National Institute for Health and Care Excellence

Final

Hyperparathyroidism (primary): diagnosis, assessment and initial management

NICE guideline: methods

NICE guideline NG132

Methods

May 2019

Final

Developed by the National Guideline Centre, hosted by the Royal College of Physicians

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, where appropriate, their carer or guardian.

Local commissioners and 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 All NICE guidance is subject to regular review and may be updated or withdrawn.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: 978-1-4731-3415-7

Contents

1 Development of the guideline

1.1 What is a NICE guideline?

NICE guidelines are recommendations for the care of individuals in specific clinical conditions or circumstances within the NHS – from prevention and self-care through primary and secondary care to more specialised services. These may also include elements of social care or public health measures. We base our guidelines on the best available research evidence, with the aim of improving the quality of healthcare. We use predetermined and systematic methods to identify and evaluate the evidence relating to specific review questions.

NICE guidelines can:

- provide recommendations for the treatment and care of people by health professionals
- be used to develop standards to assess the clinical practice of individual health professionals
- be used in the education and training of health professionals
- help patients to make informed decisions
- improve communication between patient and health professional.

While guidelines assist the practice of healthcare professionals, they do not replace their knowledge and skills.

We produce our guidelines using the following steps:

- A guideline topic is referred to NICE from NHS England.
- Stakeholders register an interest in the guideline and are consulted throughout the development process.
- The scope is prepared by the National Guideline Centre (NGC).
- The NGC establishes a guideline committee.
- A draft guideline is produced after the group assesses the available evidence and makes recommendations.
- There is a consultation on the draft guideline.
- The final guideline is produced.

The guideline is made up of a collection of documents including this Methods report and a number of evidence reports covering each of the review questions included in the guideline. These can all be downloaded from NICE at www.nice.org.uk.

NICE also publishes a summary of the recommendation in this guideline, known as ‘the NICE guideline’.

NICE Pathways brings together all connected NICE guidance.

1.2 Remit

NICE received the remit for this guideline from NHS England. NICE commissioned the NGC to produce the guideline.

The remit for this guideline is: Hyperparathyroidism (primary): diagnosis, assessment and initial management.

To prepare a clinical guideline on the diagnosis, assessment and initial management of primary hyperparathyroidism.

1.3 Who developed this guideline?

A multidisciplinary guideline committee comprising health professionals and researchers as well as lay members developed this guideline (see the list of guideline committee members and the acknowledgements).

The National Institute for Health and Care Excellence (NICE) funds the National Guideline Centre (NGC) and thus supported the development of this guideline. The committee was convened by the NGC and chaired by Jonathan Mant in accordance with guidance from NICE.

The group met approximately every 6 weeks during the development of the guideline. At the start of the guideline development process all committee members declared interests including consultancies, fee-paid work, shareholdings, fellowships and support from the healthcare industry. At all subsequent committee meetings, members declared arising conflicts of interest.

Members were either required to withdraw completely or for part of the discussion if their declared interest made it appropriate. The details of declared interests and the actions taken are shown in the declaration of interest register for this guideline published on the NICE website.

Staff from the NGC provided methodological support and guidance for the development process. The team working on the guideline included a project manager, systematic reviewers (research fellows), health economists and information specialists. They undertook systematic searches of the literature, appraised the evidence, conducted meta-analysis and cost-effectiveness analysis where appropriate and drafted the guideline in collaboration with the committee.

1.3.1 What this guideline covers

The guideline will cover adults (18 years and over) with suspected or confirmed primary hyperparathyroidism.

Specific consideration was given to women who are pregnant.

Key areas that are covered

- Identifying and diagnosing symptomatic and asymptomatic primary hyperparathyroidism.
- Indications for surgery (parathyroidectomy).
- Investigations before and during parathyroid surgery.
- Surgical management.
- Pharmacological management.
- Monitoring.
- Managing primary hyperparathyroidism during pregnancy.
- Providing information to people with primary hyperparathyroidism.

For further details please refer to the scope for this guideline (published on the NICE website) and the review questions in section 2.1.

1.3.2 What this guideline does not cover

Areas that are not covered

- Multiple endocrine neoplasia.
- Familial hyperparathyroidism.
- Parathyroid carcinoma.
- Secondary hyperparathyroidism.
- Tertiary hyperparathyroidism.
- Obstetric complications and neonatal care.
- General population screening for primary hyperparathyroidism.
- Management of long term complications of primary hyperparathyroidism.
- Management of vitamin D deficiency.

1.3.3 Relationships between the guideline and other NICE guidance

Related NICE technology appraisals:

- Bisphosphonates for treating osteoporosis. NICE technology appraisal guidance TA464 (2017)
- Alendronate, etidronate, risedronate, raloxifene and strontium ranelate for the primary prevention of osteoporotic fragility fractures in postmenopausal women. NICE technology appraisal guidance TA160 (2008)
- Cinacalcet for the treatment of secondary hyperparathyroidism in patients with end-stage renal disease. NICE technology appraisal guidance TA117 (2007)

Related NICE interventional procedures guidance:

- Minimally invasive video‑assisted parathyroidectomy. NICE interventional procedure guidance IPG501 (2014)
- Thoracoscopic excision of mediastinal parathyroid tumours. NICE interventional procedure guidance IPG247 (2007)

Related NICE guidelines:

- Renal and ureteric stones: assessment and management. NICE guideline NG118 (2019)
- Multimorbidity: clinical assessment and management. NICE guideline NG56 (2016)
- Menopause: diagnosis and management. NICE guideline NG23 (2015)
- Medicines optimisation. NICE guideline NG5 (2015)
- Vitamin D: increasing supplement use in at-risk groups. NICE public health guideline PH56 (2014)
- Osteoporosis: assessing the risk of fragility fracture. NICE guideline CG146 (2012)
- Patient experience in adult NHS services. NICE guideline CG138 (2012)
- Hypertension in pregnancy: diagnosis and management. NICE guideline CG107 (2010)
- Metastatic malignant disease of unknown primary origin in adults: diagnosis and management. NICE guideline CG104 (2010)
- Depression in adults: recognition and management. NICE guideline CG90 (2009)
- Medicines adherence. NICE guideline CG76 (2009)

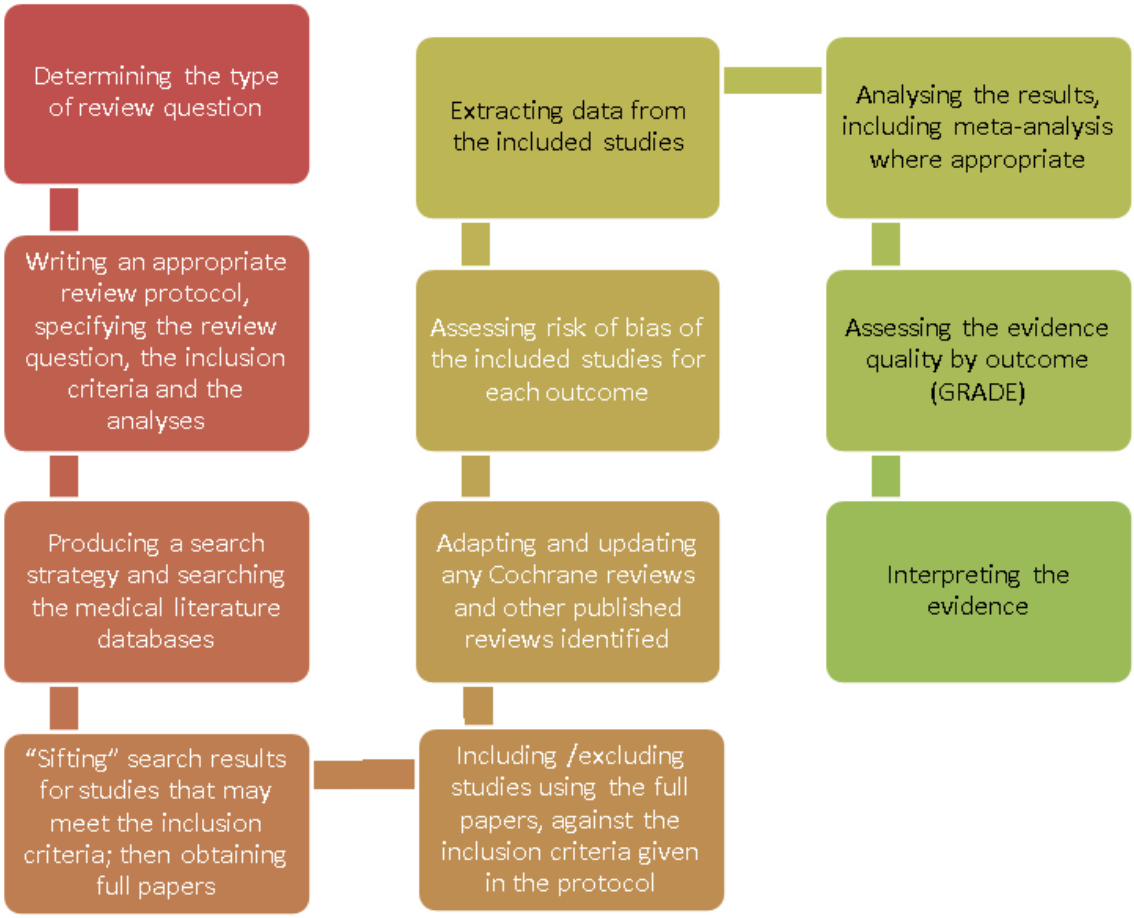
- Chronic fatigue syndrome/myalgic encephalomyelitis (or encephalopathy): diagnosis and management. NICE guideline CG53 (2007)

