National Institute for Health and Care Excellence

FINAL

Antenatal care

Supplement 1: Methods

NICE guideline NG201

Development of the guideline and methods

August 2021

Final

Developed by the National Guideline Alliance, which is part of the Royal College of Obstetricians and Gynaecologists

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN: 978-1-4731-4227-5

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to develop a guideline on antenatal care.

What this guideline covers

Key areas that are covered

- Organisation and delivery of antenatal care
- Routine antenatal clinical care
- Information and support for pregnant women and their partners
- Interventions for common problems during pregnancy

For further details of what the guideline does and does not cover see the guideline scope on the NICE website.

Methods

Introduction

This section summarises methods used to identify and review the evidence, to consider cost effectiveness, and to develop guideline recommendations. This guideline was developed in accordance with methods described in Developing NICE guidelines: the manual (NICE 2014 – updated 2018).

Declarations of interest were recorded and managed in accordance with NICE's 2018 Policy on declaring and managing interests for NICE advisory committees.

Developing the review questions and outcomes

The review questions considered in this guideline were based on the key areas identified in the guideline scope. They were drafted by the NGA technical team, and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- intervention reviews – using population, intervention, comparison and outcome (PICO)
- diagnostic reviews and reviews of prediction model accuracy – using population, diagnostic test (index test), reference standard and target condition (PIRT)
- prognostic reviews – using population, presence or absence of a prognostic, risk or predictive factor and outcome (PPO)
- qualitative reviews – using population, phenomenon of interest and context (PICo)

These frameworks guided the development of review protocols, the literature searching process, and critical appraisal and synthesis of evidence. They also facilitated development of recommendations by the committee.

Literature searches, critical appraisal and evidence reviews were completed for all review questions, except for review question G on what the content of antenatal appointments should be. For this review question, other evidence reviews, other NICE guidelines or other relevant national guidance (such as NHS screening programmes) was used to inform the recommendations, supplemented by best practice recommendations by the committee based on their knowledge and experience. See report G for more information.

The review questions and evidence reviews corresponding to each question (or group of questions) are summarised in Table 1.

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

Evidence review	Type of review
[A] Information provision	Qualitative
[B] Approaches to information provision	Intervention

Evidence review	Type of review
[C] Involving partners	Qualitative
[D] Peer support	Qualitative
[E] Antenatal classes	Intervention
[F] Accessing antenatal care	Intervention
[G] Content of antenatal appointments	Other[2]
[H] Timing of first antenatal appointment	Intervention
[I] Number of antenatal appointments	Intervention
[J] Referral and delivery of antenatal care	Qualitative
[K] Identification of hypertension in pregnancy	Intervention
[L] Identification of breech presentation	Intervention
[M] Management of breech presentation	Intervention
[N] Risk factors for venous thromboembolism in pregnancy	Prognostic
[O] Monitoring fetal growth	Diagnostic
[P] Fetal movement monitoring[1]	Intervention
[Q] Routine third trimester ultrasound for fetal growth	Intervention
[R] Management of nausea and vomiting in pregnancy	Intervention
[S] Management of heartburn in pregnancy	Intervention
[T] Management of symptomatic vaginal discharge in pregnancy	Intervention
[U] Management of pelvic girdle pain in pregnancy[1]	Intervention
[V] Management of unexplained vaginal bleeding in pregnancy	Intervention
[W] Maternal sleep position during pregnancy	Prognostic

[1] *Original health economic analysis conducted*
[2] *See report G for more information*

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

- Supplement 1 (Methods; this document)
- Supplement 2 (Economics)
- Supplement 3 (NGA staff list).

Searching for evidence

Scoping search

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

Systematic literature search

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

Databases were searched using subject headings, free-text terms and, where appropriate, study type filters. Where possible, searches were limited to retrieve studies published in English. All the searches were conducted in the following databases: Medline, Medline-in-Process, Cochrane Central Register of Controlled Trials (CCTR), Cochrane Database of Systematic Reviews (CDSR) and Embase. For intervention review questions, Database of Abstracts of Reviews of Effects (DARE) and Health Technology Assessments (HTA) were also searched. For qualitative review questions, CINAHL and PsycINFO were also searched. CINAHL was also searched for the management topics [Evidence reviews M, R-W].

