National Institute for Health and Care Excellence

Final

Heavy menstrual bleeding (update)

NICE guideline: methods

NICE guideline 88

Methods

March 2018

Final

Evidence reviews were developed by National Guideline Alliance, hosted by 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 All NICE guidance is subject to regular review and may be updated or withdrawn.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: 978-1-4731-2777-7

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to produce the update for this guideline.

The remit for this guideline update is to revise the NICE clinical guideline on the structural diagnosis of causes of heavy menstrual bleeding and the management of heavy menstrual bleeding.

What this guideline covers

Groups that are covered

The guideline update covers women with heavy menstrual bleeding, including:

- women with suspected or confirmed fibroids
- women with suspected or confirmed adenomyosis
- women with no identified pathology.

Women who wish to preserve their fertility have been identified as a subgroup needing specific consideration.

Clinical areas that are covered

The guideline update covers the following clinical issues:

- clinical and cost-effectiveness of hysteroscopy and pelvic ultrasound scan to detect causes of heavy menstrual bleeding
- clinical and cost-effectiveness of diagnostic imaging techniques to detect adenomyosis in women presenting with heavy menstrual bleeding
- clinical and cost-effectiveness of pharmacological and surgical management of heavy menstrual bleeding.

Note that guideline recommendations will normally fall within licensed indications. Exceptionally, and only if clearly supported by evidence, use outside a licensed indication may be recommended. This guideline will assume that prescribers will use a drug's summary of product characteristics to inform decisions made with individual patients.

For further details please refer to the scope on the NICE website (https://www.nice.org.uk/guidance/gid-ng10012/documents/final-scope).

What this guideline does not cover

Groups that are not covered

The guideline does not cover the following groups:

- women without heavy menstrual bleeding who have other gynaecological bleeding, for example, intermenstrual bleeding or post-coital bleeding

- women with gynaecological conditions in which heavy menstrual bleeding is not the main problem, for example, women with endometriosis.

Clinical areas that are not covered

The following areas in the published guideline were not updated:

- definition of heavy menstrual bleeding
- education and information provision
- competencies:
  - training
  - maintenance
  - governance
- the clinical and cost-effectiveness of treatment with progesterone receptor modulators for fibroids of 3 cm or more in diameter (this topic was reviewed by the NICE standing committee, and an addendum to the NICE guideline on Heavy menstrual bleeding [CG44] was published in August 2016).

Recommendations in areas that were not updated were edited to ensure that they meet the current editorial standard, and reflect the current policy and practice context.

Management of endometriosis associated with heavy menstrual bleeding is not covered by this guideline but is covered in the NICE guideline on Endometriosis: diagnosis and management published in September 2017.

Methods

This chapter sets out in detail the methods used to review the evidence and to generate recommendations in the guideline. This guideline was developed using the methods described in the 2014 NICE guidelines manual (NICE 2014).

Declarations of interest were recorded according to the 2014 NICE conflicts of interest policy.

Developing the review questions and outcomes

The 3 review questions developed for this guideline were based on the key areas identified in the guideline update scope (see https://www.nice.org.uk/guidance/gid-ng10012/documents/final-scope). They were drafted by the NGA and refined and validated by the guideline committee (see Table 1).

The review questions were based on the following frameworks:

- intervention review: population, intervention, comparator and outcome (PICO);
- diagnostic test accuracy review: population, index tests, reference standard and target condition.

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

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

Table 1: Description of review questions

Chapter or section	Type of review	Review question	Outcomes
Evidence reviews for diagnostic test accuracy in investigation for women presenting with heavy menstrual bleeding	Diagnostic	What is the diagnostic accuracy of ultrasound and hysteroscopy for investigation of women presenting with heavy menstrual bleeding?	• Sensitivity • Specificity • Positive likelihood ratio (LR+) • Negative likelihood ratio (LR-) • Area under the curve (AUC) if meta-analysis can be conducted • Patient satisfaction and acceptability of the test, including pain score
Evidence reviews for diagnostic test accuracy in investigation for women presenting with heavy menstrual bleeding	Diagnostic	What is the most clinically effective imaging strategy for diagnosing adenomyosis in women with heavy menstrual bleeding?	• Sensitivity • Specificity • LR+ • LR-

Chapter or section	Type of review	Review question	Outcomes
			• AUC if meta-analysis can be conducted • Patient satisfaction and acceptability of the test, including pain score
Evidence reviews for management of heavy menstrual bleeding	Intervention	What is the most clinically and cost-effective treatment (pharmacological/surgical) for heavy menstrual bleeding in women with: • suspected or diagnosed fibroids • suspected or diagnosed adenomyosis • no identified pathology?	• Reduction in blood loss • Quality of life • Patient satisfaction • Adverse events

AUC: area under the curve; LR+: positive likelihood ratio; LR-: negative likelihood ratio

Searching for evidence

Clinical search literature

Systematic literature searches were undertaken to identify all published clinical evidence relevant to the review questions on 13th October 2016 (Diagnosis question) and 23rd November 2016 (Management question).

