National Institute for Health and Care Excellence

Draft for consultation

Vitamin B12 deficiency in over 16s: diagnosis and management

NICE guideline: methods

NICE guideline <number>

Methods

July 2023

Draft for Consultation

Developed by NICE

1

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN:
[]

Contents

1

1 Development of the guideline

1.1 Remit

NICE received the remit for this guideline from NHS England.

The remit for this guideline is:

To develop a guideline on pernicious anaemia.

To see “What this guideline covers” and “What this guideline does not cover” please see the guideline scope Vitamin B12 deficiency, including pernicious anaemia: diagnosis and management.*

*The guideline title was changed prior to consultation to “Vitamin B12 deficiency in people aged 16 and over: diagnosis and management”.

The guideline name changed because autoimmune gastritis was thought to be a more accurate term for the condition related to vitamin B12 deficiency. Autoimmune gastritis is a chronic inflammatory disease that destroys parietal cells in the stomach, affecting vitamin B12 absorption and potentially causing a deficiency of the vitamin. Pernicious anaemia (PA) is a more variably defined condition, which commonly often has different meanings in different contexts. The term pernicious anaemia has often been used interchangeably with autoimmune gastritis and this has created confusion.

2 Methods

This guideline was developed using the methods described in the NICE guidelines manual[4] as outlined in Table 1 below.

Table 1 Versions of the NICE guidelines manual followed during guideline development and guideline validation

Stage	2018 update	2020 update	2022 update
Scoping		✓	
Development			✓
Validation			✓

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

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

2.1 Developing the review questions and outcomes

The review questions developed for this guideline were based on the key areas and draft review questions identified in the guideline scope. They were drafted by the technical team, refined and validated by the committee and signed off by NICE. A total of 13 review questions were developed in this guideline and outlined in Table 2.

The review questions were based on the following frameworks:

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

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

Table 2: Review questions

Evidence report	Type of review	Review questions	Outcomes
Evidence review A Information and support	Qualitative	What information and support is needed by people with suspected or confirmed vitamin B12 deficiency caused by a lack of vitamin B12 in their diet, and their families or carers?	• Not applicable. Themes emerged from the studies and were not predefined

Evidence report	Type of review	Review questions	Outcomes
		What information and support is needed by people with suspected or confirmed vitamin B12 deficiency caused by inadequate absorption of vitamin B12 (including pernicious anaemia), and their families or carers, and when should this be provided?	• Not applicable. Themes emerged from the studies and were not predefined
Evidence review B Risk factors and signs and symptoms	Prognostic	What are the risk factors for vitamin B12 deficiency?	• Diagnosis of vitamin B12 deficiency at any time point reported in the study (adjusted odds ratios, risk ratios or hazard ratios).
	Diagnostic	What signs and symptoms are indicative of vitamin B12 deficiency?	• Diagnostic association of signs and symptoms with a confirmed diagnosis of vitamin B12 deficiency. Measured by: - Diagnostic accuracy data (sensitivity (prioritised), specificity, PPV, NPV) - Association data (adjusted RR or OR)
Evidence review C Diagnosis	Diagnostic	What is the diagnostic accuracy of tests (including the serum cobalamin assay and holotranscobalamin, methylmalonic acid and homocysteine tests) for diagnosing vitamin B12 deficiency?	• Sensitivity o 90% for first line and 80% for second line tests • Specificity o 70% for first line and 90% for second line tests • Raw data to calculate 2x2 tables to calculate sensitivity and specificity (number of true positives, true negatives, false positives and false negatives) • Predictive values • Likelihood ratios
	Intervention	What are the most clinically and cost-effective ways to diagnose vitamin B12 deficiency, including the serum cobalamin assay and holotranscobalamin, methylmalonic acid and homocysteine tests?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values

Evidence report	Type of review	Review questions	Outcomes
			• complications and adverse events o mortality o bleeds o self-harm o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • patient concern around unexpected lab results (health anxiety score) • incorrect/delayed diagnosis • inappropriate additional tests • adherence to treatment • school/work absence
Evidence review D Identifying cause	Diagnostic	What is the diagnostic accuracy of tests and investigations (including tests for serum intrinsic factor antibody and gastric parietal cell antibody, and gastroscopy and colonoscopy), alone or in combination, for identifying the cause of vitamin B12 deficiency?	• Sensitivity (50%) • Specificity (70%) • Raw data to calculate 2x2 tables to calculate sensitivity and specificity (number of true positives, true negatives, false positives and false negatives) • Predictive values • Likelihood ratios
	Intervention	What is the clinical and cost effectiveness of tests and investigations (including tests for serum intrinsic factor antibody and gastric parietal cell antibody, and gastroscopy and colonoscopy) for identifying the cause of vitamin B12 deficiency?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values • complications and adverse events (condition related): o mortality o self-harm o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • complications and adverse events (procedure related):

Evidence report	Type of review	Review questions	Outcomes
			o bleeding o perforation o aspiration • patient concern around unexpected lab results (health anxiety score) • incorrect/delayed diagnosis • inappropriate additional tests • adherence to treatment • school/work absence
Evidence review E Vitamin B12 replacement and self-administration	Intervention	What is the clinical and cost effectiveness of vitamin B12 replacement for vitamin B12 deficiency, including the dose, frequency and route of administration?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values • complications and adverse events o mortality o bleeds o self-harm o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • adherence to treatment • school/work absence
	Intervention	What is the clinical and cost effectiveness of self-administration compared with healthcare professional administration of parenteral vitamin B12 replacement for vitamin B12 deficiency?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values • complications and adverse events o mortality

