National Institute for Health and Care Excellence

Final

Postnatal care

Supplement 1: Methods

NICE guideline NG194

Development of the guideline and methods

April 2021

Final

Developed by the National Guideline Alliance 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 All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN: 978-1-4731-4078-3

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 about postnatal care.

What this guideline covers

Groups that are covered

- Women and babies from the birth of the baby until the end of the postnatal period, which, for the purposes of this guideline, is defined as 8 weeks after the birth. Questions on babies' feeding covered all relevant time periods, including the antenatal and postnatal periods.
- Women having twins and triplets were covered as a subgroup for selected review questions.

Key areas that are covered

- Planning the content and delivery of care.
- Identifying and assessing health and wellbeing needs in women.
- Identifying and assessing health and wellbeing needs in babies.
- Planning and management of babies' feeding.

For further details of that 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). A more up to date version is now available of the NICE manual; however, development of this guideline was initiated before publication of the new manual and to ensure consistent methods the 2014 version was used throughout.

Until March 2018, declarations of interest were recorded and managed in accordance with NICE's 2014 conflicts of interest policy. From April 2018, declarations 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– 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.

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.

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	Review question	Type of review
[A] Length of stay	• How does the length of postpartum stay affect women and their babies (single births)? • How does length of postpartum stay affect women and their babies (twins or triplets)?	Intervention
[B] Information transfer	What information needs to be communicated between healthcare professionals at transfer of care from birth care team to community care?	Qualitative
[C] Timing of first contact by midwife	• When should the first postnatal contact by midwives be made after transfer from place of birth to community care single births)? • When should the first postnatal contact by midwives be made after transfer from place of birth to community care (twins or triplets)?	Intervention
[D] Timing of first contact by health visitor	When should the first postnatal contact by health visitors be made?	Intervention
[E] Timing of comprehensive assessment	When should a comprehensive, routine assessment of the woman at the end of the postnatal period occur (for example at 6 weeks, 8 weeks or not at all)?	Intervention
[F] Content of postnatal care contacts	What is the essential content of the postnatal care contacts for women and babies?	Formal consensus
[G] Provision of information about the postnatal health of women	When and how should information be given to mothers and their partners about postnatal health of the mother?	Qualitative
[H] Tools for the clinical review of women	What tools for clinical review of women are effective during the first 8 weeks after birth? (including pain scores)	Intervention
[I] Assessment of secondary postpartum haemorrhage	How should early signs and symptoms of postpartum haemorrhage be assessed?	Intervention
[J] Perineal pain	What characteristics of perineal pain suggest the need for further evaluation?	Prognostic
[K] Information on lactation suppression	What information and support should be given to women about lactation suppression? And under what	Qualitative

Evidence review	Review question	Type of review
	circumstances should the information be provided?	
[L1] Signs and symptoms of serious illness in babies	What signs and symptoms (alone or in combination) in babies are associated with serious illness or mortality?	Diagnostic and prognostic
[L2] Scoring systems for illness in babies	Which scoring systems are accurate in identifying or predicting illness severity in babies?	Clinical prediction model and diagnostic
[M] Benefits and harms of bed sharing	What are the benefits and harms of co-sleeping?	Intervention
[N] Co-sleeping risk factors	What are the risk factors in relation to co-sleeping for sudden infant death syndrome (SIDS)?	Prognostic
[O] Emotional attachment	What interventions in the postnatal period are effective at promoting emotional attachment?	Intervention
[P] Breastfeeding interventions[1]	• What interventions are effective in starting and maintaining breastfeeding (single births)? • What interventions are effective in starting and maintaining breastfeeding (twins or triplets births)?	Intervention
[Q] Breastfeeding facilitators and barriers	What are perceived by parents to be the facilitators and barriers for starting and maintaining breastfeeding?	Qualitative
[R] Tools for predicting breastfeeding difficulties	What observations or clinical tools accurately predict breastfeeding difficulties?	Prognostic
[S] Breastfeeding information and support	• What information on breastfeeding do parents find helpful (single births)? • What information on breastfeeding do parents find helpful (twins and triplets)? • What support with breastfeeding do parents find helpful (single births)? • What support with breastfeeding do parents find helpful (twins and triplets)?	Qualitative
[T] Formula feeding information and support	• What information on formula feeding do parents find helpful? • What support with formula feeding do parents find helpful?	Qualitative

[1] *Original health economic analysis conducted*

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

- Supplement 1 (Methods; this document)
- Supplement 2 (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. Searches of websites of organisations, institutional repositories and internet search engines were also undertaken for relevant policies and related documents, including grey literature.