2 Methods

This report sets out in detail the methods used to review the evidence and to develop the recommendations that are presented in each of the evidence reviews for this guideline. This guidance was developed in accordance with the methods outlined in the NICE guidelines manual, 2014 version.[2] We will be using the NICE guidelines manual, 2018 version for validation.[5]

Sections 2.1 to 2.3 describe the process used to identify and review clinical evidence (summarised in Figure 1), sections 2.2 and 2.4 describe the process used to identify and review the health economic evidence, and section 2.5 describes the process used to develop recommendations.

Figure 1: Step-by-step process of review of evidence in the guideline



2.1 Developing the review questions and outcomes

Review questions were developed using a PICO framework (population, intervention, comparison and outcome) for intervention reviews; using a framework of population, index tests, reference standard and target condition for reviews of diagnostic test accuracy; and using a framework of population, setting and context for qualitative reviews.

This use of a framework guided the literature searching process, critical appraisal and synthesis of evidence, and facilitated the development of recommendations by the guideline committee. The review questions were drafted by the NGC technical team and refined and validated by the committee. The questions were based on the key clinical areas identified in the scope.

A total of 8 key clinical areas were identified.

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)	Diagnostic	What are the indications for diagnostic testing for primary hyperparathyroidism?	Target condition: primary hyperparathyroidism • Specificity • Sensitivity • Positive and/or negative predictive value • ROC curve or area under curve
1.1 (b)	Prognostic	In adults with fragility fracture, renal stones, and/or renal tract calcification what is the incidence of primary hyperparathyroidism?	Diagnosis of PHPT
1.2	Test and treat Diagnostic	Which biochemical test or combination of tests should be used for diagnosing primary hyperparathyroidism (for example, levels of parathyroid hormone, blood calcium and phosphate, alone or in combination)?	Target condition: primary hyperparathyroidism Outcomes for test and treat review: • Mortality (dichotomous outcome) • Quality of life (continuous outcome) • Number of people receiving treatment, i.e., including people who may not have needed it such as those with false positive results (dichotomous outcome) • Repeat testing/additional testing (dichotomous outcome) • Adverse events related to test (as reported in the papers) • Adverse events related to treatment (as reported in the papers) • Preservation of end organ function (bone mineral density, fractures, renal stones and renal function) (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • Cardiovascular events (dichotomous outcome)

Evidence report	Type of review	Review questions	Outcomes
			• Cancer incidence (dichotomous outcome) Outcomes for diagnostic accuracy review: • Specificity • Sensitivity • Positive and/or negative predictive value • ROC curve or area under curve
2.1	Intervention	What is the clinical and cost effectiveness of surgery (parathyroidectomy) in people with primary hyperparathyroidism?	**Critical outcomes:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Preservation of end organ function (bone mineral density, fractures, renal stones and renal function) (dichotomous outcome) **Important outcomes:** • Deterioration in renal function (dichotomous) • Persistent hypercalcaemia (dichotomous outcome) • Cardiovascular events (dichotomous outcome) • Adverse events (to include voice change, hypoparathyroidism) (dichotomous outcome) • Cancer incidence (dichotomous outcome)
2.2	Intervention	What are the indications for surgery (parathyroidectomy) in people with primary hyperparathyroidism?	**Critical outcomes:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Preservation of end organ function (bone mineral density, fractures, renal stones and renal function) (dichotomous outcome) **Important outcomes:** • Deterioration in renal function (dichotomous) • Persistent hypercalcaemia

Evidence report	Type of review	Review questions	Outcomes
			(dichotomous outcome) • Cardiovascular events (dichotomous outcome) • Adverse events (to include voice change, hypoparathyroidism; dichotomous outcome) • Cancer incidence (dichotomous outcome)
3.1	Test and treat Diagnostic	What is the clinical and cost effectiveness of using non-invasive imaging techniques (for example parathyroid ultrasound, sestamibi scanning, CT and MRI scanning) prior to surgery?	**For test-and-treat review:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Success (cure)/failure (dichotomous outcome) • Adverse events (dichotomous outcome) • BMD of the distal radius or the lumbar spine (continuous outcome) • Deterioration in renal function (dichotomous outcome) • Fractures (vertebral or long bone) (dichotomous outcome) • Length of hospital stay (continuous outcome) • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia • Reoperation (dichotomous outcome) • Unnecessary neck exploration (dichotomous outcome) **For diagnostic accuracy review:** Target condition (for localisation studies): correct localisation of adenoma (correctly localises the region/quadrant from which an abnormal gland is removed [rather than just correctly identifies hyperactive tissue anywhere, or correctly lateralises the hyperactive gland]). Target condition (for intra-

Evidence report	Type of review	Review questions	Outcomes
			operative tests): correct prediction of removal of all abnormal tissue. Outcomes of interest: Specificity Sensitivity
3.2	Test and treat Diagnostic	What is the clinical and cost effectiveness of using invasive imaging techniques (for example parathyroid venous sampling) prior to surgery?	**For test-and-treat review:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Success (cure)/failure (dichotomous outcome) • Adverse events (dichotomous outcome) • BMD of the distal radius or the lumbar spine (continuous outcome) • Deterioration in renal function (dichotomous outcome) • Fractures (vertebral or long bone) (dichotomous outcome) • Length of hospital stay • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • Reoperation (dichotomous outcome) • Unnecessary neck exploration (dichotomous outcome) **For diagnostic accuracy review:** Target condition (for localisation studies): correct localisation of adenoma (correctly localises the region/quadrant from which an abnormal gland is removed [rather than just correctly identifies hyperactive tissue anywhere, or correctly lateralises the hyperactive gland]). Outcomes of interest: Specificity Sensitivity
3.3	Test and treat Diagnostic	What is the clinical and cost effectiveness of using	**For test-and-treat review:** • HRQOL (continuous

Evidence report	Type of review	Review questions	Outcomes
		intraoperative parathyroid hormone assays, methylene blue and intraoperative frozen sections?	outcome) • Mortality (dichotomous outcome) • Success (cure)/failure (dichotomous outcome) • Adverse events (dichotomous outcome) • BMD of the distal radius or the lumbar spine (continuous outcome) • Deterioration in renal function (dichotomous outcome) • Fractures (vertebral or long bone) (dichotomous outcome) • Length of hospital stay (continuous outcome) • Occurrence of kidney stones • Persistent hypercalcaemia (dichotomous outcome) • Reoperation (dichotomous outcome) • Unnecessary neck exploration (dichotomous outcome) **For diagnostic accuracy review:** Target condition (for intra-operative tests): correct prediction of removal of all abnormal tissue. Outcomes of interest: Specificity Sensitivity
4.1	Intervention	What is the clinical and cost effectiveness of different types of surgical intervention, for example 4-gland exploration, compared with minimally invasive techniques?	**Critical outcomes** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Success (cure)/failure (dichotomous outcome) **Important outcomes** • Adverse events (bleeding (return to theatre), severe hypocalcaemia (as defined in the study), hypercalcaemia, laryngeal nerve injury, vocal cord paralysis/laryngeal nerve

Evidence report	Type of review	Review questions	Outcomes
			injury, haematoma, infection) (dichotomous outcome) • BMD of the distal radius or the lumbar spine (continuous) • Deterioration in renal function (dichotomous – study may also report renal replacement) • Fractures (vertebral or long bone) (dichotomous outcome) • Length of hospital stay (continuous outcome) • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • Reoperation (dichotomous outcome) • Unnecessary neck exploration (dichotomous outcome)
4.2	Intervention	What are the management options for people in whom primary parathyroid surgery has failed?	**Critical outcomes:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) • Preservation of end organ function (bone mineral density, fractures, renal stones and renal function) (dichotomous) **Important outcomes:** • Deterioration in renal function (dichotomous) • Persistent hypercalcaemia (dichotomous outcome) • Cardiovascular events (dichotomous outcome) • Adverse events (dichotomous outcome) • Cancer incidence (dichotomous outcome)
5.1	Intervention	What is the clinical and cost effectiveness of calcimimetics in people with primary hyperparathyroidism?	**Critical outcomes:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) **Important outcomes:** • Deterioration in renal

