National Institute for Health and Care Excellence

Final

Twin and triplet pregnancy

Supplement C: methods

NICE guideline NG137

Supplement C

September 2019

Final

Evidence reviews were developed by the National Guideline Alliance, which is a 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 All NICE guidance is subject to regular review and may be updated or withdrawn.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: 978-1-4731-3513-0

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to partially update the NICE guideline on Multiple pregnancy: antenatal care for twin and triplet pregnancies (CG129) with a modified scope.

What this guideline covers

Groups that are covered

- All women confirmed as having a twin or triplet pregnancy by the 11–13-week ultrasound scan.

Clinical areas that are covered

The guideline update covers the following clinical issues from the published guideline:

- Fetal complications
  - screening to identify feto-fetal transfusion syndrome (FFTS)
  - screening to detect intrauterine growth restriction
- Preterm birth
  - predicting the risk of preterm birth
  - preventing preterm birth
- Timing of birth

The guideline update covers the following clinical issues that are not in the published guideline

- Fetal complications
  - screening to detect twin anaemia polycythemia sequence (TAPS)
- Intrapartum care
  - mode of birth
  - fetal monitoring during labour
  - analgesia
  - management of third stage of labour.

For further details please refer to the guideline update scope on the NICE website.

What this guideline does not cover

Groups that are not covered

The guideline does not cover the following groups:

- Women with a quadruplet or higher-order pregnancy.

Clinical areas that are not covered

This guideline does not cover the following area:

- Management of fetal complications.

Methods

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

Until April 2018, declarations of interest were recorded according to NICE's 2014 conflicts of interest policy. From April 2018 onwards declarations were recorded according and managed in accordance with NICE's 2018 Policy on declaring and 9 managing interests for NICE advisory committees.

Developing the review questions and outcomes

The 10 review questions developed for this guideline were based on the key areas identified in the guideline update scope. They were drafted by the NGA technical team and refined and validated by the committee. They cover all areas of the update scope and were signed-off by NICE (see Table 1).

The review questions were based on the following frameworks:

- intervention reviews: population, intervention, comparator and outcome (PICO)
- diagnostic test accuracy reviews: population, index test, reference standard and outcome (PIRO)
- prognostic reviews: population, presence or absence of a prognostic or predictive factor and outcome (PPO)
- incidence reviews: population, exposure and outcome

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

Full literature searches, critical appraisals and evidence reviews were completed for all review questions.

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

Table 1: Description of review questions

Chapter or section	Type of review	Review question guideline	Outcomes
A1 Screening for feto-fetal transfusion syndrome (FFTS)	Prognosis and diagnosis	1.1 What is the optimal screening programme to identify FFTS in twin and triplet pregnancy?	• Prognostic value of first trimester ultrasound tests to predict FFTS ○ Adjusted odds ratios, hazard ratios, risk ratios • Diagnostic value of first and second trimester ultrasound tests to detect FFTS **Critical** ○ Sensitivity (detection rate) and specificity

Chapter or section	Type of review	Review question guideline	Outcomes
			Important ○ Area under the receiver-operating characteristic (ROC) curve (AUC)
A2 Screening for intrauterine growth restriction (IUGR)	Diagnosis	1.2 What is the optimal screening programme to detect IUGR in twin and triplet pregnancy?	• Diagnostic value of first and second trimester ultrasound tests to detect IUGR **Critical** ○ Sensitivity (detection rate) and specificity **Important** ○ AUC
A3 Screening for twin anaemia polycythemia sequences (TAPS)	Diagnosis	1.3 What is the optimal screening programme to detect twin anaemia polycythemia sequences (TAPS)?	• Diagnostic value of second trimester ultrasound tests to detect TAPS **Critical** ○ Sensitivity (detection rate) and specificity **Important** ○ AUC
B1 Screening for spontaneous preterm birth	Diagnostic prediction and diagnostic accuracy	2.1 What is the optimal screening programme to predict the risk of spontaneous preterm birth?	• Diagnostic predictive value of screening methods to predict spontaneous preterm birth ○ Adjusted odds ratios, hazard ratios, risk ratios • Diagnostic accuracy of screening methods to detect spontaneous preterm birth **Critical** ○ Sensitivity (detection rate) and specificity
B2 Interventions to prevent spontaneous preterm birth	Intervention	2.2 What interventions are effective in preventing spontaneous preterm birth in twin and triplet pregnancy, including bed rest, progesterone and cervical cerclage?	**Critical** • Maternal: ○ Mortality • Neonatal: ○ Gestational age at birth ○ Perinatal mortality **Important** • Maternal: ○ Woman's satisfaction ○ Adverse effects • Neonatal: ○ Perinatal morbidity
C1 Mode of birth	Intervention	3.1 What is the optimal mode of birth to improve outcomes for mothers and babies?	**Critical** • Maternal: ○ Mortality • Neonatal:

Chapter or section	Type of review	Review question guideline	Outcomes
			○ Perinatal or neonatal mortality ○ Disability in childhood **Important** • Maternal: ○ Serious maternal morbidity ○ Actual mode of birth • Neonatal: ○ Serious neonatal morbidity
C2 Fetal monitoring	Intervention	3.2 What is the most effective method of fetal monitoring during labour in improving outcomes for babies and mothers?	**Critical** • Maternal: ○ Mode of birth • Neonatal: ○ Perinatal mortality (either or both twins) ○ Hypoxic-ischaemic encephalopathy Grade 2 and 3 **Important** • Maternal: ○ Infection ○ Maternal satisfaction • Neonatal: ○ Fetal acidosis/acidaemia ○ Admission to NICU
C3 Analgesia during labour	Intervention	3.3 What is the optimal method of analgesia during labour and birth?	**Critical** • Maternal: ○ Pain ○ Conversion to general anaesthesia for any operative intervention • Neonatal: ○ Major neonatal morbidities **Important** • Maternal: ○ Mode of birth ○ Women's satisfaction/experience of labour and birth ○ Mortality • Neonatal: ○ Mortality
C4 Prevention of postpartum haemorrhage (PPH)	Intervention	3.4 What is the optimal method of managing the third stage of labour to reduce the risk of PPH?	**Critical** • Maternal: ○ Mortality ○ Postpartum haemorrhage ○ Hysterectomy

Chapter or section	Type of review	Review question guideline	Outcomes
			Important • Maternal: ○ Side effects of drugs ○ Need for further intervention ○ Need for ITU or HDU ○ Women's satisfaction/experience of labour and birth
D1 Timing of birth	Incidence	What is the incidence of stillbirth and neonatal death and morbidity by gestational age in twin and triplet pregnancies according to chorionicity and amnionicity?	For the baby: **Critical** • Stillbirth • Perinatal/neonatal mortality **Important**: • Neonatal morbidities – defined as any of the following: ○ respiratory distress syndrome ○ need for respiratory support (respiratory ventilation) ○ septicaemia or meningitis ○ bronchopulmonary dysplasia ○ hypoxic ischaemic encephalopathy ○ necrotising enterocolitis ○ intraventricular haemorrhage ○ cystic periventricular leukomalacia ○ retinopathy of prematurity ○ admission to NICU

AUC: area under the curve; FFTS feto-fetal transfusion; HDU high dependency unit; ITU intensive therapy unit; IUGR intrauterine growth restriction; NICU: neonatal intensive care unit; PPH postpartum haemorrhage; ROC: receiver operating characteristic; TAPS twin anaemia polycythemia sequence

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

- Supplement A (NGA team list)
- Supplement B (Glossary and abbreviations)
- Supplement C (Methods; this document)
- Supplement D (Health economics)
-

Searching for evidence

Clinical literature search

Systematic literature searches were undertaken to identify published clinical evidence relevant to the review questions.

Databases were searched using relevant medical subject headings, free-text terms and study type filters where appropriate. Studies published in languages other than English were not reviewed. All searches were conducted in the following databases: MEDLINE, Embase and The Cochrane Library.

Any studies added to the databases after the date of the last search (even those published prior to this date) were not included unless specifically stated in the text.

Search strategies were quality assured by cross-checking reference lists of relevant papers, analysing search strategies in other systematic reviews and asking committee members to highlight key studies. Details of the search strategies, including study-design filters applied and databases searched, are presented in appendix B of each evidence review report.

All publications highlighted by stakeholders at the time of the consultation on the draft scope were considered for inclusion. During the scoping phase, searches were conducted for guidelines, health technology assessments, systematic reviews, economic evaluations, and reports on websites of organisations relevant to the topic. Formal searching for grey literature and unpublished literature was not undertaken routinely.

Economic literature search

Systematic literature searches were also undertaken to identify published health economic evidence relevant to each review questions.

The following databases were searched:

- MEDLINE (Ovid)
- EMBASE (Ovid)
- Health Technology Assessment database (HTA)
- NHS Economic Evaluations Database (NHS EED).

The search strategies for existing economic evaluations combined terms capturing the target population (women with a twin or triplet pregnancy) and, for searches undertaken in MEDLINE and EMBASE, terms to capture economic evaluations. No restrictions on language or setting were applied to any of the searches, but a standard exclusions filter was applied (letters, animals, etc.). For full details of the search strategies see appendix B in each evidence report.

Update of clinical and economics literature searches

The clinical and economic literature searches for 9 reviews (A1, A2, A3, B1, B2, C1, C2, C3, C4) were carried out in 2 stages, with initial rerun searches conducted on 19th July 2018 and final top-up searches carried out on 6th September (A1, A2, A3, B1, B2) and 11th September 2018 (C1, C2, C3 and C4).

For 1 review (D) the initial clinical and economic literature searches were conducted on 6th November 2018 and so no re-run was considered necessary.