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: EMCare, MEDLINE, MEDLINE IN-PROCESS and Embase. For review questions that included systematic reviews and RCTs among the eligibility criteria, the Cochrane Central Register of Controlled Trials (CCTR), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE) and the Health Technology Assessment (HTA) database were also searched. For review questions that were highly focused on nursing, the Cumulative Index to Nursing and Allied Health Literature (CINAHL) database was also searched. Additionally, online repositories were searched for questions where guidelines were considered.

Where possible, searches were limited to studies published in English.

Searches were run once for all reviews during development. The guideline committee and the NGA technical team considered the review questions for which the searches might need to be updated, and after prioritising against a number of criteria, made a decision to selectively rerun the searches for evidence reviews A, C, D, G, H, I and O, which were performed between 9-12 weeks in advance of the final guideline committee meetings before consultation on the draft guideline; these reruns were completed during December 2019. Any studies added to the databases after December 2019 (including those published before December 2019 but not yet indexed) were not considered for inclusion.

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 Emcare, Medline, Medline in Process and Embase.

Where wider populations were reviewed, additional searches, using the population and intervention (or equivalent) search terms used in the evidence reviews, were conducted, combined with a search filter for economic evaluations, if and as applicable.

Where possible, searches were limited to studies published in English.

The systematic economic literature searches were updated during December 2019, 11-12 weeks in advance of the final committee meeting before consultation on the draft guideline. Any studies added to the databases after December 2019 (including those published before December 2019 but not yet indexed) were not considered for inclusion.

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

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. The principal search strategies for each search were also quality assured by a second information scientist using an adaptation of the PRESS 2015 Guideline Evidence-Based Checklist (McGowan 2016). In addition, all publications highlighted by stakeholders at the time of the consultation on the draft scope were considered for inclusion.

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 priorities for economic analysis 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 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.

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 were considered for inclusion.

For prognostic reviews, prospective and retrospective cohort and case–control studies and case series were considered for inclusion. Studies that included multivariable analysis were prioritised. Review L1 was designed to include diagnostic and prognostic data and L2 was a clinical prediction model review designed to include both model performance and diagnostic accuracy data. For both reviews cross-sectional studies were considered for inclusion as well as cohort studies.

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.

For the formal consensus review (F), published guidelines were included (see formal consensus review section).

The committee were 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 K of the corresponding evidence review. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for the search of clinical evidence is presented in Appendix C of each 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 from RCTs was conducted where possible using Cochrane Review Manager (RevMan5) software. Where non-randomised evidence was used, this was not meta-analysed.

For dichotomous outcomes, such as mortality, the Mantel–Haenszel method with a random effects model was used to calculate risk ratios (RRs), due to the variability of included interventions. For all outcomes with zero events in both arms the risk difference was presented. For outcomes with low event rates (<1%), Peto odds ratios (ORs) were calculated as this method performs well when events are rare (Bradburn 2007). In meta-analysis, if some (but not all) studies have zero events in both arms, RD was presented to account for study weights, as in evidence review A on length of stay.

If a study reported only the summary statistic and 95% confidence interval (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.

No continuous data was pooled because there was not enough evidence.

Subgroups for stratified analyses were agreed for some review questions as part of protocol development.

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

Meta-regression

Meta-regression analysis was considered appropriate for evidence review P, the effectiveness of breastfeeding interventions, specifically for the interventions that started in the antenatal period and finished in the postnatal period. Meta-regression is used in meta-analysis to simultaneously investigate the impact of moderator variables on study effect size. In this case meta-regression was considered appropriate because there was a large volume of included studies (n=63) each with different intervention characteristics (or 'moderator variables'), for example where the intervention was delivered, how long it lasted for, how the intervention was delivered and how often.

For the purpose of this meta-regression analysis, each study was categorised using the following variables.

- Number of contact visits: 0, 1, 2-3, 4-8 and 9+.
- How delivered: face to face on an individual basis, face to face in a group, remote, self-help.
- Duration of contact: contact with the intervention lasted less than 8 weeks, contact with the intervention lasted more than 8 weeks.
- Where the intervention was delivered: at home, in a healthcare setting, combination of both home and healthcare setting.

The following analyses were conducted for each outcome.