Evidence report	Type of review	Review questions	Outcomes
			o bleeds o self-harm o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • adherence to treatment • school/work absence
Evidence review F Follow up	Intervention	What is the optimal frequency of follow-up for people with vitamin B12 deficiency, including pernicious anaemia?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values • complications and adverse events o mortality o bleeds o self-harm o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • adherence to treatment • school/work absence
	Intervention	What should be included in a follow-up review for people with vitamin B12 deficiency, including pernicious anaemia?	• quality of life (such as EQ5D, SF36) • patient-reported outcomes (PROM scores including some/all symptoms): o fatigue o sleep o peripheral neuropathy o cognition o psychiatric symptoms o pain • haematological values • complications and adverse events o mortality o bleeds o self-harm

Evidence report	Type of review	Review questions	Outcomes
			o nerve damage o frailty/falls o severe cognitive effects o postural hypotension • adherence to treatment • school/work absence
Evidence review G Monitoring for gastric cancer	Intervention	What monitoring should be offered to people with pernicious anaemia to identify gastric cancer?	• Mortality • Quality of life • Diagnosis of cancer • Stage of cancer at diagnosis/surgical resectability • Incidence of gastric neuroendocrine tumours (AKA carcinoid tumours/NETS/NENS) • Adverse events (procedure related): o bleeding o perforation o aspiration

2.1.1 Stratification

Stratification is applied where the committee are confident the intervention will work differently in the groups or separate recommendations are required, therefore they should be reviewed separately. For the reviews on information and support needs, analyses were stratified for pregnancy, which meant that different studies with predominantly pregnant and non-pregnant samples were not combined and analysed together. For the review on signs and symptoms, analyses were stratified for age (under 65 years and 65 years or over). For the reviews on diagnostic tests, analyses were stratified for age (under 65 years and 65 years or over), for third trimester of pregnancy (third trimester and first two trimesters or not pregnant), for ethnicity (Afro-Caribbean and all other), and for sex (for homocysteine only). For the review on vitamin B12 replacement, analyses were stratified for cause of deficiency (dietary, non-dietary and drug-induced). For the review on self-administration, analyses were stratified for physical/mental barriers to self-administration (people with barriers and people without barriers). For the reviews on follow up, analyses were stratified for treatment route (oral or intramuscular) and pregnancy/breastfeeding. For the review on monitoring for gastric cancer, analyses were stratified for previous gastric surgery. Where studies reported a mix of populations across strata, a threshold of 70% was agreed with the committee as a cut off for what would be acceptable to constitute a predominant group.

2.2 Searching for evidence

2.2.1 Clinical and health economics literature searches

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

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

Databases were searched using relevant medical subject headings, free-text terms and where appropriate study-type filters. Studies published in languages other than English were not reviewed, and where possible, searches were restricted to English language. Searches were updated between 13 and 16 December 2022. Papers published or added to databases after this date were not considered. Where new evidence was identified, for example in consultation comments received from stakeholders, the impact on the guideline was considered, and the action agreed between the technical team and NICE staff with a quality assurance role.

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

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

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

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

- Emergency Care Research Institute (ECRI) database (www.ecri.org)
- Guidelines International Network (GIN) database (www.g-i-n.net)
- National Institute for Health and Care Excellence (NICE) (www.nice.org.uk)
- Turning Research Into Practice (TRIP) database (www.tripdatabase.com)

2.3 Reviewing evidence

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

- Potentially relevant studies were identified from the search results by reviewing titles and abstracts. The full papers were then obtained.
- Full papers were evaluated against the pre-specified inclusion and exclusion criteria set out in the protocol to identify studies that addressed the review

question. The review protocols are included in an appendix to each of the evidence reports.

- Relevant studies were critically appraised using the preferred study design checklist as specified in the NICE guidelines manual.[4] The checklist used is included in the individual review protocols in each of the evidence reports.
- Key information was extracted about interventional and prognostic study methods and results into EPPI reviewer version 5. Summary evidence tables were produced from data entered into EPPI Reviewer, including critical appraisal ratings. Key information about diagnostic and qualitative study methods and results were manually extracted into standard Word evidence tables (evidence tables are included in an appendix to each of the evidence reports).
- Summaries of the evidence were generated by outcome. Outcome data were combined, analysed and reported according to study design:
  - Randomised data were meta-analysed where appropriate and reported in GRADE evidence profiles.
  - Data from non-randomised studies were meta-analysed where appropriate and reported in GRADE evidence profiles.
  - Prognostic data were meta-analysed where appropriate and reported in adapted GRADE evidence profiles.
  - Diagnostic data were meta-analysed where appropriate or presented as a range of values in GRADE evidence profiles.
  - Qualitative data were synthesised across studies using thematic analysis and presented as summary statements in GRADE CERQual tables.
- A minimum of 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.
- All of the evidence reviews were quality assured by a senior systematic reviewer. This included checking:
  - papers were included or excluded appropriately
  - a sample of the data extractions
  - a sample of the risk of bias assessments
  - correct methods were used to synthesise data.

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

2.3.1 Types of studies and inclusion and exclusion criteria

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

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

2.3.1.1 Type of studies

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

For intervention reviews, randomised controlled trials (RCTs) were included where identified as because they are considered the most robust type of study design that can produce an unbiased estimate of the intervention effects. Non-randomised intervention studies were considered appropriate for inclusion if there was insufficient randomised evidence for the committee to make a decision. In this case the committee stated a priori in the protocol that either certain identified variables must be equivalent at baseline or else the analysis had to adjust for any baseline differences. If the study did not fulfil either criterion it was excluded. Refer to the review protocols in each evidence report for full details on the study design of studies that were appropriate for each review question.

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

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

2.3.1.1.1 Qualitative studies

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

2.4 Methods of combining evidence

2.4.1 Data synthesis for intervention reviews

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

2.4.1.1 Analysis of different types of data

Dichotomous outcomes

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

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

Continuous outcomes

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

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

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

Generic inverse variance

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

2.4.1.2 Network meta-analysis

Network meta-analysis was considered for the comparison of interventional treatments but was not pursued because of differences in the study populations.