Evidence report	Type of review	Review questions	Outcomes
			function (dichotomous - study may also report renal replacement) • Fractures (vertebral or long bone) (dichotomous outcome) • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • BMD (continuous) of the distal radius or the lumbar spine • Cardiovascular events (dichotomous outcome) • Adverse events (to include discontinuation due to side effects; dichotomous outcome) • Cancer incidence (dichotomous outcome)
5.2	Intervention	What is the clinical and cost effectiveness of bisphosphonates in people with primary hyperparathyroidism?	**Critical outcomes:** • HRQOL (continuous outcome) • Mortality (dichotomous outcome) **Important outcomes:** • Deterioration in renal function (dichotomous – study may also report renal replacement) • Fractures (vertebral or long bone) (dichotomous outcome) • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • BMD (continuous) of the distal radius or the lumbar spine • Cardiovascular events (dichotomous outcome) • Adverse events (to include discontinuation due to side effects; dichotomous outcome) • Cancer incidence (dichotomous outcome)
6.1	Intervention	What is the optimum type and frequency of monitoring for people with primary hyperparathyroidism	**Critical outcomes:** • HRQOL (continuous outcome)

Evidence report	Type of review	Review questions	Outcomes
		(for example, pre-operative, postoperative, non-surgical)?	• Mortality (dichotomous outcome) **Important outcomes:** • Deterioration in renal function (continuous outcome) • Fractures (vertebral or long bone) (dichotomous outcome) • Occurrence of kidney stones (dichotomous outcome) • Persistent hypercalcaemia (dichotomous outcome) • BMD of the distal radius or the lumbar spine (continuous outcome) • Cardiovascular events (dichotomous outcome) • Adverse events (to include voice change, hypoparathyroidism, hypothyroidism/hyperthyroidism) (dichotomous outcome) • Cancer incidence (dichotomous outcome) • Reoperation (for post-surgery stratum) (dichotomous outcome)
6.2	Prognostic	What are the long-term outcomes in people with primary hyperparathyroidism?	**Critical outcomes:** • Mortality (dichotomous outcome) • Fragility fracture (dichotomous outcome) • Renal stones (dichotomous outcome) • Renal tract calcification (dichotomous outcome) • Pancreatitis (dichotomous outcome) • Stroke (dichotomous outcome) • Hypertension (dichotomous outcome) • Myocardial infarction (dichotomous outcome) • Number of people who become eligible for surgery/ meet the criteria for surgery (dichotomous) • Serum calcium (>2.85 mmol/l) (dichotomous) (continuous if dichotomous not available)

Evidence report	Type of review	Review questions	Outcomes
			• 24-hour urine for calcium (>10 mmol/dl) (dichotomous) (continuous if dichotomous not available) • BMD of proximal femur (T-score <2.5; Z score <2) (dichotomous) (continuous if dichotomous not available)
7.1	Intervention	How should the management of primary hyperparathyroidism differ in pregnant women?	Outcomes follow those in the primary reviews for surgery, surgery interventions, calcimimetics, bisphosphonates, monitoring and patient information. Additional pregnancy/neonatal outcomes: • Outcome of pregnancy – term/early/late (dichotomous outcome) • Congenital abnormalities (dichotomous outcome) • Early foetal loss (miscarriage) (dichotomous outcome) • Stillbirth (dichotomous outcome) • Admission for IV hydration (dichotomous outcome) • Complications during pregnancy (dichotomous outcome) • Eclampsia/pre-eclampsia • Complications post-partum – mother/baby – requirement for support for either (dichotomous outcome) • Apgar score baby (continuous outcome) • Calcium levels mother/baby at/around birth (continuous outcome) • Neonatal tetany or symptomatic hypocalcaemia (dichotomous outcome)
8.1	Qualitative	What information is useful for people with primary hyperparathyroidism?	Any type of information described by studies. • Content of information and how this information is delivered • Information to include pre- and post-surgery • Timing of information and

Evidence report	Type of review	Review questions	Outcomes
			support

2.2 Searching for evidence

2.2.1 Clinical and health economics literature searches

Systematic literature searches were undertaken to identify all published clinical and health economic evidence relevant to the review questions. Searches were undertaken according to the parameters stipulated within the NICE guidelines manual 2014 (see https://www.nice.org.uk/media/default/about/what-we-do/our-programmes/developing-nice-guidelines-the-manual.pdf/). Databases were searched using relevant medical subject headings, free-text terms and study-type filters where appropriate. Where possible, searches were restricted to papers published in English. Studies published in languages other than English were not reviewed. All searches were updated on 06 August 2018. Papers published or added to databases after this date were not considered. If new evidence, falling outside of the timeframe for the guideline searches, is identified, for example in consultation comments received from stakeholders, the impact on the guideline will be considered, and any further action agreed between NGC and NICE staff with a quality assurance role.

Prior to running, search strategies were quality assured using a variety of approaches. Medline search strategies were checked by a second information specialist before being run. Searches were cross-checked with reference lists of highly relevant papers, searches in other systematic reviews were analysed, and committee members were requested to highlight additional studies.

Searching for unpublished literature was not undertaken. The NGC and 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.

Detailed search strategies can be found as an appendix to each evidence review.

2.3 Identifying and analysing evidence of effectiveness

Research fellows conducted the tasks listed below, which are described in further detail in the rest of this section:

- Identified potentially relevant studies for each review question from the relevant search results by reviewing titles and abstracts. Full papers were then obtained.
- Reviewed full papers against pre-specified inclusion and exclusion criteria to identify studies that addressed the review question in the appropriate population, and reported on outcomes of interest (review protocols are included in an appendix to each of the evidence reports).
- Critically appraised relevant studies using the appropriate study design checklist as specified in the NICE guidelines manual.[2]
- Extracted key information about interventional study methods and results using 'Evibase', NGC's purpose-built software. Evibase produces summary evidence tables, including critical appraisal ratings. Key information about non-interventional study methods and results was manually extracted onto standard evidence tables and critically appraised separately (evidence tables are included in an appendix to each of the evidence reports).
- Generated summaries of the evidence by outcome. Outcome data were combined, analysed and reported according to study design:

  - Randomised data were meta-analysed where appropriate and reported in GRADE profile tables.
  - Data from non-randomised studies were meta-analysed if appropriate. Where evidence was not meta-analysed results from single studies were presented separately.
  - Diagnostic data studies were meta-analysed where appropriate. Where evidence was not meta-analysed, because of insufficient data, results from single studies were presented separately.
- A sample of a minimum of 10% of the abstract lists of the first 3 sifts by new reviewers and those for complex review questions were double-sifted by a senior research fellow and any discrepancies were rectified. All of the evidence reviews were quality assured by a senior research fellow. This included checking:
  - papers were included or excluded appropriately
  - a sample of the data extractions
  - correct methods were used to synthesise data
  - a sample of the risk of bias assessments.

2.3.1 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 another appendix to each of the evidence reports. The committee was consulted about any uncertainty regarding inclusion or exclusion.

The key population inclusion criterion was:

- Adults (18 years or over) with confirmed or suspected primary hyperparathyroidism.

The key population exclusion criterion was people:

- with secondary and tertiary HPT
- with multiple endocrine neoplasia (MEN)
- with familial hyperparathyroidism
- with parathyroid carcinoma
- on medications interfering with calcium metabolism (for example, lithium).

Literature reviews, posters, letters, editorials, comment articles, unpublished studies and studies not in English were excluded.

2.3.2 Type of studies

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

For most intervention reviews in this guideline, parallel randomised controlled trials (RCTs) were included because they are considered the most robust type of study design that can produce an unbiased estimate of the intervention effects. If non-randomised intervention studies were considered appropriate for inclusion (for example, where no randomised evidence was available for critical outcomes) 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. Please refer to the review protocols in each evidence report for full details on the study design of studies selected for each review question.

For diagnostic review questions, diagnostic RCTs, cross-sectional studies and retrospective studies were included. For prognostic review questions, prospective and retrospective cohort studies were included. Case–control studies were not included.

Where data from non-randomised studies were included, the results for each outcome were presented separately for each study or meta-analysed if appropriate.

2.3.3 Methods of combining clinical studies

2.3.3.1 Data synthesis for intervention reviews

Where possible, meta-analyses were conducted using Cochrane Review Manager (RevMan5)[9] software to combine the data given in all studies for each of the outcomes of interest for the review question.