- Initiation of breastfeeding
    - How delivered
        - Face to face as an individual versus standard care
        - Remote versus standard care
        - Self-help versus standard care
    - Number of contacts
        - 0-1 versus standard care
        - 2-3 versus standard care
        - 4-8 versus standard care
        - 9+ versus standard care
    - Duration of contact
        - Less than 8 weeks versus standard care
        - More than 8 weeks versus standard care
    - Where delivered
        - Healthcare setting versus standard care
        - Home setting versus standard care
        - Both healthcare and home setting versus standard care

    - A final model including number of contacts, how delivered and where delivered
        - 2-3 contacts versus 0-1 contacts
        - 4-8 contacts versus 0-1 contacts
        - 9+ contacts versus 0-1 contacts
        - Face to face as an individual versus standard care
        - Self-help versus standard care
        - Healthcare setting versus both healthcare and home setting
        - Home setting versus both healthcare and home setting

NB. Duration of contact was not included in the final model because there was not enough data, furthermore, duration of contacts was considered to be captured by the number of contacts.

- Any breastfeeding at 3-14 days
    - How delivered
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
    - Number of contacts
        - 0 versus standard care
        - 1 versus standard care
        - 2-3 versus standard care
        - 4-8 versus standard care
        - 9+ versus standard care
    - Duration of contact
        - Less than 8 weeks versus standard care
        - More than 8 weeks versus standard care
    - Where delivered
        - Healthcare setting versus standard care
        - Home setting versus standard care
        - Both healthcare and home setting versus standard care
    - A final model including number of contacts, how delivered and where delivered
        - 1 contact versus 0 contacts
        - 2-3 contacts versus 0 contacts
        - 4-8 contacts versus 0 contacts
        - 9+ contacts versus 0 contacts
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
        - Healthcare setting versus both healthcare and home setting
        - Home setting versus both healthcare and home setting

- Exclusive breastfeeding at 3-14 days

  - How delivered
    - Face to face as an individual versus standard care
    - Face to face as part of a group versus standard care
    - Remote versus standard care
    - Self-help versus standard care
  - Number of contacts
    - 0 versus standard care
    - 1 versus standard care
    - 2-3 versus standard care
    - 4-8 versus standard care
    - 9+ versus standard care
  - Duration of contact
    - Less than 8 weeks versus standard care
    - More than 8 weeks versus standard care
  - Where delivered
    - Healthcare setting versus standard care
    - Home setting versus standard care
    - Both healthcare and home setting versus standard care
  - A final model including number of contacts, how delivered and where delivered
    - 1 contact versus 0 contacts
    - 2-3 contacts versus 0 contacts
    - 4-8 contacts versus 0 contacts
    - 9+ contacts versus 0 contacts
    - Face to face as an individual versus standard care
    - Face to face as part of a group versus standard care
    - Remote versus standard care
    - Self-help versus standard care
    - Healthcare setting versus both healthcare and home setting
    - Home setting versus both healthcare and home setting

- Any breastfeeding at 6-12 weeks
  - How delivered
    - Face to face as an individual versus standard care
    - Face to face as part of a group versus standard care
    - Remote versus standard care
    - Self-help versus standard care
  - Number of contacts
    - 0 versus standard care
    - 1 versus standard care
    - 2-3 versus standard care
    - 4-8 versus standard care
    - 9+ versus standard care
  - Duration of contact
    - Less than 8 weeks versus standard care
    - More than 8 weeks versus standard care
  - Where delivered

        - Healthcare setting versus standard care
        - Home setting versus standard care
        - Both healthcare and home setting versus standard care
    - Model results
        - 1 contact versus 0 contacts
        - 2-3 contacts versus 0 contacts
        - 4-8 contacts versus 0 contacts
        - 9+ contacts versus 0 contacts
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
        - Healthcare setting versus both healthcare and home setting
        - Home setting versus both healthcare and home setting

- Exclusive breastfeeding at 6-12 weeks
    - How delivered
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
    - Number of contacts
        - 0 versus standard care
        - 1 versus standard care
        - 2-3 versus standard care
        - 4-8 versus standard care
        - 9+ versus standard care
    - Duration of contact
        - Less than 8 weeks versus standard care
        - More than 8 weeks versus standard care
    - Where delivered
        - Healthcare setting versus standard care
        - Home setting versus standard care
        - Both healthcare and home setting versus standard care
    - A final model including number of contacts, how delivered and where delivered
        - 1 contact versus 0 contacts
        - 2-3 contacts versus 0 contacts
        - 4-8 contacts versus 0 contacts
        - 9+ contacts versus 0 contacts
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
        - Healthcare setting versus both healthcare and home setting
        - Home setting versus both healthcare and home setting