Reviewing clinical evidence

Systematic review process

The evidence was reviewed following these steps.

- Potentially relevant articles were identified from the search results for each review question by reviewing 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 protocols (see appendix A of each evidence review chapter).
- Key information was extracted on the study's methods, according to the factors specified in the protocols and results. These were presented in summary tables (in each review chapter) and evidence tables (in appendix D of each evidence report).
- Relevant studies were critically appraised using the appropriate checklist as specified in Developing NICE guidelines: the manual 2014.
- Summaries of evidence by outcome were presented in the corresponding evidence report and discussed and presented in committee meetings.
- Results were summarised and reported in GRADE profiles (for intervention reviews), their equivalent (adapted GRADE profiles for diagnostic test accuracy [refer to estimate. Diagnostic test accuracy reviews below]) or reported in evidence tables including the risk of bias assessment for prognostic reviews.
- Model performance studies: data were presented individually by study.

Drafts of all evidence reviews were checked 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 (refer to appendix A of each evidence report). Excluded studies and the reasons for their exclusion are listed in appendix K of each evidence report.

Systematic reviews (SRs) with meta-analyses of randomised controlled trials (for intervention reviews) and cross-sectional studies (for diagnostic reviews) were considered the highest quality evidence to 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 that could produce an unbiased estimate of intervention effects. Based on their judgement, if the committee believed RCT data were not appropriate or there was limited evidence from RCTs, non-randomised controlled trials and/or observational studies were considered for inclusion, including cohort studies.

For diagnostic reviews, test-and-treat RCTs where patients undergo either a new test, or an existing test, measure the downstream health response after patients have received subsequent treatment were prioritised for inclusion. In the absence of test-and-treat RCTs, diagnostic test accuracy studies comparing a diagnostic test of interest (the 'index test') to an existing diagnostic test (the 'reference test') (cross-sectional studies and prospective or retrospective cohort studies) were considered for inclusion.

For prognostic reviews, SRs/meta-analyses of cohort studies were prioritised for inclusion. In the absence of such studies, prospective population-based cohort studies and prospective multicentre cohort studies were considered for inclusion.

For incidence reviews, prospective and/or retrospective cohort studies were prioritised for inclusion. 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 listed in appendix H of the corresponding evidence report.

Posters, letters, editorials, comment articles, unpublished studies and studies published in languages other than English were excluded. Narrative reviews were also excluded, but individual references were checked for inclusion. Conference abstracts were generally not considered for inclusion except for review questions where no other evidence was available and if published within the 2 years preceding the search, for critical outcomes only.

For quality assurance of study identification, a 10% random sample of the literature search results was sifted by a second reviewer for the following review questions:

- What is the optimal screening programme to predict the risk of spontaneous preterm birth?
- What interventions are effective in preventing spontaneous preterm birth in twin and triplet pregnancy, including bed rest, progesterone and cervical cerclage?

Discrepancies were resolved by discussion between the two systematic reviewers and with a third (senior) systematic reviewer if necessary.

Methods of combining evidence

Data synthesis for intervention reviews

Pairwise meta-analysis

Pairwise meta-analysis of homogenous randomised trails was performed using Review Manager 5 (RevMan 5) software. For binary outcomes, such as occurrence of adverse events, the Mantel-Haenszel method of statistical analysis was used to calculate risk ratios (relative risks, RRs) with 95% confidence intervals (CIs).

For all outcomes with 0 events in both arms, risk difference was presented. For outcomes with 0 events in only one arm, both Peto ORs and risk difference were calculated. Corrections for zero cell counts are not necessary when using Peto's method. For this reason, this method performs well when events are very 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 initial hospital admission stay) were analysed using an inverse-variance method for pooling weighted mean differences.

Results from multiple observational studies of the same comparison were not pooled in meta-analysis but presented as a range of effects due to the high risk of selection bias in observational studies whereby differences in participant characteristics between study groups can lead to a biased estimate of effect of a treatment or intervention.

Subgroups for stratified analyses were decided for some review questions a priori at the protocol stage if the committee identified some subgroups to be different in terms of biological or clinical characteristics and the expectation was that the interventions would have a differing effect.

Forest plots were generated to present the results of meta-analyses and stratified or subgroup analyses (see appendix E of each intervention evidence review report).

Data synthesis for reviews of diagnostic test accuracy

When data from 3 or more studies were available, a meta-analysis of diagnostic test accuracy parameters was carried out. To show the differences between study results, pairs of sensitivity and specificity were plotted for each study on an area under the receiver-operating characteristic (ROC) curve (AUC) in RevMan5 (see appendix E of each evidence report with a diagnostic component). Study results were pooled using the bivariate method for the direct estimation of summary sensitivity and specificity using a random effects approach (in WinBUGS® software). Using the output from WinBUGS®, we constructed and plotted confidence regions and, where appropriate AUC, using methods outlined by Novielli 2010. As this is a Bayesian analysis, the evidence distribution is weighted by a distribution of prior beliefs. Vague non-informative priors were used for all parameters. For each analysis, a series of 50,000 burn-in simulations was run to allow convergence and then a further 50,000 simulations were run to produce the outputs. Convergence was assessed by investigating density plots, auto-correlation plots and history plots for parameters of interest. In cases where many cell counts were 0, 1 was added to each category (true positives, false positives, true negatives, false negatives) to ensure the model was able to run, while not significantly distorting the results.