2.4.2 Data synthesis for diagnostic reviews

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

2.4.2.1 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 2.4.1.1 above).

2.4.2.2 Diagnostic accuracy studies

For diagnostic test accuracy studies, a positive result on the index test was found if the person had values of the measured quantity above or below a threshold value, and different thresholds could be used. The threshold of a diagnostic test is defined as the value at which the test can best differentiate between those with and without the target condition. In practice this usually varies across studies. If a test has a high sensitivity then very few people with the condition will be missed (few false negatives). For example, a test with a sensitivity of 97% will only miss 3% of people with the condition. Conversely, if a test has a high specificity then few people without the condition would be incorrectly diagnosed (few false positives).

Coupled forest plots of the agreed primary paired outcome measure for decision making (sensitivity and specificity) with their 95% CIs across studies (at various thresholds) were produced for each test, using RevMan5.[9] In order to do this, 2 by 2 tables (the number of true positives, false positives, true negatives and false negatives) were directly taken from the study if given, or else were derived from raw data or calculated from the set of test accuracy statistics.

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

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

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

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

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

Heterogeneity or inconsistency amongst studies was visually inspected.

2.4.3 Data synthesis for prognostic risk factor reviews

Adjusted odds ratios, risk ratios, or hazard ratios, with their 95% CIs, for the effect of the pre-specified prognostic factors were extracted from the studies. Studies were only included if the confounders pre-specified by the committee were either matched at baseline or were adjusted for in multivariate analysis. Prospective cohort studies reporting multivariable analyses that adjusted for key confounders identified by the committee at the protocol stage for that outcome were the preferred study design.

Data were not combined in meta-analyses for prognostic studies unless they had adjusted for the same confounders and were otherwise agreed to be similarly homogenous to pool.

2.4.4 Data synthesis for qualitative reviews

The main findings for each included paper were identified and thematic analysis methods were used to synthesise this information into broad overarching themes which were summarised into the main review findings. The evidence was presented in the form of a narrative summary detailing the evidence from the relevant papers and how this informed the overall review finding plus a statement on the level of confidence for that review finding. Considerable limitations and issues around relevance were listed. A summary evidence table with the succinct summary statements for each review finding was produced including the associated quality assessment.

2.5 Appraising the quality of evidence by outcomes

2.5.1 Intervention reviews

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

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

Table 3: Description of quality elements in GRADE for intervention studies

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

Details of how the 4 main quality elements (risk of bias, indirectness, inconsistency and imprecision) were appraised for each outcome are given below. Publication bias was considered with the committee. If there was reason to suspect it was present, it was explored with funnel plots. Funnel plots were constructed using RevMan5 software to assess against potential publication bias for outcomes containing more than 5 studies. This was taken into consideration when assessing the quality of the evidence.

2.5.1.1 Risk of bias

The main domains of bias for RCTs are listed in Table 4. Each outcome had its risk of bias assessed within each study first using the appropriate checklist for the study design (Cochrane RoB 2 for RCTs, or ROBINS-I for non-randomised studies or ROBIS for systematic reviews). For each study, if there was no risk of bias in any domain, the risk of bias was given a rating of 0; 'no serious risk of bias'. 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. An overall rating is calculated across all studies 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 4: Principle domains of bias in randomised controlled trials

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

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

Table 5 Principle domains of bias in non-randomised studies

Bias	Explanation
Pre-intervention	
Confounding bias	Baseline confounding occurs when one or more prognostic variables (factors that predict the outcome of interest) also predicts the intervention received at baseline. ROBINS-I can also address time-varying confounding, which occurs when post-baseline prognostic factors affect the intervention received after baseline.
Selection bias	When exclusion of some eligible participants, or the initial follow-up time of some participants, or some outcome events, is related to both intervention and outcome, there will be an association between interventions and outcome even

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

2.5.1.2 Indirectness

Indirectness refers to the extent to which the populations, interventions, comparisons and outcome measures are dissimilar to those defined in the inclusion criteria for the reviews. Indirectness is important when these differences are expected to contribute to a difference in effect size, or may affect the balance of harms and benefits considered for an intervention. As for the risk of bias, each outcome had its indirectness assessed within each study first. For each study, if there were no sources of indirectness, indirectness was given a rating of 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. An overall rating is calculated across all studies by taking into account the weighting of studies according to study precision. For example, if the most precise studies tended to have an indirectness score of −1 each for that outcome, the overall score for that outcome would tend towards −1.

2.5.1.3 Inconsistency

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

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

When heterogeneity existed within an outcome (I^2>50%), but no plausible explanation could be found, the quality of evidence for that outcome was downgraded. Inconsistency for that outcome was given a 'serious' 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%) then each of the derived subgroups were presented separately for that forest plot (providing at least 2 studies remained in each subgroup). The committee took this into account and considered whether to make separate recommendations based on the variation in effect across subgroups within the same outcome. In such a situation the quality of evidence was not downgraded.

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

2.5.1.4 Imprecision

The criteria applied for imprecision were based on the 95% CIs for the pooled estimate of effect, and the minimal important differences (MID) for the outcome. The MIDs are the threshold for appreciable benefits and harms, separated by a zone either side of the line of no effect where there is assumed to be no clinically important effect. If either end of the 95% CI of the overall estimate of effect crossed 1 of the MID lines, imprecision was regarded as serious 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 1.