- Any breastfeeding at 16 to 24 weeks:
    - How delivered
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
    - Number of contacts
        - 0 versus standard care
        - 1 versus standard care
        - 2-3 versus standard care
        - 4-8 versus standard care
        - 9+ versus standard care
    - Duration of contact
        - Less than 8 weeks versus standard care
        - More than 8 weeks versus standard care
    - Where delivered
        - Healthcare setting versus standard care
        - Home setting versus standard care
        - Both healthcare and home setting versus standard care
    - A final model including number of contacts, how delivered and where delivered
        - 1 contact versus 0 contacts
        - 2-3 contacts versus 0 contacts
        - 4-8 contacts versus 0 contacts
        - 9+ contacts versus 0 contacts
        - Face to face as an individual versus standard care
        - Face to face as part of a group versus standard care
        - Remote versus standard care
        - Self-help versus standard care
        - Healthcare setting versus both healthcare and home setting
        - Home setting versus both healthcare and home setting

Sample Win BUGS code for the analysis of any breastfeeding at 16 to 26 weeks, including the variables how the intervention was delivered, the number of contacts for the intervention and where the intervention was delivered is given in evidence report P, appendix M. Other analyses used the same substantive code as the one provided, but was modified to include the relevant predictor variables for the model under consideration.

See evidence report P for further details of the meta-regression and results.

Data synthesis for reviews of diagnostic test accuracy and prediction tools

When diagnostic test accuracy was measured dichotomously, sensitivity, specificity, and positive and negative likelihood ratios were used as outcomes. These diagnostic test accuracy parameters were calculated by the NGA technical team using data reported in the articles.

Meta-analysis of diagnostic test accuracy parameters was conducted, using the statistical software STATA (v13), if there were data from four or more studies that could be pooled.

Data synthesis for prognostic reviews

ORs or RRs 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. Ideally analyses would have adjusted for key confounders (such as age) to be considered for inclusion. Recognising variation across studies in terms of populations, risk factors, outcomes and statistical analysis methods (including adjustments for confounding factors), prognostic data were not pooled, but results from individual studies were presented in the evidence reviews.

Additional calculations were required to be undertaken by the NGA technical team for evidence review N, co-sleeping risk factors. The aim of this review was to identify risk factors in relation to co-sleeping for sudden unexpected death in infancy. The exposure of interest was co-sleeping with a risk factor and the reference standard was co-sleeping. Where the papers reported the OR for the exposures of interest (co-sleeping with a risk factor) against a reference standard not of interest (typically 'not co-sleeping') and also within the same paper, the OR for co-sleeping with no additional risk factor against the same reference standard, the NGA technical team calculated the added risk of co-sleeping with a risk factor compared to co-sleeping by using an equation by Franchini (2012) which calculates the difference between the two ORs, using similar methods to that of indirect treatment comparisons. For example, if the paper reports A versus B and also C versus B and we are interested in A versus C, the Franchini method can be used to calculate the added risk.

Data synthesis for qualitative reviews

Where possible, a meta-synthesis was conducted to combine evidence from qualitative studies. Whenever a qualitative theme relevant to the protocol was identified, these were extracted along with supporting quotes where available. When all themes had been extracted from studies, common concepts were categorised and tabulated and usually broken down into themes and subthemes, depending on the complexity of the data. Information about how many studies contributed to each theme or subtheme was also captured.

In qualitative synthesis, a theme being reported more than other themes across included studies does not necessarily mean that the theme is more important than other themes. The aim of qualitative research is to identify new perspectives on a particular topic. Study types and populations in qualitative research can differ widely, meaning that themes identified by just one or a few studies can provide important new information on a given topic. Therefore, for the purpose of the qualitative reviews in this guideline, it was planned that further studies would not be added when they reported the same themes as had already been identified from other studies because the emphasis was to be on conceptual robustness rather than quantitative completeness of the evidence. This would have implications for the types and numbers of studies included in the qualitative reviews, with study inclusion continuing until no new relevant data could be found regarding a topic that would add to or refute it. This concept is referred to in the literature as 'theoretical saturation' (Dixon-Woods 2005). In this guideline theoretical saturation was reached for 4 qualitative reviews (described in evidence reports G, Q, S and T).