All analyses were stratified for normocalcaemia and pregnant women. For some questions additional stratification was used, and this is documented in the individual review question protocols in each evidence report.

2.3.3.1.1 Analysis of different types of data

Dichotomous outcomes

Fixed-effects Mantel–Haenszel techniques using an inverse variance method for pooling were used to calculate risk ratios (relative risk, RR) for the binary outcomes, which included:

- mortality
- deterioration in renal function
- fractures
- occurrence of kidney stones
- persistent hypercalcaemia
- adverse events
- cardiovascular events
- cancer incidence
- re-operation

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. Peto odds ratios are more appropriate for data with a low number of events. When there are zero events in one or both arms, we have not used GRADEpro to calculate the absolute risk reduction, but have analysed the data using the risk difference. In such a scenario when using GRADEpro software, we switch off the auto-calculation of absolute risk and enter the risk difference and its 95% confidence interval data (calculated in Revman) manually into GRADEpro.

Continuous outcomes

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

- Heath-related quality of life (HRQoL)
- Length of stay in hospital
- Bone mineral density (BMD) of the distal radius or the lumbar spine

Where the studies within a single meta-analysis had different scales of measurement, 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. Where an outcome has been reported as a final value in some studies and as a change score in others, final values and change scores have not been meta-analysed and have been reported as separate outcomes.

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 Cochrane Review Manager (RevMan5[9] software. Where p values were reported as 'less than', a conservative approach was undertaken. For example, if a p value was reported as 'p≤0.001', the calculations for standard deviations were based on a p value of 0.001. If these statistical measures were not available then the methods described in section 16.1.3 of the Cochrane Handbook (version 5.1.0, updated March 2011) were applied.

2.3.3.1.2 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.

2.3.3.1.3 Heterogeneity

Statistical heterogeneity was assessed for each meta-analysis estimate by considering the chi-squared test for significance at an I-squared (I^2) inconsistency statistic (with an I-squared value of more than 50% indicating significant heterogeneity) as well as the distribution of effects. Where significant heterogeneity was present, predefined subgrouping of studies was carried out as specified a priori in the review protocols, and included:

- IV versus oral bisphosphonates (Evidence report H)
- Adjusted serum calcium ≥2.85mmol/L and <2.85mmol/L) (Evidence reports C and G)
- People with end-organ effects versus absence of end-organ effects (end organ effects defined as kidney stones, history of fragility fractures or osteoporosis [BMD T-score <-2.5 at any site]) (Evidence reports C and G)
- First trimester, second trimester and third trimester of pregnancy at the time of management (Evidence report J)
- Reduction in creatinine clearance to < 60 mL/min and ≥60 ml/min (Evidence report C)
- Vitamin D replete versus not vitamin D replete prior to surgery (Evidence report E)
- Age under 50 years versus ≥50 years (Evidence report C)

If the subgroup analysis resolved heterogeneity within all of the derived subgroups, then each of the derived subgroups were adopted as separate outcomes (providing at least 1 study remained in each subgroup). Assessments of potential differences in effect between subgroups were based on the chi-squared tests for heterogeneity statistics between subgroups. Any subgroup differences were interpreted with caution as separating the groups breaks the study randomisation and as such is subject to uncontrolled confounding.

If all predefined strategies of subgrouping were unable to explain statistical heterogeneity within each derived subgroup, 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 described narratively.

2.3.3.2 Data synthesis for prognostic reviews

Odds ratios (ORs), risk ratios (RRs), or hazard ratios (HRs), with their 95% CIs, for the effect of the pre-specified prognostic factors were extracted from the studies.

Studies of lower risk of bias were preferred, taking into account the analysis and the study design. In particular, prospective cohort studies were preferred if they reported multivariable analyses that adjusted for key confounders identified by the committee at the protocol stage for that outcome.

2.3.3.3 Data synthesis for diagnostic test accuracy reviews

Two separate review protocols were produced to reflect the 2 different diagnostic study designs.

2.3.3.3.1 Diagnostic RCTs

Diagnostic RCTs (sometimes referred to as test and treat trials) are a randomised comparison of 2 diagnostic tests, with study outcomes being clinically important consequences of the diagnosis (patient-related outcome measures similar to those in intervention trials, such as mortality). Patients are randomised to receive test A or test B, followed by identical therapeutic interventions based on the results of the test (so someone with a positive result would receive the same treatment regardless of whether they were diagnosed by test A or test B). Downstream patient outcomes are then compared between the 2 groups. As treatment is the same in both arms of the trial, any differences in patient outcomes will reflect the accuracy of the tests in correctly establishing who does and does not have the condition. Data were synthesised using the same methods for intervention reviews (see section 2.3.3.1.1 above).

2.3.3.3.2 Diagnostic accuracy studies

Localisation index tests

The following adapted methods were used to assess the accuracy of the localisation index tests. Localisation index tests included ultrasound (US), sestamibi scanning (including planar, subtraction, SPECT or SPECT/CT), MRI, CT, 4D-CT, venous sampling and methylene blue. All of these index tests are used pre-operatively with the exception of methylene blue which is used intra-operatively. An adapted diagnostic accuracy method was used for this part of the review, as described below.

A standard diagnostic accuracy 2x2 table could not be used for this review, as there is more than 1 possible outcome for each person (unlike a standard diagnostic accuracy study where each person either has the disease or not). As each person has more than 1 parathyroid gland, there is more than 1 possible outcome for both the index test and the reference standard (i.e. imaging could predict 1 or more possible affected glands, and the final outcome could be a single adenoma, more than 1 adenoma, or hyperplasia).

Therefore, to overcome this problem, the following 2x2 table was devised at protocol stage for this review. This method was chosen as it allows the accuracy of the tests to be determined according to whether the imaging test would have predicted the correct surgical approach in each person (focused surgery or exploratory surgery). It was agreed that this approach would give the most relevant information for determining the most clinically effective localisation test.

		By the reference standard there was a single adenoma	
		YES	NO
Index test –		True positive (correct application of focused surgery) - imaging identifies a single adenoma location correctly	False positive (either focused surgery would fail or would convert to exploratory) - Imaging shows a single adenoma but there is actually a double adenoma - Imaging shows a single adenoma but there is actually hyperplasia - single on imaging but nothing found
		False Negative - nothing on imaging so single adenoma missed (do another imaging or exploratory surgery) - Imaging incorrectly identifies the location of a single adenoma (either surgery would fail or would convert to exploratory) - multiple findings on imaging but only a single adenoma located	True Negative (correct application of exploratory surgery) - Imaging shows nothing and there are no adenomas found - Imaging correctly identifies hyperplasia - Imaging correctly identifies double adenoma - Imaging shows nothing but there is actually a double adenoma - Imaging shows nothing but there is hyperplasia - Imaging shows multiple glands but not all in hyperplasia
	TOTAL	***Number of people with a single adenoma / should have focused***	***Number of people who should have exploratory surgery (either as no adenomas, hyperplasia or double adenomas).***

If a study provided enough evidence to categorise each included participant according to the above 2x2 table (both as to the localisation of affected tissue according to the index test and the final localisation outcome from the reference standard) then it was included. For example, if a study stated that a participant had an imaging scan suggesting a single adenoma but the final outcome determined by the histology and post-operative normocalcaemia was a 4-gland hyperplasia, this person would be counted as a false positive. If it was not possible to categorise all the included participants for a given study into the above 2x2 table, then the study was excluded (for example, in people with persistent hypercalcaemia following surgery, unless the results of a further operation were provided in order to determine the final location according to the reference standard, then it would not be possible to determine whether the location of the affected tissue found on pre-operative imaging was correct or not).

The reference standard test must be the best available method to determine the actual location(s) of the affected tissue. It was agreed that the reference standard should include both histology and post-operative serum calcium levels. Histology alone was not sufficient as the reference standard, as although it can prove the presence of an adenoma, post-operative normocalcaemia is also required to prove that there was no further affected tissue remaining. Normocalcaemia in isolation is also not sufficient, unless the person was normocalcaemic after a single gland was removed. This is because, if more than 1 gland is removed, normocalcaemia could result if 1 or both of the glands were abnormal, and histology is required to determine if 1 or both were abnormal. Any studies not reporting both histology and post-operative normocalcaemia, in order to determine the actual location of abnormal tissue, were excluded.

By the above method, sensitivity and specificity would not have the same interpretation as in a standard diagnostic review. Sensitivity and specificity could be interpreted as follows:

- Sensitivity = % of people who have a single adenoma, who are correctly picked up by imaging tests (also the % of people who would get correctly applied focused surgery).

- Specificity = % of people who should get exploratory surgery (final diagnosis is >1 adenoma or hyperplasia), that do (imaging shows no adenoma, hyperplasia or double adenoma).