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

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

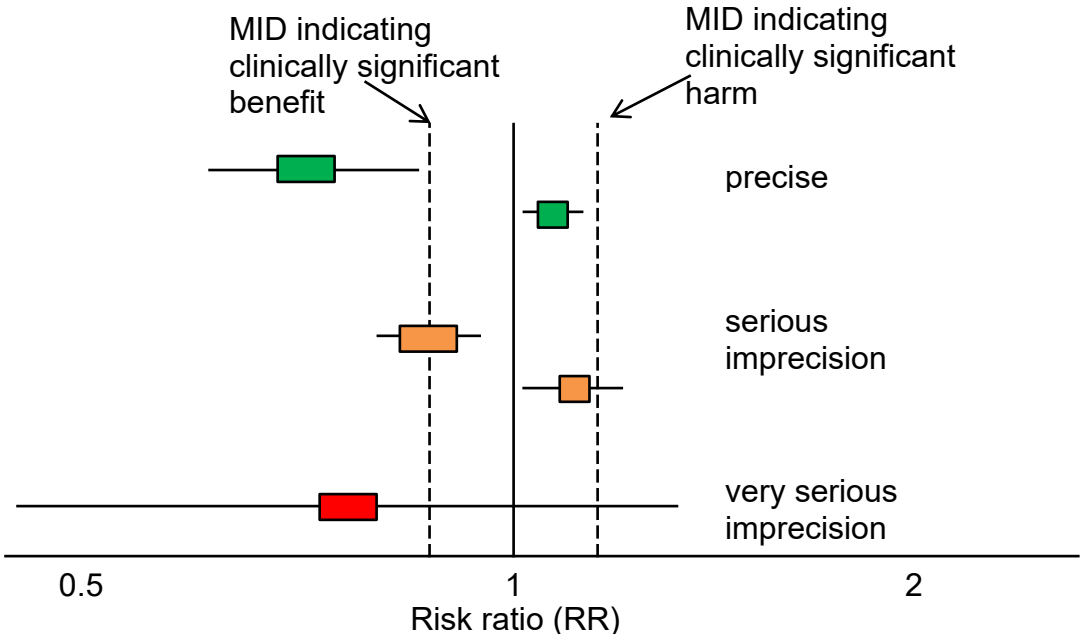
- For dichotomous outcomes the MIDs were taken to be RRs of 0.8* and 1.25. For 'positive' outcomes such as 'patient satisfaction', the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit. For 'negative' outcomes such as 'bleeding', the opposite occurs, so the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm. There aren't established default values for ORs and the same values (0.8 and 1.25) are applied here but are acknowledged as arbitrary thresholds agreed by the committee.
    - In cases where there are zero events in one arm of a single study, or some or all of the studies in one arm of a meta-analysis, the same process is followed as for dichotomous outcomes. However if there are no events in either arm in a meta-analysis (or in a single unpooled study) the sample size is used to determine imprecision using the following rule of thumb:
        - No imprecision: sample size ≥350
        - Serious imprecision: sample size ≥70 but <350
        - Very serious imprecision: sample size <70.
    - When there was more than one study in an analysis and zero events occurred in both groups for some but not all of the studies across both arms, the optimum information size was used to determine imprecision using the following guide:
        - No imprecision: >90% power
        - Serious imprecision: 80-90% power
        - Very serious imprecision: <80% power.
- For mortality any change was considered to be clinically important and the imprecision was assessed on the basis of the whether the confidence intervals crossed the line of no effect, that is whether the result was consistent with both benefit and harm.
- For continuous outcome variables the MID was taken as half the median baseline standard deviation of that variable, across all studies in the meta-analysis. Hence the MID denoting the minimum clinically important benefit was positive for a 'positive' outcome (for example, a quality of life measure where a higher score denotes better health), and negative for a 'negative' outcome (for example, a visual analogue scale [VAS] pain score). Clinically important harms will be the converse of these. If baseline values are unavailable, then half the median comparator group standard deviation of that variable will be taken as the MID. As these vary for each outcome per review, details of the values used are reported in the footnotes of the relevant GRADE summary table.

*NB GRADE report the default values as 0.75 and 1.25. These are consensus values. This guideline follows NICE process to use modified values of 0.8 and 1.25 as they are symmetrical on a relative risk scale. For this guideline, no appropriate MIDs for continuous or dichotomous outcomes were found in the literature, and so the default method was adopted.

Table 6: Published or pre-agreed MIDs

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

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



2.5.1.5 Overall grading of the quality of clinical evidence

Once an outcome had been appraised for the main quality elements, as above, an overall quality grade was calculated for that outcome. 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. RCTs start at High, 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 7. The reasons for downgrading in each case are specified in the footnotes of the GRADE tables.

Table 7: Overall quality of outcome evidence in GRADE

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

2.5.2 Diagnostic reviews

2.5.2.1 Diagnostic RCTs

Appraising the quality of evidence from diagnostic RCTs follows the same process as section 2.5.1 for intervention reviews.

2.5.2.2 Diagnostic test accuracy

2.5.2.2.1 Risk of bias

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

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

Table 8 Summary of QUADAS-2 with list of signalling, risk of bias and applicability questions.

Domain	Patient selection	Index test	Reference standard	Flow and timing
Description	Describe methods of patient selection. Describe included patients (prior testing, presentation, intended use of index test and setting)	Describe the index test and how it was conducted and interpreted	Describe the reference standard and how it was conducted and interpreted	Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2×2 table (refer to flow diagram). Describe the time interval and any interventions between index test(s) and reference standard

Domain	Patient selection	Index test	Reference standard	Flow and timing
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?	

12.5.2.2.2 Inconsistency

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

10.5.2.2.3 Imprecision

Imprecision was assessed according to the range of point estimates or, if only one study contributed to the evidence, the 95% CI around the single study. The decision thresholds set by the committee were used to determine whether imprecision is not serious, serious or very serious depending on whether confidence intervals cross zero, one or two thresholds.

2.5.2.2.4 Overall grading

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

2.5.3 Prognostic reviews

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

2.5.3.1.1 Risk of bias

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

Table 9: Description of risk of bias criteria for prognostic studies

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

2.5.3.1.2 Inconsistency

Inconsistency was assessed as for intervention studies.