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. Where possible, searches were restricted to retrieve only articles published in English. All searches were conducted in MEDLINE, Embase and The Cochrane Library.

Any studies added to the databases after this date (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 highly relevant papers, analysing search strategies in other systematic reviews and asking the group members to highlight any additional studies. The questions, the study types applied, the databases searched and the years covered can be found in Appendix E in each evidence review chapter.

Searching for grey literature or unpublished literature was not undertaken. Searches for electronic, ahead-of-print publications were not routinely undertaken unless indicated by the guideline committee. All references suggested by stakeholders at the scoping consultation were initially considered.

Health economics search literature

A global search of economic evidence was undertaken in December 2016. The following databases were searched:

- MEDLINE (Ovid)
- EMBASE (Ovid)
- Cochrane Central Register of Controlled Trials (CCTR)
- HTA database (HTA)
- NHS Economic Evaluations Database (NHS EED).

Further to the database searches, the committee was contacted with a request for details of relevant published and unpublished studies of which they may have knowledge; reference lists of key identified studies were also reviewed for any potentially relevant studies. Finally, the NICE website was searched for any recently published guidance relating to heavy menstrual bleeding that had not been already identified via the database searches.

The search strategy for existing economic evaluations combined terms capturing the target condition (heavy menstrual bleeding) and, for searches undertaken in MEDLINE, EMBASE and CCTR, 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). Conference abstracts were considered for inclusion from 1st January 2014, as high-quality studies reported in abstract form before 2014 were expected to have been published in a peer-reviewed journal. Full details of the search strategies are presented in Appendix E of each evidence review chapter.

Call for evidence

No call for evidence was made.

Reviewing clinical evidence

Systematic review process

The evidence was reviewed following these steps.

- Potentially relevant studies were identified for each review question from the relevant search results by reviewing titles and abstracts. Full papers were then obtained.
- Full papers were reviewed against pre-specified inclusion and exclusion criteria in the review protocols (in 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 F of each evidence review chapter).
- Relevant studies were critically appraised using the appropriate checklist as specified in the NICE guidelines manual (NICE 2014).
- Summaries of evidence were generated by outcome (included in the relevant review chapters) and were presented in committee meetings.

- Randomised studies: meta-analysis was carried out where appropriate and results were reported in GRADE profiles (for intervention reviews).
- Diagnostic studies: data were presented individually by study as measures of diagnostic test accuracy (sensitivity and specificity, positive and negative likelihood ratios) and were presented in modified GRADE profiles.

To assure quality of the study identification, a 10% sample of all the titles and abstracts for each review question were assessed for possible inclusion by a second independent reviewer. Possible discrepancies were resolved by discussion between the two reviewers.

All drafts of reviews were checked by a second reviewer. Any discrepancies were resolved by discussion between the 2 reviewers.

Type of studies and inclusion/ exclusion criteria

Systematic reviews (SRs) with meta-analyses were considered the highest quality evidence to be selected for inclusion.

For the review on the management of heavy menstrual bleeding, randomised controlled trials (RCTs) were prioritised for inclusion because they are considered the most robust study design to estimate the true effect of the interventions. RCTs with less than 10 participants in any intervention arm were excluded.

For the diagnostic test accuracy reviews, studies were included in which the index test and the reference standard were compared in the same individual and in which 2x2 tables could be constructed. The study designs considered for inclusion included test and treat RCTs, cross-sectional studies, and prospective cohort studies. Case-control studies were excluded.

Conference abstracts, posters, letters, editorials, comment articles, unpublished studies and studies not in the English language were excluded. Narrative reviews were also excluded, but individual references were checked for inclusion.