An index test with a low sensitivity (resulting from a high number of people in the bottom left cell) may mean that more people end up getting exploratory surgery who could have had focused surgery (if imaging shows more adenomas then there actually are), or it may mean that more people having failed surgery (if imaging shows the incorrect location of a single adenoma, although this may be picked up during the surgical procedure). An index test with a low specificity (resulting from a high number of people in the top right cell) may mean that more people would fail focused surgery and have persistent PHPT (as imaging would predict a single adenoma but they actually have >1).

Some diagnostic accuracy studies identified in the search provided accuracy data in different formats. These studies were only included in this review if it was possible to categorise all included participants in the study according to the above 2x2 table method. Some studies used a 'per-gland' method, assuming each person had 4 parathyroid glands and therefore determining 4 possible outcomes in the 2x2 table for each person. For example, if a person had 1 suspected adenoma located on imaging, and the reference standard confirmed a single adenoma at the same location, that person would have 1 true positive and 3 true negative results. Or, if a person had 1 suspected adenoma located on imaging but the final outcome according to the reference standard was 4-gland hyperplasia, then that person would be deemed to have 1 true positive and 3 false negative results. Another method adopted by some studies was an adapted 'per-person' method. If a person had all affected glands (either a single adenoma or more than 1 gland) correctly identified on imaging then they would be deemed a true positive. However, this causes problems of how to categorise people who have all their affected glands correctly identified on imaging, but the imaging also suggests further affected tissue in a location which is normal according to the reference standard. These people would be deemed to be true positives, even though relying on the imaging result alone would result in more glands being explored at surgery than was necessary.

Neither of the above methods ('per-gland' and 'per-person') were used for this review. The method used in this review was chosen as it allows the accuracy of the tests to be determined according to whether the imaging test would have predicted the correct surgical approach (focused surgery or exploratory surgery).

2.3.3.4 Intra-operative tests – diagnostic accuracy methods

The intra-operative tests of IOPTH and intra-operative frozen sections are not used to aid localisation of the affected tissue, but rather are used to determine whether all the affected tissue has been excised and whether surgery can be terminated. Therefore the method of assessing accuracy of these tests is different to the localisation tests.

The following 2x2 table was used to assess the accuracy of IOPTH and intra-operative frozen sections for predicting whether all abnormal tissue has been removed or not:

		Reference standard	
		+ve	-ve
Index test	+ve	True positive (>50% fall in PTH and all adenomas removed)	False positive (>50% fall in PTH but not all adenomas removed – person remains hypercalcaemic (up to 6 months) or requires re-op or subsequent glands resected in the same op)
	-ve	False Negative (no fall in PTH but all adenomas removed)	True Negative (no fall in PTH and not all adenomas removed – person remains hypercalcaemic (up to 6 months) or requires re-op or subsequent glands resected in the same op)
	TOTAL	***Reference standard positive***	***Reference standard negative***

Again, the reference standard was histology and post-operative serum calcium. Studies only stating the accuracy for prediction of post-operative normocalcaemia, without mention of histology, were excluded (unless all participants had normocalcaemia after removal of a single gland only). This is because, if >1 gland is removed, normocalcaemia is insufficient to determine whether 1 or both were abnormal. For example, IOPTH may not have fallen after removal of the first gland, so surgery continued and IOPTH fell after removal of the second gland. Without histology, it is not possible to classify the IOPTH result after removal of the first gland as a false negative or a true negative.

In this context:

- Sensitivity = the ability to identify people who have had all adenomas removed
- Specificity = the ability to identify people who have remaining abnormal tissue

An index test with a low sensitivity may result from a high proportion of people not having a drop in IOPTH even when all abnormal tissue has been removed, which may result in continuing to explore other glands unnecessarily if the decision to terminate surgery is based on the IOPTH alone. An index test with a low specificity may result from a high proportion of people having a drop in the IOPTH even though there is still abnormal tissue remaining and therefore, if the decision to terminate surgery is based on the IOPTH alone, the surgery would be terminated and the person would remain hypercalcaemic and require further surgery. As the outcome of a false positive result is arguably worse than the outcome of a false negative result, IOPTH would need to have a high specificity in order to be recommended.

For IOPTH, it is possible to calculate the 2x2 table values in different ways for people who had >1 gland removed (i.e. for people with multigland disease). As there will be an IOPTH result after excision of the first gland (if this is negative in people who have remaining abnormal tissue and go on to have further glands excised, then people with MGD will be counted as true negatives) and an IOPTH result after excision of all abnormal glands (if this is positive in people with MGD once all their glands have been removed then people with MGD will be counted as true positives). In some studies, both methods can be calculated as they may report (in people with MGD) a negative IOPTH after excision of their first gland (a true negative due to remaining abnormal tissue), but a positive IOPTH after excision of all the abnormal glands (a true positive if all glands are removed and the person is rendered normocalcaemic). The preferred method for this review is to find the IOPTH accuracy after excision of a single gland or excision of the first gland (in people with MGD). This is because the predominant use of IOPTH is likely to be in focused surgery and the accuracy for predicting whether further abnormal tissue remains. Therefore, if it was possible to calculate both methods from a study, the result after excision of the first gland was preferred.

There are various criteria for the IOPTH test to indicate a positive result. The criterion specified in this review was the Miami criteria (a drop in parathyroid hormone at 10 minutes post-excision of at least 50% of the highest baseline value (either pre-incision or pre-excision). However, studies were also included if they used a 50% drop in PTH from either baseline value. Studies using the criteria of a 50% drop ***and*** into the normal/reference range for PTH were excluded (unless a drop of 50% alone [regardless of whether it went into the normal range] could be calculated). The protocol also specified PTH values taken at 5 or 20 minutes post-excision.

Coupled forest plots of sensitivity and specificity with their 95% CIs across studies were produced for each index test, using RevMan5. [9] In order to do this, 2×2 tables (the number of true positives, false positives, true negatives and false negatives) were derived from raw data or calculated from the set of test accuracy statistics.

Diagnostic meta-analysis was conducted when 3 or more studies were available per index test. 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.[12] The advantage of this approach is that it produces summary estimates of sensitivity and specificity that account for the correlation between the 2 statistics. Other advantages of this method have been described elsewhere.[8, 10, 11] 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.[6]). Pooled sensitivity and specificity and their 95% CIs were reported in the clinical evidence summary tables. If values could not be pooled, then the individual sensitivity values and their coupled specificity were presented in the clinical evidence summary.

If appropriate, to allow comparison between index tests, summary ROC curves were generated for each index test from the pairs of sensitivity and specificity calculated from the 2×2 tables. An 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 index tests, 2 or more index tests were plotted on the same graph. The performance of the different index 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 index test.

Heterogeneity or inconsistency amongst studies was visually inspected in the forest plots and pooled diagnostic meta-analysis plots. For risk of bias, if a study sub selected people with single gland disease or suspected single gland disease from imaging – this was not considered as a limitation if IOPTH was the index test, but if imaging was the index test then it was downgraded for indirectness.

2.3.3.5 Data synthesis for qualitative study reviews

No evidence was available for the qualitative studies review.

2.3.4 Appraising the quality of evidence by outcomes

2.3.4.1 Intervention reviews

The evidence for outcomes from the included RCTs and, where appropriate, non-randomised intervention studies, were evaluated and presented using an adaptation of 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 2.

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 or other bias was only taken into consideration in the quality assessment if it was apparent.

2.3.4.1.1 Risk of bias

The main domains of bias for RCTs are listed in Table 3. Each outcome had its risk of bias assessed within each study first. For each study, if there were no risks of bias in any domain, the risk of bias was given a rating of 0. If there was risk of bias in just 1 domain, the risk of bias was given a 'serious' rating of −1, but if there was risk of bias in 2 or more domains the risk of bias was given a 'very serious' rating of −2. A weighted average score was then calculated across all studies contributing to the outcome, by taking into account the weighting of studies according to study precision. For example if the most precise studies tended to each have a score of −1 for that outcome, the overall score for that outcome would tend towards −1.

Table 3: Principle domains of bias in randomised controlled trials

Limitation	Explanation
Selection bias (sequence generation and	If those enrolling patients 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

Limitation	Explanation
allocation concealment)	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 of patients and healthcare professionals)	Patients, caregivers, those adjudicating or recording outcomes, and data analysts should not be aware of the arm to which patients 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 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. • 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, as they are inherently at high risk of selection bias. For this reason, GRADE requires that non-randomised evidence is initially downgraded on the basis of study design, starting with a rating of −2. This accounts for selection bias and so non-randomised intervention studies are not downgraded any further on that domain. Non-randomised evidence was assessed against the remaining domains used for RCTs in Table 3, and downgraded further as appropriate.