One advantage of this approach is that it produces summary estimates of sensitivity and specificity that account for the correlation between the 2 measures (sensitivity and specificity). Other advantages of this method have been described elsewhere (Reitsma, 2005; Van Houwelingen, 1993; Van Houwelingen, 2002).

This model also assesses variability by incorporating the precision by which sensitivity and specificity have been measured in each study. A confidence ellipse is shown in the graph that indicates the confidence region around the summary sensitivity/specificity point. A summary AUC is also presented. From the WinBUGS® output we report the summary estimate of sensitivity and specificity (plus their 95% CIs) as well as between-study variation expressed as logit sensitivity and specificity, as well as correlations between the 2 measures of variation.

Sensitivity, specificity and AUC with 95% CIs were used as outcomes for diagnostic test accuracy. These diagnostic accuracy parameters were obtained from the studies or calculated by the technical team using data from the studies.

Sensitivity and specificity are measures of the ability of a test to correctly classify a person as having a condition or not having a condition. When sensitivity is high, a negative test result rules out the condition. When specificity is high, a positive test result rules in the condition. An ideal test would be both highly sensitive and highly specific, but this is frequently not possible and typically there is a trade-off.

The GS discussed and agreed following cut-offs were used when summarising sensitivity or and specificity:

- high: more than 90%

- moderate: 75% to 90%
- low: less than 75%.

AUC shows true positive rate (sensitivity) as a function of false positive rate (1 minus specificity). The GS discussed and agreed following cut-offs for AUC were used when determining the discriminative value of a test:

- the index test is worse than chance: lower than 0.50
- very poor: 0.50–0.60
- poor: 0.61–0.70
- moderate: 0.71–0.80
- good: 0.81–0.92
- excellent or perfect test: 0.91–1.00.

Data synthesis for prognostic reviews

Identification of accurate screening measures taken in the first trimester that predict the occurrence of feto-fetal transfusion syndrome (FFTS) or that predict spontaneous preterm birth could aid early identification and management strategies. Adjusted odds ratios (ORs) or RRs with 95% CIs reported by the studies were extracted to study the relationship between a given factor and the outcome of interest. Ideally the analyses would have adjusted for key confounders (such as gestational age) in order to be considered. Because of variation across the studies in terms of population, the risk factor, outcome and statistical methods (including adjustments for confounding factors), the prognostic data were not pooled but results from individual studies were reported.

Data synthesis for incidence reviews

Data were extracted from eligible studies as number of events (stillbirths or neonatal deaths or neonatal morbidities) and number of births or ongoing pregnancies or neonates born by weeks' gestation. For each outcome, the proportion of an event of interest was calculated as the number of stillbirths or neonatal deaths divided by the total number of ongoing pregnancies, or as the number of neonatal morbidities divided by the total number of neonates born. For each outcome, data were pooled where possible, for example, where population and weeks' gestation were reported. Mean, SD, CI, median and interquartile range (IQR) were calculated (median and IQR were presented in the evidence report). In addition, the crude risk of stillbirth and neonatal mortality per 1000 pregnancies and IQR were calculated based on the raw data (number of events of interest divided by the total number of pregnancies/births) reported for each individual study. Crude risks of neonatal morbidities per 1000 births and IQR were calculated based on the raw data (number of events of interest divided by the total number of pregnancies/births) reported for each individual study. Where there was variation across the studies in terms of population (e.g. OECD vs. non-OECD country), outcome definition, and/or inconsistency in reporting these data were not pooled but results from individual studies were reported.

Graphs were generated to present the results (see appendix M of each intervention evidence review report).

Appraising the quality of evidence

Intervention reviews

Pairwise meta-analysis

GRADE methodology (the Grading of Recommendations Assessment, Development and Evaluation)

For intervention reviews, the evidence for outcomes from the included studies was evaluated and presented using GRADE, which was 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 into account individual study quality factors and the meta-analysis results. The clinical evidence profile tables include details of the quality assessment and pooled outcome data, where appropriate, an absolute measure of intervention effect and the summary of quality of evidence for that outcome. In this table, the columns for intervention and control indicate summary measures of effect and measures of dispersion (such as mean and SD or median and range) for continuous outcomes and frequency of events (n/N; the sum across studies of the number of participants with events divided by sum of the number of participants) for binary outcomes. For all outcomes with 0 events in both arms, or where there were 0 events in only one arm, assessment of imprecision in GRADE was classed as "Serious".