2.5.3.1.3 Imprecision

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

2.5.3.1.4 Overall grading

Quality rating was assigned by study. However if there was more than 1 outcome involved in a study, then the quality rating for each outcome was adjusted

accordingly. For example, if one outcome was based on an invalidated measurement method, but another outcome in the same study was not, the second outcome would be graded 1 grade higher than the first outcome.

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

2.5.4 Qualitative reviews

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

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

Table 10: Description of quality elements in GRADE-CERQual for qualitative studies

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

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

2.5.4.1 Methodological limitations

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

Table 11: Description of limitations assessed in the CASP checklist for qualitative studies

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

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

2.5.4.2 Relevance

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

2.5.4.3 Coherence

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

2.5.4.4 Adequacy

The judgement of adequacy is based on the confidence of the finding being supported by sufficient data. This is an overall determination of the richness (and quantity of the evidence supporting a review finding or theme. Rich data provide sufficient detail to gain an understanding of the theme or review finding, whereas thin data do not provide enough detail for an adequate understanding. Quantity of data is the second pillar of the assessment of adequacy. For review findings that are only supported by 1 study or data from only a small number of participants, the confidence that the review finding reasonably represents the phenomenon of interest might be decreased because there is less confidence that studies undertaken in other settings or participants would have reported similar findings. As with richness of data, quantity

of data is review dependent. Based on the overall judgement of adequacy, a rating of no concerns, minor concerns, or substantial concerns about adequacy was given.

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

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

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

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

Table 12: Overall level of confidence for a review finding in GRADE-CERQual

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

2.6 Assessing clinical importance

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

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

negative outcome. For mortality any reduction represented a clinical benefit. For adverse events 50 events or more per 1000 (5%) represented clinical harm.

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

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

2.7 Identifying and analysing evidence of cost effectiveness

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

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

- Undertook a systematic review of the published economic literature.
- Undertook new cost-effectiveness analysis in priority areas.

2.7.1 Literature review

The health economists:

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

2.7.1.1 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 2006 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 13** below and the economic evaluation checklist (appendix H of the NICE guidelines manual[4]) and the health economics review protocol, which can be found in each of the evidence reports.

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

2.7.1.2 NICE health economic evidence profiles

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

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

Table 13: 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)]

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

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

2.7.2 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 diagnosis of vitamin B12 deficiency as the highest priority area for original health economic modelling. In the absence of economic evidence to determine the most cost effective first line test, the use of a second line test was proposed for people who have a first-line test result in an indeterminate range. Elevated concentrations of MMA are considered the most representative marker of metabolic vitamin B12 hence this was selected as the second-line test for people who have an indeterminate first-line test result. By offering a second line test there would be the potential benefit of reduced missed diagnoses of B12 deficiency leading to health gains being achieved more promptly for patients, whilst also reducing inappropriate investigations or referrals. An original model was developed to analyse the cost effectiveness of MMA as a second line/confirmatory test for B12 deficiency for people who have a indeterminate first line test result.

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.[4, 5]
- 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.

Full methods and results of the cost-effectiveness analysis for 'diagnosis – the cost effectiveness of MMA as a second line test' are described in Appendix I of Evidence Report C – Diagnosis.

2.7.3 Cost-effectiveness criteria

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

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more clinically effective compared with all the other relevant alternative strategies), or
- the intervention cost less than £20,000 per QALY gained compared with the next best strategy.

If the committee recommended an intervention that was estimated to cost more than £20,000 per QALY gained, or did not recommend one that was estimated to cost less than £20,000 per QALY gained, the reasons for this decision are discussed explicitly in 'The committee's discussion of the evidence' section of the relevant evidence report, with reference to issues regarding the plausibility of the estimate or to factors set out in NICE methods manuals.[4]

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

2.7.4 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.

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

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

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

The committee considered the appropriate 'strength' of each recommendation. This takes into account the quality of the evidence but is conceptually different. Some recommendations are 'strong' in that the committee believes that the vast majority of healthcare and other professionals and patients would choose a particular intervention if they considered the evidence in the same way that the committee has.

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

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

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

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

2.8.1 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.

2.8.2 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.

2.8.3 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.

3 Acronyms and abbreviations

Acronym or abbreviation	Description
AUC	Area under the curve
BNF	British National Formulary
CI	Confidence interval
FN	False negative
FP	False positive
GRADE	Grading of Recommendations Assessment, Development and Evaluation
Hcy	Homocysteine
HoloTC	Holotranscobalamin
HR	Hazard ratio
HRQoL	Health-related quality of life
IM	Intramuscular
MD	Mean difference
MHRA	Medicines and Healthcare products Regulatory Agency
MID	Minimally important difference
MMA	Methylmalonic acid
NGC	National Guideline Centre
NICE	National Institute for Health and Care Excellence
NPV	Negative predictive value
NR	Not reported
OR	Odds ratio
PPV	Positive predictive value
PROM	Patient-reported outcome measure
QALY	Quality-adjusted life year
RCT	Randomised controlled trial
RR	Risk ratio
TN	True negative
TP	True positive