2.3.4.1.2 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 0. If there was indirectness in just 1 source (for example in terms of population), indirectness was given a 'serious' rating of −1, but if there was indirectness in 2 or more sources (for example, in terms of population and treatment) the indirectness was given a 'very serious' rating of −2. A weighted average score was then calculated across all studies contributing to the outcome by taking into account study precision. For example, if the most precise studies tended to have indirectness score of −1 each for that outcome, the overall score for that outcome would tend towards −1.

2.3.4.1.3 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. 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' score of −1 if the I^2 was 50–74% and a 'very serious' score of −2 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%), the committee took this into account and considered whether to make separate recommendations on new outcomes based on the subgroups defined by the assumed explanatory factors. In such a situation the quality of evidence was not downgraded for those emergent outcomes.

Since the inconsistency score was based on the meta-analysis results, the score represented the whole outcome and so weighted averaging across studies was not necessary.

2.3.4.1.4 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 the end of the 95% CI of the overall estimate of effect crossed 1 of the MID lines, imprecision was regarded as serious and a 'serious' score of −1 was given. 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 and a 'very serious' score of −2 was given. 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 2. As for inconsistency, since the imprecision score was based on the meta-analysis results, the score represented the whole outcome and so weighted averaging across studies was not necessary.

The 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. For binary variables, any MIDs reported in the literature will inevitably be based on expert consensus, as such MIDs relate to all-or-nothing population effects rather than measurable effects on an individual, and so are not amenable to patient-centred 'anchor' methods.

For this guideline, established MIDs from the literature were used for SF-36 (health-related quality of life) . These values are displayed below:

Table 4: MIDs used to assess imprecision for SF-36 measures

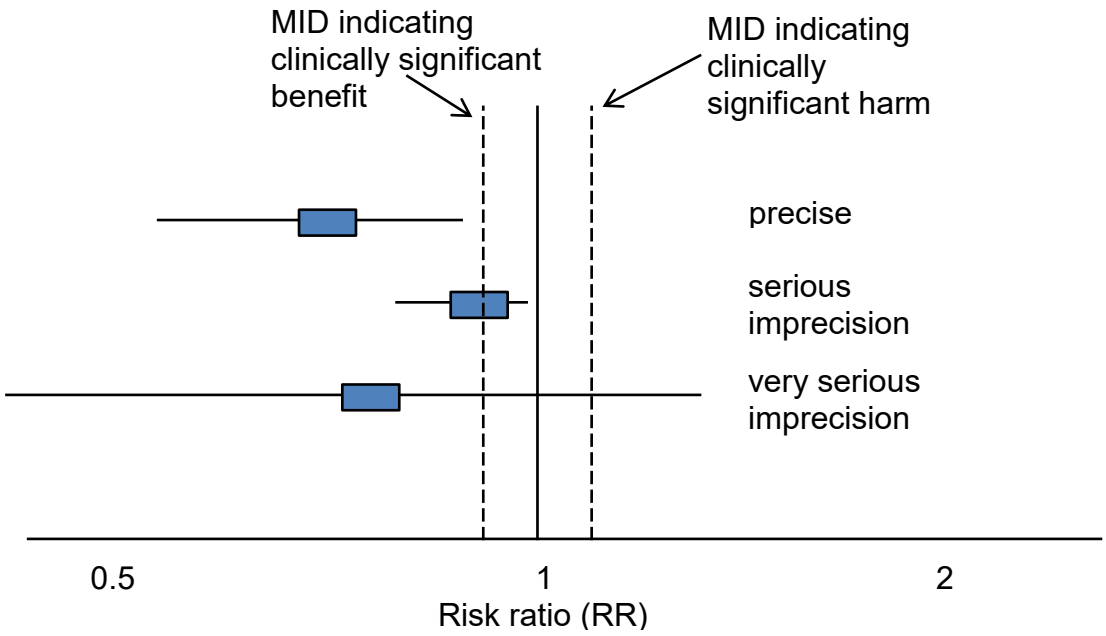
MIDs for assessing between group differences Outcome	MID for imprecision	MID for clinical importance	Source
SF-36	Physical component summary: 2 Mental component summary: 3 Physical functioning: 3 Role-physical: 3 Bodily pain: 3 General health: 2 Vitality: 2 Social functioning: 3 Role-emotional: 4 Mental health: 3		User's manual for the SF-36v2 Health Survey, Third Edition

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

- For categorical 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 significant harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically significant 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 significant benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically significant harm.
- For one or more studies reporting zero events in one arm, the aforementioned default MIDs were used to assess imprecision.
- For a pooled estimate from one or more studies reporting zero events in both arms the imprecision was assessed based on sample size. For a sample size >350 it was assessed as no imprecision, for a sample size >70 <350 imprecision was assessed as serious and for a sample size <70 imprecision was assessed as very serious.
- 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 significant 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 significant 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.
- If standardised mean differences have been used, then the MID will be set at the absolute value of +0.5. This follows because standardised mean differences are mean differences normalised to the pooled standard deviation of the 2 groups, and are thus effectively expressed in units of 'numbers of standard deviations'. The 0.5 MID value in this context therefore indicates half a standard deviation, the same definition of MID as used for non-standardised mean differences.

The default MID value was subject to amendment after discussion with the committee. If the committee decided that the MID level should be altered, after consideration of absolute as well as relative effects, this was allowed, provided that any such decision was not influenced by any bias towards making stronger or weaker recommendations for specific outcomes.

Figure 2: 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)



2.3.4.1.5 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. The scores (0, −1 or −2) from each of the main quality elements were summed to give a score that could be anything from 0 (the best possible) to −8 (the worst possible). However scores were capped at −3. This final score was then applied to the starting grade that had originally been applied to the outcome by default, based on study design. All RCTs started as High and the overall quality became Moderate, Low or Very Low if the overall score was −1, −2 or −3 points respectively. The significance of these overall ratings is explained in Table 5. The reasons for downgrading in each case were specified in the footnotes of the GRADE tables.

Non-randomised intervention studies started at Low, and so a score of −1 would be enough to take the grade to the lowest level of Very Low. Non-randomised intervention studies could, however, be upgraded if there was a large magnitude of effect or a dose-response gradient.

Table 5: 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

Level	Description
	the estimate of effect and is likely to change the estimate
Very low	Any estimate of effect is very uncertain

2.3.4.2 Prognostic reviews

The quality of evidence for prognostic studies was evaluated according to the criteria given in Table 6. 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.

Table 6: Description of quality elements for prospective studies

Quality element	Description of cases where the quality measure would be downgraded
Study design	Case–control studies rather than prospective cohort studies
Patient recruitment	If potential for selection bias
Validity of risk factor measure(s)	If non-validated and no reasonable face validity
Validity of outcome measure	If non-validated and no reasonable face validity
Blinding	If assessors of outcome not blinded to risk factor measurement (or vice versa)
Adequate duration of follow-up (or retrospective duration)	If follow-up (or retrospective) period inadequate to allow events to occur, or retrospective period so short that causality is in doubt because the outcome may have preceded the risk factor
Confounder consideration	If there is a lack of consideration of all reasonable confounders in a multivariable analysis
Attrition	If attrition is too high and there is no attempt to adjust for this
Directness	If the population, risk factors or outcome differ from that in the review question

2.3.4.2.1 Inconsistency

Inconsistency was assessed as for intervention studies.

2.3.4.2.2 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.

2.3.4.2.3 Overall grading

Quality rating started at High for prospective studies, and each major limitation brought the rating down by 1 increment to a minimum grade 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 not the appropriate study designs to answer such questions.

2.3.4.3 Diagnostic studies

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[2]). Risk of bias and applicability in primary diagnostic accuracy studies in QUADAS-2 consists of 4 domains (see Figure 3):

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

Figure 3: 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 . 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 have introduced bias?	Could the reference standard, its conduct or its interpretation have introduced bias?	Could the patient flow 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?	

2.3.4.3.1 Inconsistency

Inconsistency refers to an unexplained heterogeneity of results for an outcome across different studies. Inconsistency was assessed by inspection of the sensitivity or specificity value (based on the primary measure) using the point estimates and 95% CIs of the individual studies on the forest plots. The evidence was downgraded by 1 increment if the individual studies varied across 2 areas [(for example, 50–90% and 90–100%)] and by 2 increments if the individual studies varied across 3 areas [(for example, 0–50%, 50–90% and 90–100%)].

2.3.4.3.2 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. As a general rule (after discussion with the committee) a variation of 0–20% was considered precise, 20–40% serious imprecision, and >40% very serious imprecision. Imprecision was assessed on the primary outcome measure for decision-making.

2.3.4.3.3 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.