The inclusion and exclusion of studies was based on the review protocols, which can be found in Appendix A of each evidence review chapter. Excluded studies and the reasons for their exclusion are listed in Appendix I of each evidence review chapter. In addition, the guideline committee was consulted about any uncertainty regarding inclusion or exclusion.

Methods of combining evidence

Data synthesis for intervention review

Pairwise meta-analysis

Pairwise meta-analysis was conducted whenever it could be robustly performed to combine the results of studies 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 continuous outcomes, measures of central tendency (mean) and variation (standard deviation) are required for meta-analysis. Data for continuous outcomes

(such as health-related quality of life score or length of hospital stay) were analysed using an inverse variance method for pooling weighted mean differences. When the only evidence was based on studies summarising results by presenting medians (and interquartile ranges) or only p values were given, this information was assessed in terms of the study's sample size, and was included in the GRADE tables without calculating the relative or absolute effects. Consequently, aspects of quality assessment, such as imprecision of effect, could not be assessed for evidence of this type.

Forest plots were generated to visually present the results.

Statistical heterogeneity was assessed by visually examining the forest plots (please see Appendix H of each evidence review chapter) and by considering the chi-squared test for significance at $p<0.1$ or an I^2 squared inconsistency statistic (with an I^2 value of more than 50% indicating considerable heterogeneity). Where considerable heterogeneity was present, predefined subgroup analyses were performed.

Network meta-analysis

As is the case for ordinary pairwise meta-analysis, network meta-analysis (NMA) may be conducted using either fixed or random effect models. A fixed effect model typically assumes that there is no variation in relative effects across trials for a particular pairwise comparison and any observed differences are solely due to chance. For a random effects model, it is assumed that the relative effects are different in each trial but that they are from a single common distribution. The variance reflecting heterogeneity is often assumed to be constant across trials.

For continuous outcomes, where standard errors (SEs) could not be calculated from the data, we imputed them from other studies that reported measures of uncertainty/variance, using the median standard deviation (SD) of other study arms in the analysis that used the same treatment.

In a Bayesian analysis, for each parameter the evidence distribution is weighted by a distribution of prior beliefs. The Markov Chain Monte Carlo (MCMC) algorithm was used to generate a sequence of samples from a joint posterior distribution of 2 or more random variables and is particularly well adapted to sampling the treatment effects (known as a posterior distribution) of a Bayesian network. A non-informative prior distribution was used to maximise the weighting given to the data and to generate the posterior distribution for each log odds ratio (OR), log mean ratio (MR) or mean difference (MD) of interest in the networks. We used the median of the distribution as our point estimate and the centiles provided the 95% Credible Intervals (CrIs).

Non-informative priors were used that were normally distributed with a mean of 0 and SD of 100. However, for discontinuation due to adverse events, as there was sparse data on a number of treatments, we investigated whether the use of informative priors generated from empirical data would give a more stable between-study variance (Turner 2012).

For the analyses, a series of 40,000 burn-in simulations were run to allow the posterior distributions to convergence and then a further 100,000 simulations were run to produce the outputs. Convergence was assessed by examining the history, autocorrelation and Brooks-Gelman-Rubin plots.

Goodness-of-fit of the model was also estimated by using the posterior mean of the sum of the deviance contributions for each item by calculating the residual deviance and deviance information criteria (DIC). If the residual deviance was close to the number of unconstrained data points (the number of trial arms in the analysis) then the model was explaining the data at a satisfactory level. The choice of a fixed effect or random effects model can be made by comparing their goodness-of-fit to the data.

Incoherence in NMA between direct and indirect evidence can be assessed in closed treatment loops within the network. These closed treatment loops are regions within a network where direct evidence is available on at least 3 different treatments that form a closed 'circuit' of treatment comparisons (for example, A versus B, B versus C, C versus A). If closed treatment loops existed then discrepancies between direct and indirect evidence was assessed for each loop using node-splitting (van Valkenhoef 2016). The outputs of the NMA were as follows.

- Treatment specific log ORs, log MRs and MDs with their 95% CrIs were generated for every possible pair of comparisons by combining direct and indirect evidence in each network.
- The probability that each treatment is ranked within the best 3 or worst 3 treatments, based on the proportion of Markov chain iterations in which the treatment effect for an intervention is ranked best, second best and so forth. This was calculated by taking the treatment effect of each drug compared to placebo and counting the proportion of simulations of the Markov chain in which each intervention had the highest treatment effect.
- The ranking of treatments compared to the reference treatment (typically placebo or levonorgestrel-releasing intrauterine system (LNG-IUS) presented as median rank and its 95% CrI.

