroxyprogesterone , resulting in various symptoms depending on the specific type of CAH. Treatment aims to manage hormone imbalances and related complications.
Cortisol	A glucocorticoid (steroid) hormone produced by the adrenal glands, which are located on top of each kidney. It plays a crucial role in various bodily functions and is often referred to as the "stress hormone" because its levels rise in response to physiological stress such as infections and and in response to trauma or surgery.
Crohn’s disease	A chronic inflammatory bowel disease causing inflammation and ulcers in the gastrointestinal tract, leading to symptoms like abdominal pain, diarrhoea, and weight loss.
Cushingoid features	Physical characteristics or symptoms resembling those seen in Cushing's syndrome, which is caused by excessive levels of cortisol in the body. These features include weight gain, rounded face, buffalo hump (fat accumulation between the shoulders), and thinning of the skin. They can be caused by prolonged use of treatment doses of glucocorticoid medications or other conditions leading to increased cortisol production.
Demyelinating spastic paraparesis	A neurological condition characterized by damage to the myelin sheath, the protective covering of nerve fibres, leading to muscle weakness, stiffness, and spasticity (increased muscle tone). It is often associated with conditions like multiple sclerosis. Treatment aims to manage symptoms and slow disease progression.
Electrochemiluminescence method (ECLIA)	A method used in clinical laboratories for the quantification of various substances in biological samples, such as blood or serum. It is commonly employed for measuring hormones, tumour markers, infectious disease markers, and other analytes of interest. The method combines principles of electrochemistry and chemiluminescence.
Haemochromatosis	A genetic disorder characterized by excessive absorption and accumulation of iron in the body. The condition leads to the gradual buildup of iron in various organs, particularly the liver, heart, and pancreas and a late diagnosis and treatment can result in serious health problems.

Term	Definition
HbA1c	A blood test that measures the average blood sugar level over the past 3 months. It is often used to assess diabetes
Histiocytosis X	A group of rare disorders characterized by the overproduction and accumulation of white blood cells called histiocytes. These cells normally play a role in the immune system and are responsible for engulfing and digesting foreign substances, as well as helping to regulate the immune response.
Hypoglycaemia	A condition in which blood sugar (glucose) levels drop too low. Symptoms include sweating, shakiness, anxiety, blurred vision, headache, and confusion. It can result from medications including medication used to manage diabetes, certain or diseases. Treatment involves consuming or administering glucose.
Hyponatraemia	A medical condition characterized by an abnormally low concentration of sodium in the blood- less than 135mmol/L. When sodium levels in the blood become too low, it can lead to various symptoms and, in severe cases, pose serious health risks.
Hypotension	A medical condition characterized by abnormally low blood pressure, which may lead to dizziness, fainting, or inadequate blood flow to vital organs.
Hypotensive crisis	A hypotensive crisis, also known as severe hypotension or hypotensive emergency, refers to a sudden and severe drop in blood pressure that requires immediate medical attention.
Hypothalamic Pituitary Adrenal (HPA) axis	A vital neuroendocrine system in the body responsible for regulating the physiological stress response and maintaining hormonal balance, involving interactions between the hypothalamus, pituitary gland, and adrenal glands.
Hypothalamus	A small region located at the base of the brain, responsible for regulating various bodily functions and maintaining homeostasis by controlling the release of hormones and influencing behaviours like hunger, thirst, body temperature, and sleep.
Hypovolemia	A medical condition characterized by an abnormally low volume of blood circulating in the body. It can result from severe dehydration, bleeding, or fluid loss and may lead to symptoms such as dizziness, rapid heart rate, and low blood pressure. Prompt medical attention is necessary to address the underlying cause and restore fluid balance.
Insulin tolerance test	A medical diagnostic test used to check how well the hypothalamus, pituitary gland and therefore also the adrenals glands respond to a low blood sugar. A usual response is to have a rise in ACTH and hence a rise is cortisol.
Lupus	An autoimmune disease where the immune system attacks healthy tissues and organs, leading to inflammation and damage. Symptoms can vary widely

Term	Definition
	and may affect the skin, joints, kidneys, and other organs. Treatment involves medications to manage inflammation and immune response.
Lymphocytic hypophysitis	A rare disorder that involves inflammation of the pituitary gland due to infiltration of lymphocytes (a type of white blood cell) into the gland.
Multiple Sclerosis	A chronic autoimmune disease of the central nervous system, where the immune system attacks the protective covering of nerve fibres, leading to various neurological symptoms such as muscle weakness, coordination difficulties, and cognitive impairments.
Myasthenia Gravis	An autoimmune neuromuscular disorder causing muscle weakness and fatigue due to the immune system attacking the neuromuscular junction. Managed with medications and sometimes surgery.
Polymyalgia	A medical condition characterized by muscle pain and stiffness, typically affecting the shoulders, neck, and hips. It is often associated with another condition called giant cell arteritis. Treatment involves anti-inflammatory medications to relieve symptoms and manage underlying causes.
Primary Adrenal insufficiency	A medical condition where the adrenal glands do not function characterised by insufficient production of adrenal hormones, glucocorticoids such as cortisol, and mineralocorticoids such as aldosterone. This can lead to symptoms like fatigue, weakness, low blood pressure, and electrolyte imbalances which can be life threatening. Treatment involves glucocorticoid and mineralocorticoid replacement and investigating and managing the underlying cause. addressing the underlying cause.
Prolonged jaundice	Jaundice is a yellowing of the skin and eyes due to an accumulation of bilirubin, a yellow pigment produced during the breakdown of red blood cells. While jaundice is common in newborns and often resolves within the first two weeks of life, on its own. If jaundice persists beyond this period, it is considered prolonged jaundice.
Rheumatoid Arthritis	A chronic autoimmune disease that primarily affects the joints, causing inflammation, pain, stiffness, and joint deformities. It can also affect other body systems. Treatment aims to reduce inflammation, manage symptoms, and slow disease progression through medications and lifestyle modifications.
Sarcoidosis	A rare inflammatory disease that can affect multiple organs in the body, most commonly the lungs and lymph nodes. The condition is characterized by the formation of small clumps of inflammatory cells, known as granulomas, in various tissues. These granulomas can interfere with the normal function of affected organs.
Secondary adrenal insufficiency	A medical condition resulting from pituitary gland dysfunction causing insufficient production of ACTH hormone. This then causes decreased production of