2.3.4.4 Qualitative reviews

No evidence was available for the qualitative studies evidence review.

2.3.5 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 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 the critical outcome of 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 the outcome mortality assessment for clinically important difference was carried out by the committee on a case by case basis.

This assessment was carried out by the committee for each critical outcome, and an evidence summary table was produced to compile the committee's assessments of clinical importance per outcome, alongside the evidence quality and the uncertainty in the effect estimate (imprecision).

2.3.6 Clinical evidence statements

Clinical evidence statements are summary statements that are included in each evidence report, and which summarise the key features of the clinical effectiveness evidence presented. The evidence statements are presented by outcome and encompass the following key features of the evidence:

- The number of studies and the number of participants for a particular outcome.
- An indication of the direction of clinical importance (if one treatment is beneficial or harmful compared to the other or whether there is no difference between the 2 tested treatments).
- A description of the overall quality of the evidence (GRADE overall quality).

2.4 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.[2]

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.
- Undertook new cost-effectiveness analysis in priority areas.

2.4.1 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 pre-specified 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.[2]
- 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.

2.4.1.1 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 2002 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. 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 7 below and the economic evaluation checklist (appendix H of the NICE guidelines manual[2]) 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.

2.4.1.2 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.[2] 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 7 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.[7]

Table 7: 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. • 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.

Item	Description
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[2]

2.4.2 Undertaking new health economic analysis

As well as reviewing the published health economic literature for each review question, as described above, new health economic analysis was undertaken by the health economist in selected areas. Priority areas for new analysis were agreed by the committee after formation of the review questions and consideration of the existing health economic evidence.

The committee identified intraoperative parathyroid hormone (IOPTH) testing as the highest priority area for original health economic modelling. At present, IOPTH tests are not routinely used as part of parathyroid surgery, and only a limited number of hospitals have the required equipment. Given the high costs of this equipment as well as that of running the tests, any recommendation that will increase the use of IOPTH tests could potentially lead to a large increase in healthcare resource use.

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

- Methods were consistent with the NICE reference case for interventions with health outcomes in NHS settings.[2, 4]
- The committee was involved in the design of the model, selection of inputs and interpretation of the results.
- Model inputs were based on the systematic review of the clinical literature supplemented with other published data sources where possible.
- When published data were not available committee expert opinion was used to populate the model.
- Model inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.
- The model was peer-reviewed by another health economist at the NGC.

Full methods and results of the cost-effectiveness analysis for IOPTH testing are described in a separate economic analysis report.

2.4.3 Cost-effectiveness criteria

NICE's report 'Social value judgements: principles for the development of NICE guidance' sets out the principles that committees should consider when judging whether an intervention offers good value for money.[3] 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 the factors set out in 'Social value judgements: principles for the development of NICE guidance'.[3]

If a study reported the cost per life year gained but not QALYs, the cost per QALY gained was estimated by multiplying by an appropriate utility estimate to aid interpretation. The estimated cost per QALY gained is reported in the health economic evidence profile with a footnote detailing the life-years gained and the utility value used.

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.

2.4.4 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.

2.5 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–K]).
- 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 and summary ROC curves (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).

Recommendations were drafted 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. Firstly, the net clinical benefit over harm (clinical effectiveness) was considered, focusing on the critical outcomes. 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 clinical and health economic evidence was of poor quality, conflicting or absent, the committee drafted recommendations 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 2.5.1 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[2]).

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

2.5.1 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.

2.5.2 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.

2.5.3 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.

2.5.4 Disclaimer

Healthcare providers need to use clinical judgement, knowledge and expertise when deciding whether it is appropriate to apply guidelines. The recommendations cited here are a guide and may not be appropriate for use in all situations. The decision to adopt any of the recommendations cited here must be made by practitioners in light of individual patient circumstances, the wishes of the patient, clinical expertise and resources.

The National Guideline Centre disclaims any responsibility for damages arising out of the use or non-use of this guideline and the literature used in support of this guideline.

2.5.5 Funding

The National Guideline Centre was commissioned by the National Institute for Health and Care Excellence to undertake the work on this guideline.

3 Acronyms and abbreviations

Acronym or abbreviation	Description
25OHD	25-hydroxyvitamin D
4DCT	4-dimensional computed tomography
AUC	Area under the curve
BMD	Bone mineral density
CCCR	Calcium/creatinine clearance ratio
CE	Renal calcium excretion
CEA	Cost-effectiveness analysis
CI	Confidence interval
CR	Renal calcium/creatinine excretion ratio
CUA	Cost-utility analysis
DXA	Dual-energy X-ray absorptiometry
FHH	Familial hypocalciuric hypercalcaemia
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HPT	Hyperparathyroidism
HRQOL	Health-related quality of life
ICER	Incremental cost-effectiveness ratio
IOPTH	Intra-operative parathyroid hormone
MDT	Multidisciplinary team
MEN	Multiple endocrine neoplasia
MGD	Multigland disease
MIBI	Sestamibi
MIP	Minimally invasive parathyroidectomy
MIPUSS	Minimally invasive parathyroidectomy with intra-operative surgical sonography
NGC	National Guideline Centre
NICE	National Institute for Health and Care Excellence
NIH	National Institutes of Health
NRS	Non-randomised studies
OECD	Organisation for Economic Co-operation and Development
PAS	Parathyroid assessment of symptoms
PHPT	Primary hyperparathyroidism
PTH	Parathyroid hormone
QALY	Quality-adjusted life year
RCT	Randomised controlled trial
ROC	Receiver operating characteristic
SF-36	Short-form 36 questionnaire
SPECT	Single-photon emission computed tomography
VAP	Video assisted parathyroidectomy
VFA	Vertebral fracture assessment

4 Glossary

The NICE Glossary can be found at www.nice.org.uk/glossary.

4.1 Guideline-specific terms

Term	Definition
4-gland exploration	Surgical treatment for primary hyperparathyroidism involving identification of all 4 parathyroid glands.
Adenoma	A benign (non-cancerous) tumour.
Bilateral neck exploration	See 4-gland exploration.
Bisphosphonates	A class of drugs that reduce bone loss and increase bone mineral density.
Calcimimetics	A class of drugs that reduce serum levels of PTH and calcium through their effect on the calcium-sensing receptor on parathyroid cells.
Cardiovascular disease (CVD)	Diseases that involve the heart, the blood vessels, or both, caused by atherosclerosis.
Dual-energy X-ray absorptiometry (DXA) scan	A bone density X-ray.
Eclampsia	One or more convulsions (seizures) in pregnant women with pre-eclampsia (see below).
Ectopic parathyroid adenoma	Any parathyroid gland that is not located next to the thyroid.
End-organ disease	Renal stones, fragility fractures or osteoporosis.
Familial hypocalciuric hypercalcaemia (FHH)	A rare inherited condition with slightly increased levels of plasma calcium, low urinary calcium excretion, and normal to moderately elevated plasma parathyroid hormone (PTH).
Focused parathyroidectomy	Surgical treatment for primary hyperparathyroidism involving the surgeon targeting only the hyperactive gland that is identified on preoperative localisation tests.
Hypercalcaemia	An albumin-adjusted serum calcium level of 2.6 mmol/litre or above.
Hyperplasia	Enlargement of an organ or tissue (in the context of primary hyperparathyroidism, enlargement of the parathyroid gland[s]).
Intra-operative frozen sections	Histological evaluation of tissue specimens during surgery. This is used to determine whether all the affected tissue has been excised during surgery and whether surgery can be terminated.
Intra-operative parathyroid hormone (IOPTH)	A technique where PTH is monitored intraoperatively with the aim of determining whether all the affected tissue has been excised and whether surgery can be terminated.
Minimally invasive parathyroidectomy	A term often used to refer to focused parathyroidectomy (see above) which can be used for any parathyroid operation done through a very small incision.
Nephrolithiasis	Kidney stones or renal stones – these are small deposits that build up in the kidneys.
Nephrocalcinosis	Also known as Albright's calcinosis or Anderson-Carr kidneys, a term used to describe deposition of calcium salts in the renal parenchyma due to hyperparathyroidism.
Neonatal tetany	A disorder that occurs in newborns, characterised by periodic painful muscular spasms and tremors caused by faulty calcium metabolism and associated with diminished function of the parathyroid glands.
Normocalcaemic primary hyperparathyroidism	An albumin-adjusted serum calcium of below 2.6 mmol/litre and an elevated parathyroid hormone level that cannot be explained by

Term	Definition
	abnormal renal function or low vitamin D.
Parathyroid surgery	Surgery targeted at the parathyroid.
Parathyroidectomy	Removal of parathyroid tissue.
Parathyroid hormone (PTH)	A hormone secreted by the 4 parathyroid glands which regulates calcium levels in the blood.
Parathyroid venous sampling	An invasive imaging technique involving inserting a catheter in the femoral vein and selective catheterisation and sampling of PTH in multiple neck and mediastinal veins.
Persistent hypercalcaemia	Persistent elevation of serum calcium levels (calcium ≥2.6 mmol/litre) following surgery.
Polyuria	Excessive urination volume.
Pre-eclampsia	A disorder of pregnancy characterised by high blood pressure and protein in the urine.
Primary hyperparathyroidism (PHPT)	A disorder of the parathyroid glands whereby one or more of the glands produce too much parathyroid hormone (PTH), causing blood calcium levels to rise (hypercalcaemia).
Serum adjusted calcium	The reference range for serum adjusted calcium is 2.2 to 2.6 mmol/litre.
Sestamibi (MIBI)	A non-invasive imaging technique in nuclear medicine. Images can be obtained of diseased parathyroid glands which have absorbed a radioactive marker.
Single-photon emission computed tomography (SPECT)	A three-dimensional variant of sestamibi scanning.
Urolithiaisis	The process of stone formation in the kidney, bladder, and/or urethra (urinary tract).