One of the main advantages of the Bayesian approach is that the method leads to a decision framework that supports decision making. The Bayesian approach also allows the probability that each intervention is best for achieving a particular outcome, as well as its ranking, to be calculated.

We adapted a random effects model template for continuous and dichotomous data available from NICE Decision Support Unit (DSU) technical support document number 2: http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series (2391675).htm. This model accounts for the within-study correlation between treatment effects induced by multi-arm trials.

For further description of outcomes and the specific results of the NMA please see the evidence review chapter for the management of heavy menstrual bleeding.

Data synthesis for diagnostic test accuracy reviews

Diagnostic data and outcomes

Sensitivity, specificity, positive and negative likelihood ratios, and area under the curve (AUC) were used as outcomes for diagnostic test accuracy reviews in this guideline. These diagnostic accuracy parameters (with 95% CIs) were obtained from the studies or calculated by the technical team using data from the studies (see Table 2).

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 target condition. When specificity is high, a positive test result rules in the target 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 following cut-offs were used when summarising the levels of sensitivity or specificity for the guideline committee:

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

Positive and negative likelihood ratios are measures of the association between a test result and the target condition. A positive likelihood ratio (LR+) greater than 1 indicates a positive test result and is associated with having the disorder, whilst a negative likelihood ratio (LR-) less than 1 indicates a negative test result and is associated with not having the disorder. A high LR+ would indicate that the test is useful in ruling in the condition whereas a low LR- would indicate that the test is useful in ruling out the condition.

The following cut-offs were used when summarising the likelihood ratios for the guideline committee:

- very useful test: LR+ higher than 10.0, LR- lower than 0.1
- moderately useful test: LR+ 5.0 to 10.0, LR- 0.1 to 0.2
- not a useful test: LR+ lower than 5.0, LR- higher than 0.2.

Table 2: 2x2 table for calculating diagnostic test accuracy parameters

	Condition present (according to reference standard)	No condition (according to reference standard)	Total
Index test positive	True positive (TP)	False positive (FP)	TP+FP = total number of subjects positive index test result
Index test negative	False negative (FN)	True negative (TN)	FN+TN = total number of subjects with negative index test result
Total	TP+FN = total number of subjects with condition	FP+TN = total number of subject without condition	TP+FP+FN+TN = Total number of subjects in study
Calculations for diagnostic test accuracy parameters:			
Sensitivity = TP/(TP+FN) Specificity = TN/(TN+FP)		LR+ = sensitivity/(1-specificity) LR- = (1-sensitivity)/specificity	

FN: false negative; FP: false positive; LR+: positive likelihood ratio; LR-: negative likelihood ratio; TP: true negative; TP: true positive

Diagnostic meta-analysis

When data from 5 or more studies were available, a diagnostic meta-analysis was carried out by using statistical software STATA with metandi package (Harbord and Whiting 2009; Harbord 2008). The metandi package performs bivariate meta-analysis of sensitivity and specificity using a generalised linear mixed model approach.

Forest plots and hierarchical summary receiver operating characteristic (HSROC) plots were created to visually present the results.

Appraising the quality of evidence

Intervention reviews

Pairwise analysis

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

For intervention reviews, the evidence for outcomes from the included RCTs were evaluated and presented using GRADE, which was developed by the international GRADE working group.

The 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/economic 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 patients with events divided by sum of the number of completers) for binary outcomes. 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 selection of outcomes for each review question was decided when each review protocol was discussed with the guideline committee, and was informed by committee discussion and by key papers, for example, previous NMAs. The systematic review by Herman (2016) describing the outcomes used in published systematic reviews and RCTs was also used to ensure all the main primary and secondary outcomes reported in trials were considered.

The evidence for each outcome in the intervention reviews was examined separately for the quality elements listed and defined in Table 3. Each element was graded using the quality levels listed in Table 4.

The main criteria considered in the rating of these elements are discussed below. Footnotes were used to describe reasons for grading a quality element as having serious or very serious limitations. The ratings for each component were summed to obtain an overall assessment for each outcome (Table 5).

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

Quality element	Description
Risk of bias (study limitations)	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.

Quality element	Description
Indirectness	Indirectness refers to differences in study population, intervention, comparator and outcomes between the available evidence and the review question, or recommendation made, such that the effect estimate is changed. This is also related to applicability or generalisability of findings.
Imprecision	Results are imprecise when studies include relatively few patients and 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.
Publication bias	Publication bias is a systematic underestimate or an overestimate of the underlying beneficial or harmful effect due to the selective publication of studies.