4 Glossary

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

4.1 Guideline-specific terms

Term	Definition
Active B12	The metabolically active portion of B12, also known as holotranscobalamin (holoTC)
Anti-gastric parietal cell antibody test	A blood test used to detect autoimmune gastritis (also known as pernicious anaemia). Anti-gastric parietal cell antibodies destroy the cells that make a glycoprotein named intrinsic factor which is required for B12 absorption.
Anti-intrinsic factor antibody test	A blood test used to detect autoimmune gastritis (also known as pernicious anaemia). Two types of antibody to intrinsic factor have been detected in the sera of patients with autoimmune gastritis (also known as pernicious anaemia). Type I (blocking) blocks the binding of vitamin B12 to intrinsic factor, whereas type II prevents the attachment of intrinsic factor or the intrinsic factor–B12 complex to ileal receptors.
Atrophic gastritis	A chronic form of gastritis (inflammation of the gastric body).
Autoimmune gastritis	Autoimmune gastritis (AIG) is an autoimmune condition characterised by inflammation in the body of the stomach. The main target of the inflammation is the proton pump in the gastric parietal cells. Because the parietal cells secrete gastric acid and intrinsic factor (which is essential for vitamin B12 absorption), autoimmune gastritis can lead to vitamin B12 deficiency. Usually, the inflammation in the stomach does not cause any symptoms directly. In the early stages of AIG, there is inflammation detectable by histological sampling of the stomach (biopsy), in the later stages there is loss of the specialised cells of the body of the stomach (atrophy). The reduction of vitamin B12 absorption in AIG is complex. In the early stages the reduction in gastric acid reduces the release of vitamin B12 from food. In the later stages the loss of parietal cells and intrinsic factor are more important. In addition, in some cases the intrinsic factor antibodies can directly impair vitamin B12 absorption.
CobaSorb test	A test based on changes in circulating holo-transcobalamin before and after vitamin B12 administration.
Dyspepsia	Dyspepsia describes a range of symptoms arising from the upper gastrointestinal (GI) tract, but it has no universally accepted definition. The British

Term	Definition
	Society of Gastroenterology (BSG) defines dyspepsia as a group of symptoms that alert doctors to consider disease of the upper GI tract, and states that dyspepsia itself is not a diagnosis. These symptoms, which typically are present for 4 weeks or more, include upper abdominal pain or discomfort, heartburn, gastric reflux, nausea or vomiting.
Gastrectomy	The removal of all, or part of, the stomach.
Gastric adenocarcinoma	The most common type of stomach cancer. Adenocarcinoma starts in the glandular cells of the stomach lining.
Gastric neuroendocrine tumours	Rare stomach tumours that start in the neuroendocrine cells of the stomach. They often develop slowly and don't always have specific symptoms.
Gastroscopic endoscopy	A gastroscopy is a test to check inside a person's throat, food pipe (oesophagus) and stomach. This test can help determine what is causing a person's symptoms.
Gastroscopy with gastric body biopsy	A gastroscopy (see 'gastroscopic endoscopy') can also be used to remove tissue for testing (biopsy).
Macrocytosis	When a person's red blood cells are larger than normal. They may have simple macrocytosis or they may go on to have a macrocytic anaemia.
Major gastric resection	Refers to total gastrectomy, partial gastrectomy of various approaches and large segmental resections.
Malabsorption	A state arising from abnormalities in the absorption of food nutrients.
Methylmalonic acid (MMA)	An organic acid of which the blood levels are usually raised where there is a vitamin B12 deficiency.
Pernicious anaemia	Pernicious anaemia (PA) is a variably defined condition, which often has different meanings in different contexts. As originally described pernicious anaemia referred to a severe, progressive, often fatal megaloblastic anaemia frequently associated with neurological problems. It was subsequently discovered that the anaemia and neurological compromise were due to vitamin B12 deficiency, and that autoimmune gastritis was a common cause of this specific syndrome. The term pernicious anaemia has often been used interchangeably with autoimmune gastritis and this has created confusion. Overt vitamin B12 deficiency is only seen in about 20% of cases of AIG. There are many other causes of the vitamin B12 deficiency that can cause the anaemia characteristic of PA, and most people with low B12 levels do not have any anaemia. Even if they do have anaemia and /or neurological disease, with

Term	Definition
	modern management using vitamin B12 replacement, the disease is not “pernicious” (tending to fatal, progressive). In view of these confusions about terminology, the guideline group felt that it was appropriate to use the terms autoimmune gastritis to refer to the specific stomach disease and vitamin B12 deficiency when referring to a biochemical or clinical syndrome.
Plasma homocysteine	An amino acid of which the blood levels are usually raised where there is a vitamin B12 deficiency. Homocysteine can also be raised in folate, vitamin B6 and vitamin B2 deficiency.
Reference range	The range of values containing the central 95% of the healthy population.
Severe megaloblastic anaemia	A type of anaemia of which the most common causes are folate (vitamin B9) deficiency and cobalamin (vitamin B12) deficiency. Megaloblastic anaemia causes macrocytic anaemia from ineffective red blood cell production and intramedullary haemolysis.
Sub-acute combined degeneration of the spinal cord	A neurological complication of vitamin B12 deficiency characterized by degeneration of the dorsal columns and the lateral columns of the spinal cord due to demyelination. It commonly presents with sensory deficits, paresthesia, weakness, ataxia, and gait disturbance. In severe untreated cases, it can lead to spasticity and paraplegia. It is crucial to promptly identify and treat vitamin B12 deficiency to prevent the development of this serious neurological complication.
Total B12	The amount of all cobalamins in human serum (holotranscobalamin and B12-haptocorrins).

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.
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.

Term	Definition
Association	Statistical relationship between 2 or more events, characteristics or other variables. The relationship may or may not be causal.
Base case analysis	In an economic evaluation, this is the main analysis based on the most plausible estimate of each input. In contrast, see Sensitivity analysis.
Baseline	The initial set of measurements at the beginning of a study (after run-in period where applicable), with which subsequent results are compared.
Bayesian analysis	A method of statistics, where a statistic is estimated by combining established information or belief (the 'prior') with new evidence (the 'likelihood') to give a revised estimate (the 'posterior').
Before-and-after study	A study that investigates the effects of an intervention by measuring particular characteristics of a population both before and after taking the intervention, and assessing any change that occurs.
Bias	Influences on a study that can make the results look better or worse than they really are. (Bias can even make it look as if a treatment works when it does not.) Bias can occur by chance, deliberately or as a result of systematic errors in the design and execution of a study. It can also occur at different stages in the research process, for example, during the collection, analysis, interpretation, publication or review of research data. For examples see selection bias, performance bias, information bias, confounding factor, and publication bias.
Blinding	A way to prevent researchers, doctors and patients in a clinical trial from knowing which study group each patient is in so they cannot influence the results. The best way to do this is by sorting patients into study groups randomly. The purpose of 'blinding' or 'masking' is to protect against bias. A single-blinded study is one in which patients do not know which study group they are in (for example whether they are taking the experimental drug or a placebo). A double-blinded study is one in which neither patients nor the researchers and doctors know which study group the patients are in. A triple blind study is one in which 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.