4.2 General terms

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

Term	Definition
	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 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

Term	Definition
	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)	There is always some uncertainty in research. This is because a small group of patients is studied to predict the effects of a treatment on the wider population. The confidence interval is a way of expressing how certain we are about the findings from a study, using statistics. It gives a range of results that is likely to include the 'true' value for the population. The CI is usually stated as '95% CI', which means that the range of values has a 95 in a 100 chance of including the 'true' value. For example, a study may state that "based on our sample findings, we are 95% certain that the 'true' population blood pressure is not higher than 150 and not lower than 110". In such a case the 95% CI would be 110 to 150. A wide confidence interval indicates a lack of certainty about the true effect of the test or treatment – often because a small group of patients has been studied. A narrow confidence interval indicates a more precise estimate (for example, if a large number of patients have been studied).
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	Cost–consequences analysis is one of the tools used to carry out an economic evaluation. This compares the costs (such as treatment and

Term	Definition
analysis (CCA)	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.
Decision analysis	An explicit quantitative approach to decision-making under uncertainty, based on evidence from research. This evidence is translated into probabilities, and then into diagrams or decision trees which direct the clinician through a succession of possible scenarios, actions and outcomes.
Deterministic analysis	In economic evaluation, this is an analysis that uses a point estimate for each input. In contrast, see Probabilistic analysis
Discounting	Costs and perhaps benefits incurred today have a higher value than costs and benefits occurring in the future. Discounting health benefits reflects individual preference for benefits to be experienced in the present rather than the future. Discounting costs reflects individual preference for costs to be experienced in the future rather than the present.
Disutility	The loss of quality of life associated with having a disease or condition. See Utility
Dominance	A health economics term. When comparing tests or treatments, an option that is both less effective and costs more is said to be ‘dominated’ by the alternative.
Drop-out	A participant who withdraws from a trial before the end.
Economic evaluation	An economic evaluation is used to assess the cost effectiveness of healthcare interventions (that is, to compare the costs and benefits of a healthcare intervention to assess whether it is worth doing). The aim of an economic evaluation is to maximise the level of benefits – health effects – relative to the resources available. It should be used to inform and support the decision-making process; it is not supposed to replace the judgement of healthcare professionals. There are several types of economic evaluation: cost–benefit analysis, cost–consequences analysis, cost-effectiveness analysis, 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

Term	Definition
	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.
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.
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 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 profile.
Harms	Adverse effects of an intervention.
Health economics	Study or analysis of the cost of using and distributing healthcare resources.
Health-related quality of life (HRQoL)	A measure of the effects of an illness to see how it affects someone's day-to-day life.
Heterogeneity or Lack of homogeneity	The term is used in meta-analyses and systematic reviews to describe when the results of a test or treatment (or estimates of its effect) differ significantly in different studies. Such differences may occur as a result of differences in the populations studied, the outcome measures used or because of different definitions of the variables involved. It is the opposite of homogeneity.
Imprecision	Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of effect.
Inclusion criteria (literature review)	Explicit criteria used to decide which studies should be considered as potential sources of evidence.
Incremental analysis	The analysis of additional costs and additional clinical outcomes with

Term	Definition
	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.
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 more healthy 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).
Logistic regression or Logit model	In statistics, logistic regression is a type of analysis used for predicting the outcome of a binary dependent variable based on one or more predictor variables. It can be used to estimate the log of the odds (known as the ‘logit’).
Loss to follow-up	A patient, or the proportion of patients, actively participating in a clinical trial at the beginning, but whom the researchers were unable to trace or contact by the point of follow-up in the trial
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)
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

Term	Definition
	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.
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.
Odds ratio	Odds are a way to represent how likely it is that something will happen (the probability). An odds ratio compares the probability of something in one group with the probability of the same thing in another. An odds ratio of 1 between 2 groups would show that the probability of the event (for example a person developing a disease, or a treatment working) is the same for both. An odds ratio greater than 1 means the event is more likely in the first group. An odds ratio less than 1 means that the event is less likely in the first group. Sometimes probability can be compared across more than 2 groups – in this case, one of the groups is chosen as the 'reference category', and the odds ratio is calculated for each group compared with the reference category. For example, to compare the risk of dying from lung cancer for non-smokers, occasional smokers and regular smokers, non-smokers could be used as the reference category. Odds ratios would be worked out for occasional smokers compared with non-smokers and for regular smokers compared with non-smokers. See also confidence interval, risk ratio.
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 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.

Term	Definition
	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.
Polypharmacy	The use or prescription of multiple medications.
Posterior distribution	In Bayesian statistics this is the probability distribution for a statistic based after combining established information or belief (the prior) with new evidence (the likelihood).
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.
Post-test probability	In diagnostic tests: The proportion of patients with that particular test result who have the target disorder (post-test odds/[1 plus post-test odds]).
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.
Pre-test probability	In diagnostic tests: The proportion of people with the target disorder in the population at risk at a specific time point or time interval. Prevalence may depend on how a disorder is diagnosed.
Prevalence	See Pre-test probability.
Prior distribution	In Bayesian statistics this is the probability distribution for a statistic based on previous evidence or belief.
Primary care	Healthcare delivered outside hospitals. It includes a range of services provided by GPs, nurses, health visitors, midwives and other healthcare professionals and allied health professionals such as dentists, pharmacists and opticians.
Primary outcome	The outcome of greatest importance, usually the one in a study that the power calculation is based on.
Probabilistic analysis	In economic evaluation, this is an analysis that uses a probability distribution for each input. In contrast, see Deterministic analysis.
Product licence	An authorisation from the MHRA to market a medicinal product.
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

Term	Definition
	studies showing that a treatment works well and don't publish those showing it did not have any effect. If this happens, analysis of the published results will not give an accurate idea of how well the treatment works. This type of bias can be assessed by a funnel plot.
Quality of life	See 'Health-related quality of life'.
Quality-adjusted life year (QALY)	A measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life. One QALY is equal to 1 year of life in perfect health. QALYS are calculated by estimating the years of life remaining for a patient following a particular treatment or intervention and weighting each year with a quality of life score (on a scale of 0 to 1). It is often measured in terms of the person's ability to perform the activities of daily life, freedom from pain and mental disturbance.
Randomisation	Assigning participants in a research study to different groups without taking any similarities or differences between them into account. For example, it could involve using a random numbers table or a computer-generated random sequence. It means that each individual (or each group in the case of cluster randomisation) has the same chance of receiving each intervention.
Randomised controlled trial (RCT)	A study in which a number of similar people are randomly assigned to 2 (or more) groups to test a specific drug or treatment. One group (the experimental group) receives the treatment being tested, the other (the comparison or control group) receives an alternative treatment, a dummy treatment (placebo) or no treatment at all. The groups are followed up to see how effective the experimental treatment was. Outcomes are 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.

Term	Definition
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 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.

Term	Definition
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.
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.
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).

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: 17/09/2018.

2. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. London. National Institute for Health and Care Excellence, 2014. Available from: http://www.nice.org.uk/article/PMG20/chapter/1%20Introduction%20and%20overview

3. 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

4. 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

5. National Institute for Health and Clinical Excellence. The guidelines manual. London. National Institute for Health and Clinical Excellence, 2018. Available from: https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview

6. 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

7. Organisation for Economic Co-operation and Development (OECD). Purchasing power parities (PPP). Available from: http://www.oecd.org/std/prices-ppp/ Last accessed: 13/12/2017.

8. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58(10):982-990

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. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002; 21(4):589-624

11. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine. 1993; 12(24):2273-2284

12. 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/