Table 4: Levels of quality elements in GRADE

Levels of quality elements in GRADE	Description
None/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 5: 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 overestimated.

Risk of bias in intervention studies was assessed using the Cochrane Risk of Bias Tool ((see Appendix H in the NICE guidelines manual 2014).

The possible sources of bias in RCTs in the Cochrane risk of bias tool fit with these 5 categories:

- selection bias
- performance bias
- attrition bias
- detection bias

- reporting bias

It should be noted that 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 this poor design will impact on the estimation of the intervention effect.

More details about this tool can be found here:
http://cobe.paginas.ufsc.br/files/2014/10/Cochrane.RCT_.pdf

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). When outcomes derived from a single study 'no inconsistency' was used when assessing this domain, as per GRADE methodology (Santesso 2016). .

Heterogeneity was assessed by calculating the I^2 statistic for the meta-analysis. An I^2 of more than 50% was considered to indicate high heterogeneity. When high heterogeneity was observed, possible reasons for it were explored and subgroup analyses were performed as pre-specified in the review protocol.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE by 1 or 2 levels for the domain of inconsistency, depending on the extent of heterogeneity in the results.

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.

Assessing imprecision and clinical significance in intervention reviews

Imprecision in guidelines concerns whether the uncertainty (CI) around the effect estimate means that it is not clear whether there is a clinically important difference between interventions or not (that is, whether the evidence would clearly support one recommendation or appear to be consistent with several different types of recommendations). Therefore, imprecision differs from the 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 actually is. 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 this procedure to be repeated. The larger the trial, the smaller the 95% CI and the more certain the effect estimate.

Imprecision in the evidence reviews was assessed by considering whether the width of the 95% CI of the effect estimate was relevant to decision-making, taking each

outcome in isolation. This is explained in Figure 1, which considers a positive outcome for the comparison of treatment A versus treatment B. Three decision-making zones can be identified, bounded by the thresholds for clinical importance (minimally important difference, MID) for benefit and for harm. The MID for harm for a positive outcome means the threshold at which drug A is less effective than drug B by an amount that is clinically important to patients (favours B).

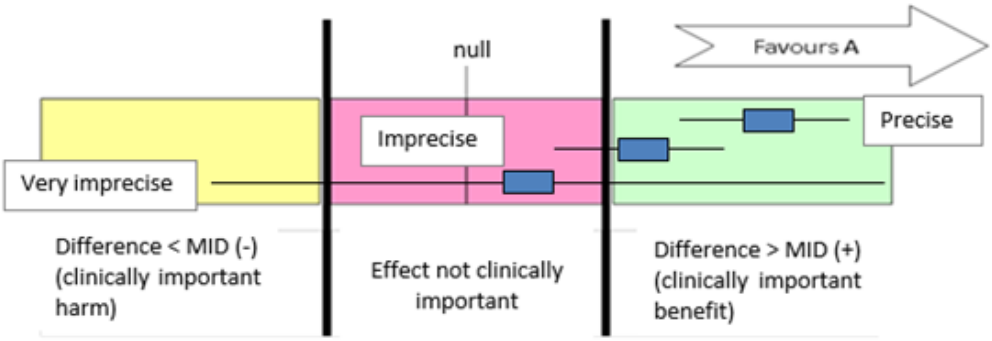
When the CI of the effect estimate is wholly contained in 1 of the 3 zones (for example, clinically important benefit), we are not uncertain about the size and direction of effect (whether there is a clinically important benefit, or the effect is not clinically important, or there is a clinically important harm), so there is no imprecision.

When a wide CI lies partly in each of 2 zones, it is uncertain in which zone the true value of effect estimate lies and therefore there is uncertainty over which decision to make (based on this outcome alone). The CI is consistent with 2 possible decisions and so this is considered to be imprecise in the GRADE analysis and the evidence is downgraded by 1 level ('serious imprecision').

If the CI of the effect estimate crosses into 3 zones, this is considered to be very imprecise evidence because the CI is consistent with 3 possible clinical 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 the CI is in, or partially in, a clinically important zone, requires the committee to estimate an MID or to say whether they would make different decisions for the 2 confidence limits.