The selection of outcomes for each review question was decided when each review protocol was discussed with the committee, and was informed by committee discussion and by key papers.

The evidence for each outcome in the intervention reviews was examined separately for the quality elements listed and defined in Table 2. Each element was graded using the quality levels listed in Table 3. Reporting or publication bias was taken into consideration in the quality assessment and reported in the clinical evidence profile tables if it was apparent.

The main criteria considered in the rating of these elements are discussed below. Footnotes were used in the GRADE profiles to describe reasons for grading a quality element as having serious or very serious limitations. The ratings for each component were combined to obtain an overall assessment for each outcome (Table 4).

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

Quality element	Description
Risk of bias	Limitations in the study design and implementation may bias the estimates of the treatment effect. High risk of bias for the majority of the evidence decreases confidence in the estimate of the effect.
Inconsistency	Inconsistency refers to an unexplained heterogeneity of results or findings.
Indirectness	Indirectness refers to differences in study population, intervention, comparator and outcomes between the available evidence and the review question, such that the effect estimate is changed. This is also related to applicability or generalisability of findings.

Quality element	Description
Imprecision	Results are imprecise when studies include relatively few patients and / or few events and thus have wide confidence intervals around the estimate of the effect. Imprecision results if the confidence interval includes the clinically important threshold (minimally important difference – see below).
Publication bias	Publication bias is a systematic underestimate or an overestimate of the underlying beneficial or harmful effect due to selective publication of studies.

Table 3: Levels of quality elements in GRADE

Levels of quality elements in GRADE	Description
None, or no serious	There are no serious issues with the evidence.
Serious	The issues are serious enough to downgrade the outcome evidence by 1 level.
Very serious	The issues are serious enough to downgrade the outcome evidence by 2 levels.

Table 4: Levels of overall quality of outcome evidence in GRADE

Overall quality of outcome evidence in GRADE	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.

Assessing risk of bias in intervention reviews

Bias is a systematic error, or a consistent deviation from the truth in the results. 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 tool for RCTs was used (see appendix H in the Developing NICE guidelines: the manual 2014).

The Cochrane risk of bias tool assesses the following possible sources of bias:

- selection bias
- performance bias
- attrition bias
- detection bias
- reporting bias.

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 design will impact on the estimation of the intervention effect.

More details about the Cochrane risk of bias tool can be found in section 8 of the Cochrane handbook of Systematic Reviews of Interventions (Higgins, 2011).

For systematic reviews of RCTs the AMSTAR checklist was used to assess risk of bias and for systematic reviews of other study types the Cochrane ROBIS checklist was used.

For observational studies the Cochrane risk of bias tool for non-randomised studies (ROBINS-I) was used (see appendix H in the Developing NICE guidelines: the manual 2014).

Assessing inconsistency in intervention reviews

Inconsistency refers to unexplained heterogeneity of results of meta-analysis. When estimates of the 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 applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled). For outcomes derived from a single study 'no serious inconsistency' was used when assessing this domain, as per GRADE methodology (Santesso 2016).

Statistical heterogeneity was assessed by visually examining the forest plots, and by considering the chi-squared test for significance at $p<0.1$ or an I-squared inconsistency statistic. Where considerable heterogeneity was present (an I-squared value of 66% or more), predefined subgroup analyses were performed. In the case of unexplained heterogeneity, possible causes were discussed with the committee before the final decision to pool data or not was made. If the heterogeneity still remained, a random effects (DerSimonian 2015) model was employed to provide a more conservative estimate of the effect.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency. Where heterogeneity (I^2) was >50% the evidence was downgraded by 1 level and where I^2 was >80% the evidence was downgraded by 2 levels.

Assessing indirectness in intervention reviews

Directness refers to the extent to which the populations, intervention, comparisons and outcome measures are similar 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. It was downgraded by one level of it affected only one part of the PICO characteristics and downgraded by 2 levels if indirectness was identified in more than one PICO element.

Assessing imprecision and clinical importance in intervention reviews

Imprecision in GRADE methodology refers to uncertainty around the effect estimate and whether or not there is a clinically important difference between interventions (that is, whether the evidence would clearly support 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 really concerned with whether the point estimate is accurate or correct (has internal or

external validity). Instead, it is concerned with the uncertainty about what the point estimate 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 mean value will fall on 95% of repeated samples, were this procedure to be repeated. The larger the trial, the smaller the 95% CI 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 comparison of treatment 'A' versus treatment 'B'. Three decision-making zones can be differentiated, bounded by the thresholds of clinical importance (minimally important differences; MIDs) for benefit and harm. The MID for harm for a positive outcome means the threshold at which treatment A is less effective than treatment B by an amount that is clinically important to people with the condition of interest (favours B).