Term	Definition
Clinical efficacy	The extent to which an intervention is active when studied under controlled research conditions.
Clinical effectiveness	How well a specific test or treatment works when used in the 'real world' (for example, when used by a doctor with a patient at home), rather than in a carefully controlled clinical trial. Trials that assess clinical effectiveness are sometimes called management trials. Clinical effectiveness is not the same as efficacy.
Clinician	A healthcare professional who provides patient care. For example, a doctor, nurse or physiotherapist.
Cochrane Review	The Cochrane Library consists of a regularly updated collection of evidence-based medicine databases including the Cochrane Database of Systematic Reviews (reviews of randomised controlled trials prepared by the Cochrane Collaboration).
Cohort study	A study with 2 or more groups of people – cohorts – with similar characteristics. One group receives a treatment, is exposed to a risk factor or has a particular symptom and the other group does not. The study follows their progress over time and records what happens. See also observational study.
Comorbidity	A disease or condition that someone has in addition to the health problem being studied or treated.
Comparability	Similarity of the groups in characteristics likely to affect the study results (such as health status or age).
Concordance	This is a recent term whose meaning has changed. It was initially applied to the consultation process in which doctor and patient agree therapeutic decisions that incorporate their respective views, but now includes patient support in medicine taking as well as prescribing communication. Concordance reflects social values but does not address medicine-taking and may not lead to improved adherence.
Confidence interval (CI)	A range of values for an unknown population parameter with a stated 'confidence' (conventionally 95%) that it contains the true value. The interval is calculated from sample data, and generally straddles the sample estimate. The 'confidence' value means that if the method used to calculate the interval is repeated many times, then that proportion of intervals will actually contain the true value.
Confounding factor	Something that influences a study and can result in misleading findings if it is not understood or appropriately dealt with. For example, a study of heart disease may look at a group of people that exercises regularly and a group that does not exercise. If the ages of the people in the 2 groups are different, then any difference in heart disease rates between the 2 groups could be because of age rather than exercise. Therefore age is a confounding factor.
Consensus methods	Techniques used to reach agreement on a particular issue. Consensus methods may be used to develop NICE guidance if there is not enough good quality research evidence to give a clear answer to a question. Formal consensus methods include Delphi and nominal group techniques.
Control group	A group of people in a study who do not receive the treatment or test being studied. Instead, they may receive the standard treatment (sometimes called 'usual care') or a dummy treatment (placebo). The results for the control group are compared with those for a group receiving the treatment being tested. The aim is to check for any differences.

Term	Definition
	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-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.
Credible interval (CrI)	The Bayesian equivalent of a confidence interval.
Decision analysis	An explicit quantitative approach to decision-making under uncertainty, based on evidence from research. This evidence is translated into probabilities, and then into diagrams or decision trees which direct the clinician through a succession of possible scenarios, actions and outcomes.
Deterministic analysis	In economic evaluation, this is an analysis that uses a point estimate for each input.
Diagnostic odds ratio	The diagnostic odds ratio is a measure of the effectiveness of a diagnostic test. It is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease.
Discounting	Costs and perhaps benefits incurred today have a higher value than costs and benefits occurring in the future. Discounting health benefits reflects individual preference for benefits to be experienced in the present rather than the future. Discounting costs reflects individual preference for costs to be experienced in the future rather than the present.
Disutility	The loss of quality of life associated with having a disease or condition. See Utility
Dominance	A health economics term. When comparing tests or treatments, an option that is both less effective and costs more is said to be ‘dominated’ by the alternative.
Drop-out	A participant who withdraws from a trial before the end.
Economic evaluation	An economic evaluation is used to assess the cost effectiveness of healthcare interventions (that is, to compare the costs and benefits of a healthcare intervention to assess whether it is worth doing). The aim of an economic evaluation is to maximise the level of benefits – health effects – relative to the resources available. It should be used to inform and support the decision-making process; it is not supposed to replace the judgement of healthcare professionals. There are several types of economic evaluation: cost–benefit analysis, cost–consequences analysis, cost-effectiveness analysis, cost-minimisation analysis and cost–utility analysis. They use similar methods to define and evaluate costs, but differ in the way they estimate the benefits of a particular drug, programme or intervention.
Effect (as in effect measure, treatment effect, estimate of effect, effect size)	A measure that shows the magnitude of the outcome in one group compared with that in a control group. For example, if the absolute risk reduction is shown to be 5% and it is the outcome of interest, the effect size is 5%.