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.

Themes and subthemes were listed in the evidence reports for the committee to understand the relationship between them.

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. More information about this tool can be found on the developer's website.

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 clinical evidence profile tables include details of the quality assessment and pooled outcome data, where appropriate, a relative and 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 the sum across studies of the number of participants in each arm 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 dichotomous outcomes.

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
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 tool (see Appendix H in Developing NICE guidelines: the manual; NICE 2014). The original version of the tool was updated in August 2019 and from this point, version 2 was used to assess included RCTs, with the result that it was used in review P, breast feeding interventions.

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 chosen design and methodology will impact on the estimation of the intervention effect. It is sometimes argued that if the nature or design of a study makes it difficult or unethical to reduce the risk of bias, for example through blinding participants, then the study should not be penalised (downgraded). However, the NGA technical team took the view that a lack of blinding should lead to downgrading on performance bias, because regardless of the reason, a lacking of blinding could still potentially lead to deviations from the intended interventions and bias in measurement of the outcomes. The same standard was applied across all reviews.

More details about the Cochrane risk of bias tool can be found in Section 8 of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins and Thomas 2019).

For systematic reviews of RCTs the AMSTAR checklist was used and for systematic reviews of other study types 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 considerable heterogeneity, and more than 75% 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. In the case of unexplained heterogeneity, sensitivity analyses were planned based on the quality of studies, eliminating studies at high risk of bias (in relation to randomisation, allocation concealment and blinding, and/or missing outcome data).

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.

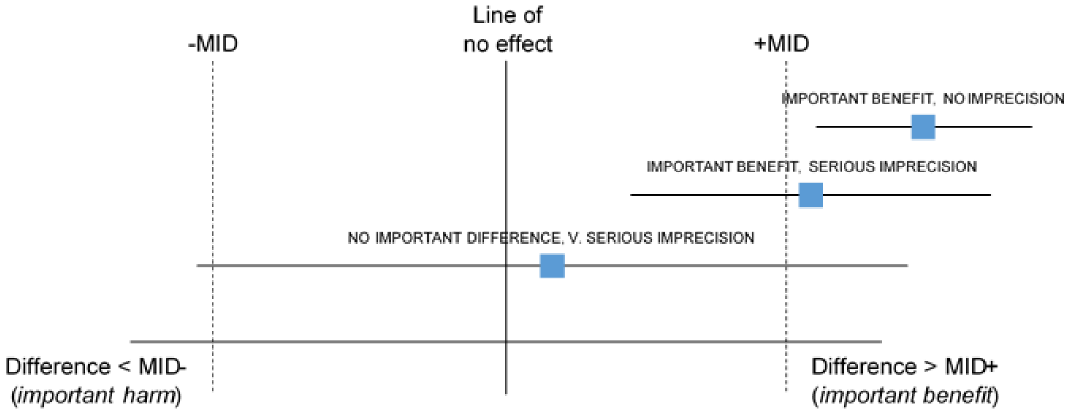
When the CI crosses 2 zones, it is uncertain in which zone the true value of the effect estimate lies and therefore there is uncertainty over which decision to make. The CI is consistent with 2 possible decisions, therefore, the effect estimate is considered to be imprecise in the GRADE analysis and 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.

Exceptions to the above approach are described in the section below.

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


Line of no effect
-MID
+MID
IMPORTANT BENEFIT, NO IMPRECISION
IMPORTANT BENEFIT, SERIOUS IMPRECISION
NO IMPORTANT DIFFERENCE, V. SERIOUS IMPRECISION
Difference < MID-
(important harm)
Difference > MID+
(important benefit)

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 most outcomes. For dichotomous outcomes minimally important thresholds for a RR of 0.8 and 1.25 respectively were used as default MIDs in most of the evidence reviews. The committee also chose to use 0.8 and 1.25 as the MIDs for ORs 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, as low event rates OR are mathematically similar to RR making the extrapolation appropriate. There were some exceptions when the default MIDs were not used and these are outlined below.

When there were zero events in both arms, for which the risk difference was presented, as in review A. In this context imprecision cannot be assessed against relative CIs (such as 0.8 to 1.25) so instead, a sample size rule (akin to optimal information size criteria) of 300 was used, with MID as the null effect. Outcomes were downgraded once if 1 of these applied and twice if they both applied to the risk difference. 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 300 and 500 as cut-offs for very serious and serious imprecision respectively. The committee used these numbers based on commonly used optimal information size thresholds.