Searches were run once for all reviews during development. Searches for evidence reviews E-H and O-W were updated in May 2020. Due to the atypical prolongation of guideline development because of suspension of committee work due to COVID-19, the searches were updated again in September 2020 in advance of the final committee meeting before consultation. During this second update, new evidence was only formally included in reviews where it affected recommendations or the overall conclusions of the evidence report. If new evidence matched inclusion criteria but did not impact on the report, it was referenced in brief in the appendix M of the relevant reports.

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

Economic systematic literature search

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

A single search, using the population search terms used in the evidence reviews, was conducted to identify economic evidence in the NHS Economic Evaluation Database (NHS EED) and HTA. Another single search, using the population search terms used in the evidence reviews combined with an economic evaluations search filter, was conducted in Medline, Medline in Process and Embase. Where possible, searches were limited to studies published in English.

The economic literature searches were updated in July 2020.

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

Quality assurance

Search strategies were quality assured by cross-checking reference lists of relevant studies, analysing search strategies from published systematic reviews and asking members of the committee to highlight key studies.

Reviewing evidence

Systematic review process

The evidence was reviewed in accordance with the following approach.

- Potentially relevant articles were identified from the search results for each review question by screening titles and abstracts. Full-text copies of the articles were then obtained.
- Full-text articles were reviewed against pre-specified inclusion and exclusion criteria in the review protocol (see appendix A of each evidence review).
- Key information was extracted from each article on study methods and results, in accordance with factors specified in the review protocol. The information was presented in a summary table in the corresponding evidence review and in a more detailed evidence table (see appendix D of each evidence review).
- Included studies were critically appraised using an appropriate checklist as specified in Developing NICE guidelines: the manual (NICE 2014). Further detail on appraisal of the evidence is provided below.
- Summaries of evidence by outcome were presented in the corresponding evidence review and discussed by the committee.

Review questions selected as high priorities for economic analysis (and those selected as medium priorities and where economic analysis could influence recommendations) and complex review questions were subject to dual screening and study selection through a 10% random sample of articles. Any discrepancies were resolved by discussion between the first and second reviewers or by reference to a third (senior) reviewer. For the remaining review questions, internal (NGA) quality assurance processes included consideration of the outcomes of screening, study selection and data extraction and the committee reviewed the results of study selection and data extraction. The review protocol for each question specifies whether dual screening and study selection was undertaken for that particular question.

Drafts of all evidence reviews were quality assured by a senior reviewer.

Type of studies and inclusion/exclusion criteria

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

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

For intervention reviews, randomised controlled trials (RCTs) were prioritised for inclusion because they are considered to be the most robust type of study design that could produce an unbiased estimate of intervention effects. Where there was limited evidence from RCTs, non-randomised studies (NRS) were considered for inclusion.

For diagnostic reviews single gate test accuracy studies were considered for inclusion.

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

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

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

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

Methods of combining evidence

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

Data synthesis for intervention reviews

Pairwise meta-analysis

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

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

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

If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into RevMan5. If the control event rate was reported this was used to generate the absolute risk difference in GRADEpro. If

multivariable analysis was used to derive the summary statistic but no adjusted control event rate was reported, no absolute risk difference was calculated.

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

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

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

Handling of cluster randomised trials

Where cluster randomised trials were included in evidence reviews they were analysed to minimise the potential for unit-of-analysis error. If studies reported contrast level outcomes (for example risk ratios, mean differences) that appeared to have been calculated taking into account the cluster study design, these were preferentially extracted over raw data (for example counts of events in each arm or mean and standard deviation of each arm). However if raw data was used, a design effect adjustment was made (Higgins 2020) using an appropriate estimate of the intracluster correlation coefficient, details on the calculation are provided in the relevant evidence reviews.