Term	Definition
	The effect size is usually tested, using statistics, to find out how likely it is that the effect is a result of the treatment and has not just happened by chance (that is, to see if it is statistically significant).
Effectiveness	How beneficial a test or treatment is under usual or everyday conditions, compared with doing nothing or opting for another type of care.
Efficacy	How beneficial a test, treatment or public health intervention is under ideal conditions (for example, in a laboratory), compared with doing nothing or opting for another type of care.
EQ-5D (EuroQol 5 dimensions)	A standardised instrument used to measure health-related quality of life. It provides a single index value for health status.
Evidence	Information on which a decision or guidance is based. Evidence is obtained from a range of sources including randomised controlled trials, observational studies, expert opinion (of clinical professionals or patients).
Exclusion criteria (literature review)	Explicit standards used to decide which studies should be excluded from consideration as potential sources of evidence.
Exclusion criteria (clinical study)	Criteria that define who is not eligible to participate in a clinical study.
Extended dominance	If Option A is both more clinically effective than Option B and has a lower cost per unit of effect, when both are compared with a do-nothing alternative then Option A is said to have extended dominance over Option B. Option A is therefore cost effective and should be preferred, other things remaining equal.
Extrapolation	An assumption that the results of studies of a specific population will also hold true for another population with similar characteristics.
First line test	The first test that the person receives within the diagnostic pathway.
Follow-up	Observation over a period of time of an individual, group or initially defined population whose appropriate characteristics have been assessed in order to observe changes in health status or health-related variables.
Generalisability	The extent to which the results of a study hold true for groups that did not participate in the research. See also external validity.
Gold standard	A method, procedure or measurement that is widely accepted as being the best available to test for or treat a disease.
GRADE, GRADE evidence profile	A system developed by the GRADE Working Group to address the shortcomings of present grading systems in healthcare. The GRADE system uses a common, sensible and transparent approach to grading the quality of evidence. The results of applying the GRADE system to clinical trial data are displayed in a table known as a GRADE evidence profile.
Harms	Adverse effects of an intervention.
Hazard Ratio	The hazard or chance of an event occurring in the treatment arm of a study as a ratio of the chance of an event occurring in the control arm over time.
Health economics	Study or analysis of the cost of using and distributing healthcare resources.
Health-related quality of life (HRQoL)	A measure of the effects of an illness to see how it affects someone's day-to-day life.
Heterogeneity	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

Term	Definition
or Lack of homogeneity	significantly in different studies. Such differences may occur as a result of differences in the populations studied, the outcome measures used or because of different definitions of the variables involved. It is the opposite of homogeneity.
Imprecision	Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of effect.
Inclusion criteria (literature review)	Explicit criteria used to decide which studies should be considered as potential sources of evidence.
Incremental analysis	The analysis of additional costs and additional clinical outcomes with different interventions.
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.
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.

Term	Definition
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 studies, where allocation to groups occurs through usual treatment decisions or people's preferences. Non-randomised studies can also be experimental, where the investigator has some degree of control over the allocation of treatments. Non-randomised intervention studies can use a number of different study designs, and include cohort studies, case–control studies, controlled before-and-after studies, interrupted-time-series studies and quasi-randomised controlled trials.
Number needed to treat (NNT)	The average number of patients who need to be treated to get a positive outcome. For example, if the NNT is 4, then 4 patients would have to be treated to ensure 1 of them gets better. The closer the NNT is to 1, the better the treatment. For example, if you give a stroke prevention drug to 20 people before 1 stroke is prevented, the number needed to treat is 20. See also number needed to harm, absolute risk reduction.
Observational study	Individuals or groups are observed or certain factors are measured. No attempt is made to affect the outcome. For example, an observational study of a disease or treatment would allow 'nature' or usual medical care to take its course. Changes or differences in one characteristic (for example, whether or not people received a specific treatment or intervention) are studied without intervening. There is a greater risk of selection bias than in experimental studies.
Odds ratio	A measure of treatment effectiveness. The odds of an event happening in the treatment group, expressed as a proportion of the odds of it happening in the control group. The 'odds' is the ratio of events to non-events.
Opportunity cost	The loss of other healthcare programmes displaced by investment in or introduction of another intervention. This may be best measured by the health benefits that could have been achieved had the money been spent on the next best alternative healthcare intervention.
Outcome	The impact that a test, treatment, policy, programme or other intervention has on a person, group or population. Outcomes from interventions to improve the public's health could include changes in knowledge and behaviour related to health, societal changes (for example, a reduction in crime rates) and a change in people's health and wellbeing or health status. In clinical terms, outcomes could include the number of patients who fully recover from an illness or the number of hospital admissions, and an improvement or deterioration in someone's health, functional ability, symptoms or situation. Researchers should decide what outcomes to measure before a study begins.

Term	Definition
P value	The p value is a statistical measure that indicates whether or not an effect is statistically significant. For example, if a study comparing 2 treatments found that one seems more effective than the other, the p value is the probability of obtaining these, or more extreme results by chance. By convention, if the p value is below 0.05 (that is, there is less than a 5% probability that the results occurred by chance) it is considered that there probably is a real difference between treatments. If the p value is 0.001 or less (less than a 1% probability that the results occurred by chance), the result is seen as highly significant. If the p value shows that there is likely to be a difference between treatments, the confidence interval describes how big the difference in effect might be.
Placebo	A fake (or dummy) treatment given to participants in the control group of a clinical trial. It is indistinguishable from the actual treatment (which is given to participants in the experimental group). The aim is to determine what effect the experimental treatment has had – over and above any placebo effect caused because someone has received (or thinks they have received) care or attention.
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.
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.
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.
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

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

Term	Definition
	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.
Second line test	A subsequent diagnostic test that a person may receive depending upon the result of the first line test.
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

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

1

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: 15/08/2022.

2. Maruish ME. User's Manual for the SF-36v2 Health Survey. 3rd ed. Quality Metric Incorporated. 2011. Available from: https://books.google.co.uk/books?id=a0vYnQEACAAJ

3. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016; 75:40-46

4. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual [updated January 2022]. London. National Institute for Health and Care Excellence, 2014. Available from: http://www.nice.org.uk/article/PMG20/chapter/1%20Introduction%20and%20overview

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

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

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

8. Organisation for Economic Co-operation and Development (OECD). Purchasing power parities (PPP). 2021. Available from: https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm Last accessed: 01/06/2022.

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

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