Where the committee felt that any improvement represented an important difference (as in breastfeeding outcomes in review P and infant mortality in review N) the line of no statistically significant effect was applied. That is, any statistically significant change was considered to be important in practice and in that case, there was no imprecision. If there was no statistically significant effect (that is that the line of null effect was crossed), the effect estimate was considered to have serious imprecision. If the point estimate was greater than the MID and the 95% CI crossed the line of no effect but the 90% CI did not, it was judged that while there may be an important effect, there is uncertainty around the estimate.

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

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 (sensitivity, specificity and likelihood ratios).

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.

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

Quality element	Description
	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 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 visually by inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis. This was supported 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. If there was considerable or very serious heterogeneity 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.

The thresholds were:

- sensitivity: low threshold 75%, high threshold 90%
- specificity: low threshold 75%, high threshold 90%.

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.

In review L2, the following cut-offs were used when summarising the performance of the scoring systems:

- very useful test: sensitivity ≥90%
- moderately useful test: sensitivity 75% to 89%
- not a useful test: sensitivity ≤75%

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. Adapted GRADE was not used for evidence from case series; instead quality of case series evidence was assessed using the Checklist for Case Series developed by the Joanna Briggs Institute. More information about this tool can be found on the developer's website.

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.

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. Prognostic 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 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

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.

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 dichotomous outcomes (RR) 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.

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 according to themes 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 (as with other types of evidence) to obtain an overall assessment of quality for each theme as described in Table 9.

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. As noted above, it is not equivalent to the number of studies contributing to a theme, but rather to the depth of evidence and whether sufficient quotations or observations were provided to underpin the findings.

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.

Formal consensus reviews

Formal consensus was carried out using the nominal group technique (Murphy 1998) for evidence review F. This is a structured method focusing on the opinions of individuals within a group. Due to this focus on individuals it is referred to as a ‘nominal group’ technique. It usually involves anonymous voting with an opportunity to provide comments. It is usually conducted by an iterative process in which options with low agreement are eliminated and options with high agreement are retained. Using the comments that individuals provided, options with medium agreement are revised and then considered in a second round.

Details of the nominal group technique as used in this guideline

A search was conducted for relevant published guidelines and systematic reviews. Systematic reviews were included in order to plug any gaps highlighted by the guidelines in terms of coverage of the areas that the committee agreed would be essential content of postnatal care contacts. These were listed a priori in the review protocol.

Only international or national guidelines that had been developed on the basis of an evidence review were included. It was agreed that these would be more generalisable than locally developed guidelines and would have a more robust basis for drafting recommendations. In order to identify the most relevant literature, only those guidelines published since the previous NICE recommendations on the content of postnatal care were published in NG37 (2006) were considered. For consistency, the same date cut-off was used in a search for published systematic reviews which were considered in parallel with the published guidelines. All potentially relevant guidelines were assessed for quality using the Appraisal of Guidelines for Research and Evaluation (AGREE II) instrument (see assessing quality below). The 2 published systematic reviews that were identified were assessed for quality using the Risk of Bias in Systematic Reviews (ROBIS) checklist.

Once the guidelines had been assessed, the NGA technical team extracted relevant recommendations from these guidelines and derived a set of statements for all included topics. All statements were checked for practical content by the NGA clinical advisor and the committee chair. If no recommendations existed within the included guidelines for a particular element of postnatal care content, then findings from the published systematic reviews were used to produce statements.

The formal consensus exercise was conducted over 2 committee meetings. At the initial meeting the statements were presented to the committee in a questionnaire format. All committee members were invited to take part in the formal consensus exercise (this did not include the chair or guideline clinical advisor as they had been involved in deriving the statements, nor co-opted members). Committee members were asked to rate each statement based on their personal opinion of what they believed 'best practice' would be. The statements were rated using a 9-point Likert scale, where 1 represents 'strongly disagree', 5 represents 'neither agree nor disagree', and 9 represents 'strongly agree'. The participants were also able to record for any statement that they believed they had insufficient knowledge to provide a rating. There was also space for written comments about each statement. Once this first round of voting had been conducted, the NGA technical team calculated overall percentage agreement for each individual statement. Statements with 80% or greater agreement were kept, and were to be used to inform recommendations. Statements with less than 60% agreement were discarded unless there were obvious and addressable issues identified from any comments. Those statements with 60% to 80% agreement were redrafted by the NGA technical team (using the written comments if provided).