Data synthesis for reviews of diagnostic test accuracy and prediction tools

When diagnostic test accuracy was measured dichotomously, sensitivity, specificity, positive and negative predictive values were used as outcomes. These diagnostic test accuracy parameters were obtained directly from results reported in the source articles or calculated by the NGA technical team using data reported in the articles.

Meta-analysis of diagnostic test accuracy parameters was conducted if there was data from two or more studies that could be pooled using WinBUGS.

Data synthesis for prognostic reviews

ORs, RRs or hazard ratios (HRs) with 95% CIs reported in published studies were extracted or calculated by the NGA technical team to examine relationships between

risk factors and outcomes of interest. Recognising variation across studies in terms of populations, risk factors, outcomes and statistical analysis methods (including adjustments for confounding factors), prognostic data were not meta-analysed, but results from individual studies were presented in the evidence reviews.

Data synthesis for qualitative reviews

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

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

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

Appraising the quality of evidence

Intervention studies

Pairwise meta-analysis

GRADE methodology for intervention reviews

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

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

The selection of outcomes for each review question was agreed during development of the associated review protocol in discussion with the committee. The evidence for each outcome was examined separately for the quality elements summarised in Table 2. Criteria considered in the rating of these elements are discussed below. Each element was graded using the quality ratings summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a ‘serious’ or ‘very serious’ quality issue. The ratings for each

component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

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

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

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

Table 3: GRADE quality ratings (by quality element)

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

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

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

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

Assessing risk of bias in intervention reviews

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

Risk of bias in RCTs was assessed using the Cochrane risk of bias v2 tool (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

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

For systematic reviews the ROBIS checklist was used (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

For non-randomised studies the ROBINS-I checklist was used (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

Assessing inconsistency in intervention reviews

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

Inconsistency was assessed visually by inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis (for example if the point estimates of the individual studies consistently showed benefits or harms). This was supported by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating serious heterogeneity, and more than 80% indicating very serious heterogeneity. When serious or very serious heterogeneity was observed, possible reasons were explored and subgroup analyses were performed as pre-specified in the review protocol where possible.

When considerable heterogeneity was present, the meta-analysis was re-run using the Der-Simonian and Laird method with a random effects model and this was used for the final analysis.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency.

Assessing indirectness in intervention reviews

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

Assessing imprecision and importance in intervention reviews

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

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

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

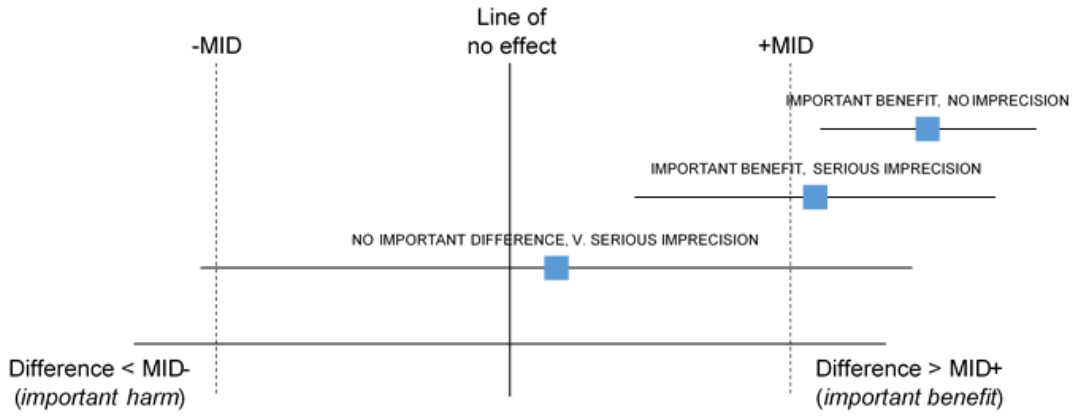
When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

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

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

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

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



MID, minimally important difference

Defining minimally important differences for intervention reviews

The committee was asked whether there were any recognised or acceptable MIDs in the published literature and community relevant to the review questions under consideration. The committee was not aware of any MIDs that could be used for the guideline.