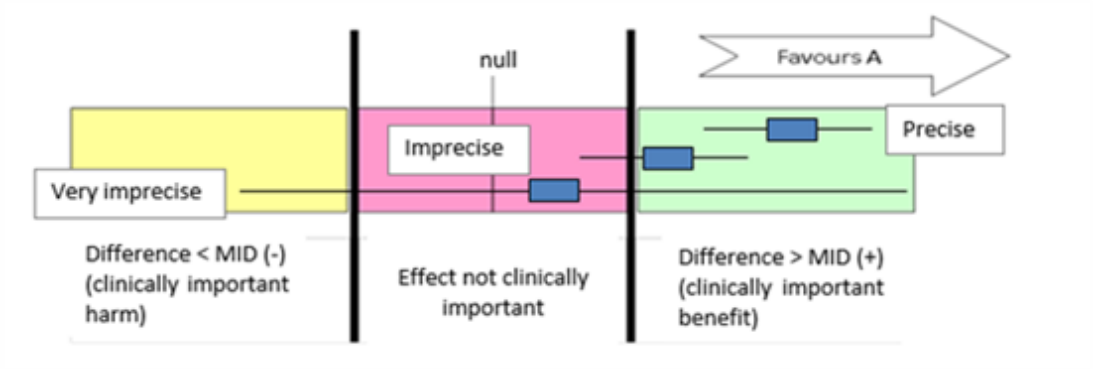
When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

If the effect estimate CI includes clinically important benefit (or harm) there is uncertainty over which decision to make (based on this outcome alone); i.e. the CI crosses 2 zones and 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').

An effect CI including clinically important benefit, clinically important harm and no effect is consistent with 3 possible decisions; i.e. the CI crosses all 3 zones and the effect estimate is considered to be very imprecise because the CI is consistent with 3 possible clinical decisions and therefore there is a considerable lack of confidence in the results. The evidence is therefore downgraded by 2 levels in the GRADE analysis ('very serious imprecision').

For all outcomes with 0 events in both arms, or where there were 0 events in only one arm, the risk difference was calculated. However, there was no agreement on the equivalent to an MID for these cases. Due to the low event rate which is usually associated wide confidence intervals it was decided to give those cases a 'serious' imprecision rating to prevent quality inflation for these outcomes.

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



MID, minimally important difference

Minimally important differences

The literature was searched for established MIDs for the selected outcomes in the evidence reviews. In addition, the committee members were asked whether they were aware of any accepted MIDs in the clinical community.

If no published or accepted MIDs were identified, the committee considered whether it was clinically acceptable to use the GRADE default MIDs to assess imprecision. The GRADE default MIDs for dichotomous outcomes are 0.8 and 1.25 (due to the statistical distribution of this measure this means that this is symmetrical on a log [RR] scale), and for continuous outcomes they 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). As no published MID values were identified, the committee agreed that GRADE default MID values were to be used as a starting point for all outcomes and any exception to their application based on the committee's consideration of clinical acceptability were noted and explained in the 'committee's discussion of the evidence' sections of the evidence reviews.

Where the point estimate of the effect of an outcome (RR or mean difference) was above the MID for a positive outcome or below the MID for a negative outcome the CI was used to assess clinical importance. If the 95% CI crosses the null effect a clinically important beneficial/harmful effect of one intervention compared with the other was determined. In marginal cases the 90% CI was also taken into consideration. If the 90% CI crossed the null effect no clinically important difference between the two interventions for a given outcome was determined, and where the 90% CI did not cross the null effect it was determined that there 'may be' a clinically important beneficial/harmful effect of one intervention compared with another due to the uncertainty around the estimate. Diagnostic test accuracy reviews

Modified GRADE methodology for diagnostic test accuracy reviews

The GRADE approach was modified to assess the quality of evidence about diagnostic test accuracy by adapting the principles of GRADE for intervention reviews as described below.

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. These 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:

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 test accuracy reviews

Risk of bias in diagnostic test accuracy studies was assessed using the risk of bias items from the QUADAS-2 checklist (see appendix H in Developing NICE guidelines: the manual 2014).

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

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

An overall risk of bias judgement was for each study was reached by considering the QUADAS-2 bias domains together. The risk of bias for the body of diagnostic test accuracy evidence was based on the risk of bias from the individual studies but with consideration of how much each study contributed to the overall evidence base.

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

Assessing inconsistency in diagnostic test accuracy reviews

Where there were multiple studies, the body of evidence was downgraded for serious inconsistency if there was unexplained variability between studies, when viewed on a forest plot or AUC. 'No serious inconsistency' is nevertheless used to describe this quality assessment in the GRADE tables for outcomes from single studies.

Assessing indirectness in diagnostic test accuracy reviews

Indirectness in diagnostic studies 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.

The indirectness for the body of diagnostic test accuracy evidence was based on the indirectness of the individual studies but with consideration of how much each study contributed to the overall evidence base.