The redrafted statements were placed into the same questionnaire format as round 1 of the formal consensus process. Committee members were sent the revised statements electronically and asked to rate them in the same way as in the first round. Responses were emailed back to the NGA technical team, who calculated agreement as above.

At the following committee meeting, statements with 80% or greater agreement (from rounds 1 and 2) were presented as the evidence to inform the development of recommendations. The statements were discussed and the committee used them in combination with their knowledge and experience to develop the recommendations.

Assessing quality of guidelines in formal consensus reviews

Potentially relevant guidelines were assessed for quality using AGREE II instrument (Table 11). The tool assesses 6 domains: scope and purpose, stakeholder involvement, rigour of development, clarity of presentation, applicability and editorial independence.

Within each domain there is a set of questions, each of which is scored using a 7-point scale (1 – 'strongly disagree' to 7 – 'strongly agree'). Each section is rated and then an overall score for that domain is calculated. Two reviewers independently rated all identified guidelines using this method (see the AGREE II for detailed instructions). The committee took account of the quality ratings during discussions and when they agreed recommendations based on the formal consensus review findings.

Table 11: Assessing quality of guidelines

Domain	Description
Scope and purpose	Assesses the aim of the guideline, the specific health questions, and the target population
Stakeholder involvement	Assesses the extent to which the guideline involved the appropriate stakeholders, and whether it represents the views of intended users
Rigour of development	Assesses the methods used to gather and synthesise the evidence and to construct the recommendations
Clarity of presentation	Assesses the language, format and structure of the guideline
Applicability	Assesses likely barriers and facilitators of implementation, uptake and resource implications of the guideline
Editorial independence	Assesses the likelihood of the recommendations being biased and potential conflict of interests

Reviewing economic evidence

Systematic reviews of economic literature were conducted for all review questions covered in the guideline, unless economic evidence was not relevant to a review question. In addition, modelling studies that estimated long-term benefits to women and babies and related cost-savings associated with breastfeeding (any or exclusive) were reviewed in order to identify modelling components such as model structures, clinical outcomes associated with breastfeeding, clinical and cost data and further assumptions that could be adopted or used after adaptation when developing the guideline economic model on the same topic.

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 12.

Table 12: Inclusion and exclusion criteria for systematic reviews of economic evaluations

Inclusion criteria
Only studies from the Organisation for Economic Co-operation and Development (OECD) member countries were included, as the aim of the review was to identify economic information transferable to the UK context.
Only studies published from 2004 onwards were included in the review. This date restriction was imposed so that retrieved economic evidence was relevant to current healthcare settings and costs.
Selection criteria regarding the populations and interventions assessed were identical to the clinical literature review.
Full economic evaluations that compared 2 or more relevant options and considered both costs and consequences as well as costing analyses that compared only costs between 2 or more interventions.
Clinical effectiveness data utilised in the analysis should have been derived from a literature review, a clinical trial, a prospective or retrospective cohort study, or a study with a before-and-after design.
The outcome measure of the economic analysis should be the quality-adjusted life-year (QALY) or one of the measures considered in the clinical review.
Studies should be reporting separately costs for each option assessed, from a healthcare perspective.
Exclusion criteria
Poster presentations and abstracts in conference proceedings.
Non-English language papers.
Non-comparative studies.
Studies that adopted a non-healthcare perspective and did not consider healthcare costs.

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.

Evidence tables of included economic studies and lists of economic studies excluded after obtaining full text with reasons for exclusion, are provided in Appendix H and Appendix K, respectively, of the relevant evidence reviews. The PRISMA for the search of economic evaluations is presented in Appendix G of each evidence review.

Consideration of modelling studies that estimated long-term benefits to women and babies and related cost-savings associated with breastfeeding

A systematic review of modelling studies that assessed the long-term benefits and cost-savings associated with breastfeeding was undertaken, to identify parameters that could inform the economic model that was developed to inform the cost-effectiveness of interventions for starting and maintaining breastfeeding, assessed in evidence review P.

The titles and abstracts of papers identified through the searches were assessed for inclusion in this review using broad eligibility criteria defined in Table 13.

Table 13: Inclusion and exclusion criteria for the systematic review of modelling studies that estimated long-term benefits to women and babies and related cost-savings associated with breastfeeding

Inclusion criteria
Modelling studies from any country were considered.
The study population should be women breastfeeding term babies that were healthy at birth (single or multiple births).
The studies should estimate projected benefits to women and/or babies and/or related cost-savings associated with breastfeeding.
Exclusion criteria
Poster presentations and abstracts in conference proceedings.
Non-English language papers.