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

If risk difference was used for meta-analysis, for example if the majority of studies had zero events in either arm, imprecision was assessed based on sample size using 200 and 400 as cut-offs for very serious and serious imprecision respectively. The committee used these numbers based on commonly used optimal information size thresholds.

The same thresholds were used as default MIDs in the guideline for all dichotomous outcomes considered in intervention evidence reviews except for those related to mortality (maternal or child). For these mortality outcomes, any statistically significant difference was considered to be important (and p values were quoted in the corresponding evidence statements), although the default MIDs were still used to guide precision ratings.

For continuous outcomes default MIDs are equal to half the median SD of the control groups at baseline (or at follow-up if the SD is not available a baseline).

In this guideline by default a finding was considered important when the point estimate lay outside the MID boundaries and the 95% CI did not cross the line of no effect.

Assessing publication bias in intervention reviews

Where 10 or more studies were included as part of a single meta-analysis, a funnel plot was produced to graphically assess the potential for publication bias. Where fewer than 10 studies were included for an outcome, the committee subjectively assessed the likelihood of publication bias based on factors such as the proportion of trials funded by industry and the propensity for publication bias in the topic area.

Diagnostic reviews

Adapted GRADE methodology for diagnostic reviews

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

The evidence for each outcome in the diagnostic reviews was examined separately for the quality elements listed and defined in Table 5. The criteria considered in the rating of these elements are discussed below. Each element was graded using the quality levels summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a 'serious' or 'very serious' quality issue. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

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

Table 5: Adaptation of GRADE quality elements for diagnostic reviews

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

Assessing risk of bias in diagnostic reviews

Risk of bias in diagnostic reviews was assessed using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklist (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

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

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

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

Assessing inconsistency in diagnostic reviews

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

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

Assessing indirectness in diagnostic reviews

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

- participant selection
- index test
- reference standard.

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

Assessing imprecision and importance in diagnostic reviews

The judgement of precision for diagnostic evidence was based on the CIs of the sensitivity and specificity. The committee defined 2 decision thresholds for each measure, a value above which the test could be recommended and a value below which the test would be considered of no use. These thresholds were based on the committee's experience and consensus and defined in the relevant evidence review protocol.

Outcomes were downgraded for imprecision when their 95% CI crossed at least 1 threshold. If the CI crossed 1 threshold, the outcome was downgraded once for imprecision. If the CI crossed 2 thresholds, the outcome was downgraded twice for

imprecision. These assessments were made on the meta-analysed outcomes where applicable or if outcomes were not meta-analysed, on the individual study results themselves.

Prognostic studies

Adapted GRADE methodology for prognostic reviews

For prognostic reviews with evidence from comparative studies an adapted GRADE approach was used. As noted above, GRADE methodology is designed for intervention reviews but the quality assessment elements were adapted for prognostic reviews.

The evidence for each outcome in the prognostic reviews was examined separately for the quality elements listed and defined in Table 6. The criteria considered in the rating of these elements are discussed below. Each element was graded using the quality levels summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having 'serious' or 'very serious' quality issues. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

The phase of prognostic study design was used, alongside the specific context of the review, to guide the starting quality in GRADE as per Huguet 2013. For the review on sleep position due to the relative rarity of the critical outcome (stillbirth) case control studies were still considered to justify a starting high quality rating. For the review on venous thromboembolism, all evidence was from studies aiming to test independent associations between potential risk factors and the outcome of venous thromboembolism and therefore started at a high quality rating.