Figure 1: Illustration of precise, imprecise and very imprecise evidence based on the confidence interval of outcomes in forest plots



Minimally important differences

The literature was searched for established MIDs for the selected outcomes in the evidence reviews, such as blood loss or quality of life. In addition, the committee was asked whether they were aware of any acceptable MIDs in the clinical community.

If no published or acceptable MIDs were identified, the committee considered whether it was clinically acceptable to use the GRADE default MID to assess imprecision. For binary outcomes clinically important thresholds for a RR of 0.8 and 1.25 respectively were used (due to the statistical distribution of this measure this means that this is not a symmetrical interval). This default MID was used for all the

binary outcomes in the intervention evidence reviews as a starting point and decisions on clinical importance were then considered based on the absolute risk difference. For continuous outcomes GRADE default MIDs were half of the median SD of the control group.

Network meta-analysis

For the NMAs, quality was assessed by looking at risk of bias across the included evidence (using the standard GRADE approach for this domain), as well as heterogeneity and incoherence.

The following limits of the upper 95% CrI for between-study standard deviation were used to assess heterogeneity for NMAs in which a random effects model was used:

- less than 0.3 – low heterogeneity
- 0.3 to 0.6 – moderate heterogeneity
- more than 0.6 to 0.9 – high heterogeneity
- more than 0.9 to 1.2 – very high heterogeneity.

Where significant incoherence was found it was considered to be serious when the direction of effect for both direct and indirect estimates was the same (for example, an OR of greater than 1 in both the direct and indirect estimates), and very serious when the direction of effect was different (for example, an OR of greater than 1 for the direct estimate but less than 1 for the indirect estimate).

For fixed-effect NMAs that did not model heterogeneity, or for networks in which incoherence could not be assessed as no closed treatment loops existed, these criteria were not considered to impact the quality of evidence.

Diagnostic reviews

Adapted GRADE methodology

The GRADE toolbox is designed for RCTs and observational studies, but we adapted the quality assessment elements and outcome presentation for diagnostic test accuracy reviews. For example, the GRADE clinical evidence tables were modified to include the most appropriate measures of diagnostic accuracy (sensitivity, specificity, and likelihood ratios).

The evidence for each outcome in the diagnostic test accuracy reviews was examined separately for the quality elements listed and defined in Table 6. Each element was graded using the quality levels listed in Table 4.

The main criteria considered in the rating of these elements are discussed below. Footnotes were used to describe reasons for grading a quality element as having serious or very serious limitations. The ratings for each component were summed to obtain an overall assessment for each outcome (Table 5).

Table 6: Description of the elements in GRADE and how they are used to assess the quality for diagnostic accuracy reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in the study design and implementation may bias the estimates of the diagnostic accuracy. High risk of bias for the majority of the evidence decreases confidence in the estimate of

Quality element	Description
	the effect. Diagnostic accuracy studies are not usually randomised and therefore would not be downgraded for study design from the outset and start as high level evidence.
Inconsistency	Inconsistency refers to an unexplained heterogeneity of test accuracy measures, for example sensitivity or specificity, between studies.
Indirectness	Indirectness refers to differences in study population, index tests, reference standards and outcomes between the available evidence and the review question.
Imprecision	Results are considered imprecise when studies include relatively few patients and the confidence intervals were wide. Imprecision results if the confidence interval includes the clinically important threshold.

Assessing risk of bias and indirectness in diagnostic test accuracy reviews

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

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

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

More details about this tool can be found here: http://www.bristol.ac.uk/social-community-medicine/projects/quadas/quadas-2/

Assessing inconsistency in diagnostic test accuracy reviews

Inconsistency refers to the unexplained heterogeneity of the results in meta-analysis. When estimates of diagnostic accuracy parameters vary widely across studies (that is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only applicable when statistical meta-analysis was conducted (that is, results from different studies were pooled). However, 'no inconsistency' is nevertheless used to describe this quality assessment in the GRADE profiles for outcomes from single studies.

For the diagnostic test accuracy reviews, the heterogeneity of the pooled result was assessed by visually inspecting the size of the 95% CI prediction region in the HSROC plot. When considerable heterogeneity was observed, possible reasons for it were explored and subgroup analyses were performed, when possible, according to the pre-specified subgroups in the review protocol.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE by 1 or 2 levels for the domain of inconsistency, depending on the extent of heterogeneity in the results.