Once the screening of titles and abstracts was complete, full-text versions of the selected papers were acquired.

Modelling studies that met inclusion criteria and those that were excluded after full text was obtained are reported in Appendix N of evidence review P.

Appraising the quality of economic evidence

The applicability and quality of economic evaluations in this guideline were appraised using the methodology checklist reported in Developing NICE guidelines: the manual (NICE 2014), Appendix H, for all studies that met the inclusion criteria.

The methodological assessment of economic studies considered in this guideline has been summarised in economic evidence profiles that were developed for each review question for which economic evidence was available. All studies that fully or partially met the applicability and quality criteria described in the methodology checklist were considered during the guideline development process.

Economic profiles of all economic studies that were considered during guideline development, including de novo economic analyses undertaken for this guideline, are provided in Appendix I of the respective evidence reviews.

Economic modelling

The aims of the economic input to the guideline were to inform the guideline committee of potential economic issues related to the postnatal care of women and their babies in order to ensure that recommendations represented a cost-effective use of healthcare resources. Health economic evaluations aim to integrate data on care benefits (ideally in terms of quality-adjusted life-years, QALYs) with the costs of different care options. In addition, the economic input aimed to identify areas of high resource impact, as these need to be supported by robust evidence on cost effectiveness.

Areas for economic modelling were prioritised by the committee. 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 NGA technical team. Economic modelling was undertaken in areas with likely major resource implications, where the current extent of uncertainty over cost effectiveness was significant and economic analysis was expected to reduce this uncertainty. The following economic questions were selected as key issues that were addressed by economic modelling.

- Cost-effectiveness of interventions for the initiation and maintenance of breastfeeding. The methods and results of the de novo economic analysis are fully reported in Appendix J of evidence review P.
- Cost-effectiveness of shorter versus longer postpartum stay. This question was originally prioritised for economic modelling, however, clinical evidence suggested that there were no significant differences in outcomes between early and late discharge; therefore, this question was de-prioritised at a later stage as there was no need for economic modelling.
- Cost-effectiveness of clinical tools for the clinical review of women postnatally. This question was not possible to model, due to lack of relevant clinical evidence.
- Cost-effectiveness of different approaches for the assessment of early signs and symptoms of postpartum haemorrhage. This question was not possible to model, due to lack of relevant clinical evidence.

When new economic analysis was not prioritised or was not possible to conduct, the committee made a qualitative judgement regarding cost effectiveness by considering expected differences in resource use and costs between options, alongside clinical effectiveness evidence identified from the clinical evidence review.

Cost effectiveness criteria

NICE's report The NICE Principles 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 committee's considerations of cost effectiveness are discussed explicitly under the heading 'The committee's discussion of the evidence' under subheading 'Cost effectiveness and resource use' in the relevant evidence reviews.

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).

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.

Dixon-Woods 2005

Dixon-Woods M, Agarwal S, Jones D et al. Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research & Policy 10(1), 45–53, 2005

Franchini 2012

Franchini, A. J., S. Dias, A. E. Ades, J. P. Jansen, N. J. Welton. Accounting for correlation in network meta-analysis with multi-arm trials. Research Synthesis Methods 3, 142-160, 2012

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. 158, 280–286, 2013

Higgins 2011

Higgins JPT, Thomas, J (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 6 The Cochrane Collaboration, 2019. Available from https://training.cochrane.org/handbook/current (accessed 18 March 2020)

Lewin 2015

Lewin S, Glenton C, Munthe-Kaas H et al. 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, 2015

McGowan 2016

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

Murphy 2016

Murphy MK, Black NA, Lamping DL, McKee CM, Standerson CFB, Askam, J. Consensus development methods, and their use in clinical guideline development. Heath Technology Assessment, 2, 1998

NICE 2014

National Institute for Health and Care Excellence (NICE). Developing NICE guidelines: the manual, 2014 (updated 2018). Available from https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview (accessed 18 March 2020)

NICE 2018

National Institute for Health and Care Excellence (NICE). NICE Policy on conflicts of interest, 2014 (updated 2017). Available from https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/declaration-of-interests-policy.pdf (accessed 18 March 2020)

Santesso 2016

Santesso N, Carrasco-Labra A, Langendam M et al. 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