Table 6: Adaptation of GRADE quality elements for prognostic reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in study design and implementation may bias estimates and interpretation of the effect of the prognostic/risk factor. High risk of bias for the majority of the evidence reduces confidence in the estimated effect
Inconsistency	This refers to unexplained heterogeneity between studies looking at the same prognostic/risk factor, resulting in wide variability in estimates of association (such as RRs or ORs), with little or no overlap in confidence intervals
Indirectness	This refers to any departure from inclusion criteria listed in the review protocol (such as differences in study populations or prognostic/risk factors), that may affect the generalisability of results
Imprecision	This occurs when a study has relatively few participants and also when the number of participants is too small for a multivariable analysis (as a rule of thumb, 10 participants are needed per variable). This was assessed by considering the confidence interval in relation to the point estimate for each outcome reported in the included studies

Quality element	Description
Publication bias	This refers to systematic under- or over-estimation of the association between a risk factor and an outcome resulting from selective publication of study results

RR, relative risk; OR, odds ratio

Assessing risk of bias in prognostic reviews

The Quality in Prognosis Studies (QUIPS) tool developed by Hayden 2013 was used to assess risk of bias in studies included in prognostic reviews (see Appendix H in the Developing NICE guidelines: the manual; NICE 2014). The risk of bias in each study was determined by assessing the following domains:

- selection bias
- attrition bias
- prognostic factor bias
- outcome measurement bias
- control for confounders
- appropriate statistical analysis.

Assessing inconsistency in prognostic reviews

Where multiple results were deemed appropriate to meta-analyse (that is, there was sufficient similarity between risk factor and outcome under investigation) inconsistency was assessed by visually inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis. This was assessed by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating considerable heterogeneity, and more than 80% indicating very serious heterogeneity. When considerable or very serious heterogeneity was observed, possible reasons were explored and subgroup analyses were performed as pre-specified in the review protocol where possible. If meta-analysis was not appropriate, inconsistency was assessed by comparing the point estimates and confidence intervals of studies reporting on similar risk factor/outcome pairings.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency.

Assessing indirectness in prognostic reviews

Indirectness in prognostic reviews was assessed by comparing the populations, prognostic factors and outcomes in the evidence to those defined in the review protocol.

Assessing imprecision and importance in prognostic reviews

Prognostic studies may have a variety of purposes, for example, establishing typical prognosis in a broad population, establishing the effect of patient characteristics on prognosis, and developing a prognostic model. While by convention MIDs relate to intervention effects, the committee agreed to use GRADE default MIDs for RRs as a starting point from which to assess whether the size of an outcome effect in a

prognostic study would be large enough to be meaningful in practice. As for the intervention reviews in this guideline, where prognostic evidence related to mortality outcomes (for example stillbirth), statistical significance was used as the delineator of clinical importance and whether the evidence was seriously imprecise. Evidence of this category could be downgraded a second time if the reviewer subjectively considered that the 95% CI were 'very wide' in addition to crossing the line of no effect.

Assessing publication bias in prognostic reviews

As per Huguet 2013, publication bias was considered to be present in reviews of prognostic evidence unless the association had been repeatedly investigated in phase 2 (cohort study design that seeks to confirm independent associations between the prognostic factor and the outcome) or phase 3 (cohort study design that seeks to generate understanding of the underlying processes for the prognosis of a health condition) prognostic studies.

Qualitative reviews

GRADE-CERQual methodology for qualitative reviews

For qualitative reviews an adapted GRADE Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) approach (Lewin 2015) was used. In this approach the quality of evidence is considered for each theme in the evidence. The themes may have been identified in the primary studies or they may have been identified by considering the reports of a number of studies. Quality elements assessed using GRADE-CERQual are listed and defined in Table 7. Each element was graded using the levels of concern summarised in Table 8.

The ratings for each component were combined to obtain an overall assessment of confidence in each review finding or 'theme' as described in Table 9. 'Confidence' in this context refers to the extent to which the review finding is a reasonable representation of the phenomenon of interest set out in the protocol. Similar to other types of evidence all review findings start off with 'high confidence' and are rated down by one or more levels if there are concerns about any of the individual CERQual components. In line with advice from the CERQual developers, the overall assessment does not involve numerical scoring for each component but in order to ensure consistency across and between guidelines, the NGA established some guiding principles for overall ratings. For example, a review finding would not be downgraded (and therefore would be assessed with 'high' confidence) if all 4 components had 'no or very minor' concerns or 3 'no or very minor' and 1 'minor'. At the other extreme, a review finding would be downgraded 3 times (to 'very low') if at least 2 components had serious concerns or at least 3 had moderate concerns. A basic principle was that if any components had serious concerns then overall confidence in the review finding would be downgraded at least once (potentially more depending on the other ratings). Transparency about overall judgements is provided in the CERQual tables, including a brief reference to components for which there were concerns in the 'overall confidence' cell.