Assessing indirectness in diagnostic test accuracy reviews

Indirectness in diagnostic test accuracy 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 (see **Error! Reference source not found.**):

- patient selection
- index test
- reference standard.

Assessing imprecision and clinical significance in diagnostic test accuracy reviews

In diagnostic accuracy measures, it was first considered whether sensitivity, specificity, positive likelihood ratios or negative likelihood ratios would be given more weight in the decision-making process. If one measure was given more importance than the other, then imprecision was rated on this statistical measure using the following MID thresholds:

- sensitivity and specificity
  - high: more than 90%
  - moderate: 75-90%
  - low: less than 75%
- positive likelihood ratio:
  - very useful test: more than 10
  - moderately useful test: 5-10
  - not a useful test: less than 5
- negative likelihood ratio:
  - very useful test: less than 0.1
  - moderately useful test: 0.1 to 0.2
  - not a useful test: more than 0.2.

Evidence statements

Evidence statements are summary statements that are presented after the GRADE profiles, summarising 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 and encompass the following key features of the evidence:

- the quality of the evidence (GRADE rating)
- the number of studies and the number of participants for a particular outcome
- a brief description of the participants
- the clinical significance of the effect and an indication of its direction (for example, if a treatment is clinically significant [beneficial or harmful] compared with another, or whether there is no clinically significant difference between the tested treatments).

Reviewing economic evidence

Inclusion and exclusion of economic studies

The titles and abstracts of papers identified through the searches were independently assessed for inclusion using pre-defined eligibility criteria defined 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
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
Cost-of-illness type studies
Conference papers pre January 2014

Once the screening of titles and abstracts was complete, full versions of the selected papers were acquired for assessment. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for this search on economic evaluations is presented in the Health Economics Chapter.

The quality of evidence was assessed using the economic evaluations checklist as specified in the NICE guidelines manual (NICE 2014). Quality assessments of included studies and data extraction tables are provided in Appendix B of the evidence review chapters. The excluded economic studies list is presented in the management evidence review chapter.

Health economic modelling

The aims of the health economic input to the guideline were to inform the guideline committee of potential economic issues related to the diagnosis and management of heavy menstrual bleeding 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; recommendations which – while nevertheless cost-effective – might have a large impact on Clinical Commissioning Group or Trust finances and so need special attention.

The guideline committee prioritised a single economic model on diagnosis and management where it was thought that economic considerations would be particularly important in formulating recommendations and a review of the health economic literature was undertaken. This model covered multiple review questions, as a complete health economic analysis of the treatment pathway required consideration of all possible combinations of diagnostic strategy and treatment strategy together.

Cost effectiveness criteria

NICE's report Social value judgements: principles for the development of NICE guidance (https://www.nice.org.uk/media/default/about/what-we-do/research-and-development/social-value-judgements-principles-for-the-development-of-nice-guidance.pdf) 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 either 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 significant 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 'Consideration of economic benefits and harms' heading of the relevant sections.

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 their 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 'Recommendations and link to evidence' headings within each chapter.

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

Validation process

This guidance is subject to a 6-week public consultation and feedback as part of the quality assurance and peer review of the document. All comments received from registered stakeholders are responded to in turn and posted on the NICE website at publication. For further details please refer to the NICE guidelines manual (NICE 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 the NICE guidelines manual (NICE 2014).

Funding

The NGA was commissioned by NICE to undertake the work on this guideline.

References

Harbord 2008

Harbord, R., metandi: Stata module for meta-analysis of diagnostic accuracy, Statistical Software Components, Boston College Department of Economics, Revised 15 Apr 2008.

Harbord and Whiting 2009

Harbord, R. M., Whiting, P., metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression, Stata Journal, 9, 211-29, 2009

Herman 2016

Herman, M.C., Penninx, J.P., Geomini, P.M., Mol, B.W., Bongers, M.Y., Choice of primary outcomes evaluating treatment for heavy menstrual bleeding, BJOG: An International Journal of Obstetrics & Gynaecology, 1, 123(10), 1593-8, 2016

NICE 2014

National Institute for Health and Care Excellence (NICE), NICE guidelines manual, 2014 [online, accessed 8th June 2017]

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.

Turner 2012

Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., Higgins, J. P. T., Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews, International Journal of Epidemiology, 41, 818-827, 2012

Van Valkenhoef 2016

Van Valkenhoef, G., Dias, S., Ades, A. E., Welton, N. J., Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis, Research Synthesis Methods, 7, 80-93, 2016