Term	Definition
	cortisol from the adrenal glands. It can be picked up on testing, or by similar symptoms to primary adrenal insufficiency. Treatment involves glucocorticoid replacement and investigating and managing the underlying cause.
Steroid-Sensitive Nephrotic Syndrome	A kidney disorder primarily affecting children, characterized by excessive protein in the urine, low blood albumin levels, swelling, and high blood lipid levels. It responds well to treatment with corticosteroid medications.
Synacthen test	A medical diagnostic test used to assess the function of the adrenal glands. (See above ACTH stimulation test).
Tertiary adrenal insufficiency	A medical condition resulting from dysfunction of the hypothalamus causing insufficient production of CRH hormone. This then causes decreasing production of ACTH and cortisol form the adrenal glands. It can be picked up by testing for symptoms similar to primary adrenal insufficiency. Treatment involves glucocorticoid replacement and investigating and managing the underlying cause.
Thrombocytopenic purpura (TTP)	A rare blood disorder characterized by low platelet levels (thrombocytopenia) and the formation of small blood clots throughout the body. It can lead to purpura (purple or red spots on the skin caused by bleeding under the skin) and can affect various organs, potentially causing serious complications. TTP requires immediate medical attention and treatment to prevent life-threatening outcomes.
TDS	Related to medical prescriptions, ‘ter in die’ is a Latin term meaning ‘three times a day’.

2.10.3 Acronyms

Acronym	Meaning
AD	Addison’s Disease
ACTH	Adrenocorticotropic hormone
AI	Adrenal Insufficiency
BD	Related to medical prescriptions, ‘bis in die’ is a Latin term meaning ‘twice a day’.
CAH	Congenital Adrenal Hyperplasia
CARES	Congenital Adrenal Hyperplasia Research Education and Support
CMV	Cytomegalovirus
COPD	Chronic Obstructive Pulmonary Disease
DSD	Differences in Sex Development

Acronym	Meaning
ECLIA	Electrochemiluminescence method
FMSF	Family Management Style Framework
GC-MS	Gas Chromatography Mass Spectrometry
HbA1c	Haemoglobin A1c
HCP	Health Care Professional
HDT	High-dose Synacthen Test
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
HPA Axis	Hypothalamic Pituitary Adrenal Axis
IM	Intramuscular
ITT	Insulin Tolerance Test
IV	Intravenous
LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
LDSST	Low Dose Short Synacthen Test
LDST	Low Dose Synacthen Test
PAI	Primary Adrenal Insufficiency
PVM	Photovoice method
SSST	Standard Short Synacthen Test
SST	Short Synacthen Test
TB	Tuberculosis
TDS	Related to medical prescriptions, ‘ter in die’ is a Latin term meaning ‘three times a day’.
TTP	Thrombocytopenic purpura

References

1. GRADE Working Group. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group website. 2011. Available from: http://www.gradeworkinggroup.org/ Last accessed: 18/12/2023.

2. Maruish M, Kosinski M, Bjorner J, Gandek B, Turner-Bowker D, Ware J. User’s manual for the SF-36v2 Health Survey. 2011.

3. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. Journal of Clinical Epidemiology. 2016; 75:40-46

4. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. London. National Institute for Health and Care

Excellence, 2014. Available from: https://www.nice.org.uk/process/pmg20/chapter/introduction

5. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal 2013. London. National Institute for Health and Clinical Excellence, 2013. Available from: http://publications.nice.org.uk/pmg9

6. National Institute for Health and Clinical Excellence. Social value judgements: principles for the development of NICE guidance. London. National Institute for Health and Clinical Excellence, 2008. Available from: https://www.nice.org.uk/media/default/about/what-we-do/research-and-development/social-value-judgements-principles-for-the-development-of-nice-guidance.pdf

7. Novielli N, Cooper NJ, Abrams KR, Sutton AJ. How is evidence on test performance synthesized for economic decision models of diagnostic tests? A systematic appraisal of Health Technology Assessments in the UK since 1997. Value in Health. 2010; 13(8):952-957

8. Organisation for Economic Co-operation and Development (OECD). Purchasing power parities (PPP). 2012. Available from: http://www.oecd.org/std/ppp Last accessed: 18/12/2023.

9. Review Manager (RevMan) [Computer program]. Version 5. Copenhagen. The Nordic Cochrane Centre, The Cochrane Collaboration, 2015. Available from: http://tech.cochrane.org/Revman

10. WinBUGS [Computer programme] version 1.4. Cambridge. MRC Biostatistics Unit University of Cambridge, 2015. Available from: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/