Table 7: Adaptation of GRADE quality elements for qualitative reviews

Quality element	Description
Risk of bias ('Methodological limitations')	Limitations in study design and implementation may bias interpretation of qualitative themes identified. High risk of bias for the majority of the evidence reduces confidence in review findings. Qualitative studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Relevance (or applicability) of evidence	This refers to the extent to which the evidence supporting the review findings is applicable to the context specified in the review question
Coherence of findings	This refers to the extent to which review findings are well grounded in data from the contributing primary studies and provide a credible explanation for patterns identified in the evidence
Adequacy of data (theme saturation or sufficiency)	This corresponds to a similar concept in primary qualitative research, that is, whether a theoretical point of theme saturation was achieved, at which point no further citations or observations would provide more insight or suggest a different interpretation of the particular theme. Individual studies that may have contributed to a theme or sub-theme may have been conducted in a manner that by design would have not reached theoretical saturation at an individual study level

Table 8: CERQual levels of concern (by quality element)

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

Table 9: Overall confidence in the evidence in CERQual (by review finding)

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

Assessing methodological limitations in qualitative reviews

Methodological limitations in qualitative studies were assessed using the Critical Appraisal Skills Programme (CASP) checklist for qualitative studies (see appendix H in Developing NICE guidelines: the manual; NICE 2014). Overall methodological limitations were derived by assessing the methodological limitations across the 6 domains summarised in Table 10.

Table 10: Methodological limitations in qualitative studies

Aim and appropriateness of qualitative evidence	This domain assesses whether the aims and relevance of the study were described clearly and whether qualitative research methods were appropriate for investigating the research question
Rigour in study design or validity of theoretical approach	This domain assesses whether the study approach was documented clearly and whether it was based on a theoretical framework (such as ethnography or grounded theory). This does not necessarily mean that the framework has to be stated explicitly, but a detailed description ensuring transparency and reproducibility should be provided
Sample selection	This domain assesses the background, the procedure and reasons for the method of selecting participants. The assessment should include consideration of any relationship between the researcher and the participants, and how this might have influenced the findings
Data collection	This domain assesses the documentation of the method of data collection (in-depth interviews, semi-structured interviews, focus groups or observations). It also assesses who conducted any interviews, how long they lasted and where they took place
Data analysis	This domain assesses whether sufficient detail was documented for the analytical process and whether it was in accordance with the theoretical approach. For example, if a thematic analysis was used, the assessment would focus on the description of the approach used to generate themes. Consideration of data saturation would also form part of this assessment (it could be reported directly or it might be inferred from the citations documented that more themes could be found)
Results	This domain assesses any reasoning accompanying reporting of results (for

	example, whether a theoretical proposal or framework is provided)

Assessing relevance of evidence in qualitative reviews

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

Assessing coherence of findings in qualitative reviews

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

Assessing adequacy of data in qualitative reviews

Adequacy of data (theme saturation or sufficiency) corresponds to a similar concept in primary qualitative research in which consideration is made of whether a theoretical point of theme saturation was achieved, meaning that no further citations or observations would provide more insight or suggest a different interpretation of the theme concerned. Adequacy rating was assessed subjectively and informed by a combination of the number of studies, the number of participants and the richness of the evidence from each study/participant (for example the amount of supporting quotes or depth of observation).

Assessing importance in qualitative reviews

For themes stemming from qualitative findings, importance was agreed by the committee taking account of the generalisability of the context from which the theme was derived and whether it was sufficiently convincing to support or warrant a change in current practice, as well as the quality of the evidence.