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

Assessing imprecision in diagnostic test accuracy reviews

Imprecision was judged by comparing the CI of the estimate of sensitivity or specificity to clinical decision thresholds agreed beforehand by the committee. The committee decided whether sensitivity or specificity was the most important for decision making and agreed 2 threshold values. First a threshold for high sensitivity/specificity (above which the test would be definitely recommended) and second a threshold for low sensitivity/specificity (below which the test would not be recommended). If the CI of the estimate of sensitivity or specificity included 1 of these thresholds then the evidence was downgraded for serious imprecision, because it was consistent with two possible decisions. If the CI included both these thresholds then the evidence was downgraded for very serious imprecision because it was consistent with 3 possible decisions.

In the case of the screening reviews the judgement of precision was based on the CI for test sensitivity as this was considered by the committee to be the primary measure of interest. If the 95% CI included either 75% or 90%, the result was judged to be seriously imprecise (90% was considered to be the cut-off for the test to be highly sensitive and if the sensitivity was less than 75% the test was considered to be of low sensitivity). If the 95% CI included both 75% and 90%, the test was judged to be very seriously imprecise.

Prognostic reviews

Methodology for prognostic reviews

The GRADE approach was not used to assess the quality of evidence for prognostic reviews. Quality assessment was therefore conducted at the individual study level for these reviews, rather than according to outcome.

Assessing risk of bias in prognostic reviews

Risk of bias in individual prognostic studies was assessed using the risk of bias items from the QUIPS checklist (see appendix H in Developing NICE guidelines: the manual 2014). An overall risk of bias judgement was for each study was reached by considering the QUIPS bias domains together.

Incidence reviews

Methodology for incidence reviews

The GRADE approach was not used to assess the quality of evidence for incidence reviews. Quality assessment was therefore conducted at the individual study level for these reviews, rather than according to outcome..

Assessing risk of bias in incidence reviews

Risk of bias in individual incidence studies was assessed using an adapted version of the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Studies Reporting Prevalence Data (Munn 2015) for incidence studies. Many of these steps needed to be tailored for this type of evidence, particularly surrounding the stages of critical appraisal and synthesis. Individual parameters such as study design, method of sampling, adequacy of follow-up, ascertainment of the outcome, and appropriate determination of gestational age and chorionicity were assessed. Adapted criteria for assessment are summarised in Table 6. An overall risk of bias judgement was for each study was reached by considering the items together.

Table 6: Adapted Critical Appraisal Checklist for Incidence Studies

Item	Question
1	Was the sample frame appropriate to address the target population?
2	Were the study participants sampled in an appropriate way?
3	Were the criteria for inclusion in the sample clearly defined?
4	Were the study subjects and the setting described in detail?
5	Was the exposure measured in a valid and reliable way?
6	Were the outcome measures clearly defined, valid, reliable, and implemented consistently across all study participants?
7	Other limitations?

Evidence statements

Evidence statements are summary statements highlighting the key features of the clinical evidence presented. The wording of the evidence statements reflects the certainty or uncertainty in the estimate of effect. The evidence statements are presented by outcome or theme. They encompass the following key features of the evidence:

- the quality of the evidence (or in case of prognostic and incidence evidence the 'risk of bias' which was assessed for each study)
- study design
- the number of studies and the number of participants for the outcome concerned or prognostic/risk factor (quantitative evidence)
- where relevant an indication of the direction of effect (for example, if a treatment is beneficial or harmful compared with another, or whether there is no difference between the tested treatments or a summary of the effect size of the prognostic/risk factor). This was based on the MID of the relative effect.
- where relevant, whether or not the estimate of effect is clinically important

- in the case of diagnostic accuracy measures the evidence statement also included a summary of the effect size of the sensitivity and specificity.

The exceptions to the above were the evidence statements related to the 'timing of birth' evidence review. In this review the pattern of findings was described as a narrative according to groups divided by chorionicity and amnionicity (based on the graphical presentations in appendix M of evidence review D). They also encompassed some of the features highlighted above (such as risk of bias, study design, number of studies and number of participants).

Economic evidence

The aim of the economic input to the guideline was to inform the committee of potential economic issues related to twin and triplet pregnancy and to ensure that recommendations represented a cost effective use of healthcare resources. Health economic evaluations aim to integrate data on healthcare benefits (ideally in terms of quality-adjusted life-years (QALYs)) with the costs of different care options. In addition, the health economic input aimed to identify areas of high resource impact. These are recommendations which might have a large impact on Clinical Commissioning Groups' or Trusts' finances and so need special attention.

Reviewing economic evidence

Systematic reviews of economic literature were conducted for all review questions covered in the guideline update.

Inclusion and exclusion of economic studies

Titles and abstracts of articles identified through the searches were independently assessed for inclusion using the predefined eligibility criteria summarised in Table 7.