Reviewing economic evidence

Inclusion and exclusion of economic studies

Titles and abstracts of articles identified through the economic literature searches were independently assessed for inclusion using the predefined eligibility criteria listed in Table

Table 11: Inclusion and exclusion criteria for systematic reviews of economic evaluations

Inclusion criteria
Economic evaluations that compare costs and health consequences of interventions (i.e. true cost-effectiveness analyses)
Population, interventions, comparators and outcomes match those specified in the PICO
Quality of life based outcomes were used as the measure of effectiveness in at least one of the analyses presented
Incremental results reported or enough information for incremental results to be derived
Conducted from the perspective of a healthcare system in an OECD country
Exclusion criteria
Conference abstracts, poster presentations or dissertation abstracts with insufficient methodological details for quality assessment
Non-English language papers

OECD: Organisation for Economic Co-operation and Development; PICO: Population, Intervention, Comparison, and Outcome

Once the screening of titles and abstracts was completed, full-text copies of potentially relevant articles were requested for detailed assessment. Inclusion and exclusion criteria were applied to articles obtained as full-text copies.

Details of economic evidence study selection and lists of excluded studies are presented in Supplement 2: Health Economics. Economic evidence tables and health economic evidence profiles are presented in appendix H and appendix I of the relevant evidence reviews respectively.

Appraising the quality of economic evidence

The quality of economic evidence was assessed using the economic evaluations checklist specified in Developing NICE guidelines: the manual (NICE 2014). See Supplement 2: Health Economics for further details.

Economic modelling

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

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

- Is fetal movement monitoring from 28 weeks effective? (Evidence review P)

- What interventions are effective in treating pelvic girdle pain during pregnancy? (Evidence review U)

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

Cost effectiveness criteria

Developing NICE guidelines: the manual (NICE 2014) sets out the principles that committees should consider when judging whether an intervention offers good value for money. In general, an intervention was considered to be cost effective if any of the following criteria applied (provided that the estimate was considered plausible):

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

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

Details of the cost effectiveness analyses undertaken for the guideline are presented in appendix J of the relevant evidence report.

Developing recommendations

Guideline recommendations

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

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

For further details refer to Developing NICE guidelines: the manual (NICE 2014).

Research recommendations

When areas were identified for which evidence was lacking, the committee considered making recommendations for future research. For further details refer to Developing NICE guidelines: the manual (NICE 2014) and the Research Recommendations Process and Methods guide.

Validation process

This guideline was subject to a 6-week public consultation and feedback process. All comments received from registered stakeholders were responded to in writing and posted on the NICE website at publication. For further details refer to Developing NICE guidelines: the manual (NICE 2014).

Updating the guideline

Following publication, NICE will undertake a surveillance review to determine whether the evidence base has progressed sufficiently to consider altering the guideline recommendations and warrant an update. For further details refer to Developing NICE guidelines: the manual (NICE 2014).

Funding

The NGA was commissioned by NICE to develop this guideline.

References

Bradburn 2007

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

Hayden 2013

Jill A. Hayden, Danielle A. van der Windt, Jennifer L. Cartwright, Pierre Côté, Claire Bombardier. Assessing Bias in Studies of Prognostic Factors. Ann Intern Med. 2013;158:280–286. doi: 10.7326/0003-4819-158-4-201302190-00009

Higgins 2020

Higgins JPT, Eldridge S, Li T (editors). Chapter 23: Including variants on randomized trials. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook. (accessed 1 February 2021)

Huguet 2013

Huguet, A., Hayden, J. A., Stinson, J., McGrath, P. J., Chambers, C. T., Tougas, M. E., & Wozney, L. (2013). Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework. Systematic reviews, 2(1), 71.

Lewin 2015

Lewin S, Glenton C, Munthe-Kaas H et al. (2015) Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 12(10), e1001895

McGowan 2016

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

NICE 2014

National Institute for Health and Care Excellence (NICE) (2014) Developing NICE guidelines: the manual (updated 2018). Available from https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview (accessed 1 February 2021)

NICE 2018

National Institute for Health and Care Excellence (NICE) (2014) NICE Policy on conflicts of interest (updated 2017). Available from

https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/declaration-of-interests-policy.pdf (accessed 1 February 2021)

Santesso 2016

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