Table 7: Inclusion and exclusion criteria for the systematic reviews of economic evaluations

Inclusion criteria
Intervention or comparators according to the scope
Study population according to the scope
Only studies published in or after 2011. This date restriction was imposed in order to limit the evidence reviewed to that published since the previous NICE guideline (CG129)
Full economic evaluations (cost utility, cost effectiveness, cost benefit or cost consequence analyses) that assess both the costs and outcomes associated with the interventions of interest
Exclusion criteria
Abstracts with insufficient methodological details. Conference abstracts, poster presentations or dissertation abstracts were excluded
Cost-of-illness type studies

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

The quality of evidence was assessed using the economic evaluations checklist as specified in Developing NICE guidelines: the manual 2014. The Preferred Reporting

Items for Systematic Reviews and Meta-Analyses (PRISMA) for the search of economic evaluations is presented in the respective evidence reviews.

Details of economic evidence study selection, and lists of excluded studies, are provided in appendix H and appendix K of the respective evidence reports.

Economic modelling

The aims of the economic input to the guideline were to inform the guideline committee of potential economic issues related to the management of women with twin and triplet pregnancies in order 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 QALYs) with the costs of different care options

As well as reviewing the published economic literature, as described above, a new economic analysis was undertaken in areas prioritised by the committee in conjunction with the health economist. Topics were prioritised on the basis of the following criteria, in accordance with Developing NICE guidelines: the manual 2014:

- the overall importance of the recommendation, which may be a function of the number of people affected and the potential impact on costs and health outcomes per patient
- the current extent of uncertainty over cost effectiveness, and the likelihood that economic analysis will reduce this uncertainty
- the feasibility of building an economic model.

The rationale for prioritising review questions for economic modelling was set out in an economic plan agreed between NICE, the committee, and members of the technical team.

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

- What is the optimal screening programme to predict the risk of spontaneous preterm birth? (review B1)
- What interventions are effective in preventing spontaneous preterm birth in twin and triplet pregnancy? (review B2)

These review questions were considered in a single economic analysis to assess the cost-effectiveness of screening and prevention of spontaneous preterm birth.

The full methods and results of the original economic analysis are reported in appendix J of evidence report for review B1 and appendix J of evidence report for review B2. 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

NICE's report Social value judgements: principles for the development of NICE guidance sets out the principles that committees should consider when judging whether an intervention offers good value for money. In general, an intervention was

considered to be cost effective if any of the following criteria applied (given 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 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, or
- the intervention provided clinically 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 'Cost effectiveness and resource use' headings of the relevant sections in the evidence review chapters.

Developing recommendations

Guideline recommendations

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

The main considerations specific to each recommendation are outlined under the 'The committee's discussion of the evidence' headings within each evidence review chapter as well as the 'rationale and impact' sections in the short guideline.

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

Research recommendations

When areas were identified for which good evidence was lacking, the committee considered making recommendations for future research. For further details please refer to Developing NICE guidelines: the manual 2014.

Validation process

This guideline 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 at publication. For further details please refer to Developing NICE guidelines: the manual 2014.

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. For further details please refer to Developing NICE guidelines: the manual 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.

DerSimonian 2015

DerSimonian, R., Laird, N. Meta-analysis in clinical trials revisited. Contemporary clinical trials. 2015 Nov 30; 45:139-45.

Higgins 2011

Higgins JPT, Green S (editors) (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] The Cochrane Collaboration. Available from www.handbook.cochrane.org (accessed 22 January 2019)

Munn, 2015

Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. International journal of evidence-based healthcare. 2015;13(3):147-53.

NICE 2014

National Institute for Health and Care Excellence (NICE) (2014) Developing NICE guidelines: the manual (updated 2017). Available from https://www.nice.org.uk/process/pmg20/resources (accessed 22 January 2019)

NICE 2018

National Institute for Health and Care Excellence (NICE) (2014) NICE Policy on conflict 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 15 January 2019)

Novielli 2010

Novielli, N., Cooper, N. J., Abrams, K. R., Sutton, A. J., 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, 13, 952-7, 2010

Reitsma 2005

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., Zwinderman, A. H., Bivariate analysis of sensitivity and specificity produces

informative summary measures in diagnostic reviews, Journal of clinical epidemiology, 58, 982-90, 2005

Santesso 2016

Santesso, N., Carrasco-Labra, A., Langendam, M., Brignardello-Petersen, R., Mustafa, R.A., Heus, P., Lasserson, T., Opiyo, N., Kunnamo, I., Sinclair, D. and Garner, P., 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, 2016.

Schünemann 2013

Schünemann, H., Brożek, J, Guyatt, G., Oxman, A., (eds.). Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Working Group, 2013. Available from gdt.guidelinedevelopment.org/app/handbook/handbook.html (accessed 26 November 2017)

Van Houwelingen 1993

Van Houwelingen, H. C., Zwinderman, K. H., Stijnen, T., A bivariate approach to meta-analysis, Statistics in medicine, 12, 2273-84, 1993

Van Houwelingen 2002

Van Houwelingen, H. C., Arends, L. R., Stijnen, T., Advanced methods in meta-analysis: multivariate approach and meta-regression, Statistics in medicine, 21, 589-624, 2002