National Institute for Health and Care Excellence

FINAL

Menopause (update)

Supplement 1: methods

NICE guideline NG23

Methods

November 2024

FINAL

These supplements were developed by NICE

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 All NICE guidance is subject to regular review and may be updated or withdrawn.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

© NICE 2024. All rights reserved. Subject to Notice of rights.

ISBN: 978-1-4731-6569-4

Contents

Development of the guideline

What this guideline covers

The guideline was partially updated in 2023 and the methods outlined in this methods document relate only to the following sections of the updated guideline:

- Managing troublesome menopausal symptoms, for the following situations only:
  - Cognitive behavioural therapy to manage symptoms associated with the menopause
  - Interventions to manage genitourinary symptoms associated with the menopause.
- Long-term benefits and risks of hormone replacement therapy
  - Venous thromboembolism for people with early menopause (40 to 44 years)
  - Cardiovascular disease
  - Type 2 diabetes for people with early menopause (40 to 44 years)
  - Breast cancer
  - Endometrial cancer
  - Ovarian cancer
  - Osteoporosis for people with early menopause (40 to 44 years)
  - Dementia
  - Loss of muscle mass and strength for people with early menopause (40 to 44 years)
  - All-cause mortality

What this guideline does not cover

The following sections of the guideline were not updated in 2023 and were developed using methods outlined in the 2015 full guideline document:

- Individualised care
- Diagnosis of perimenopause and menopause
- Managing short-term menopausal symptoms (other than cognitive behavioural therapy or treatments for genitourinary symptoms)
- Review and referral
- Starting and stopping hormone replacement therapy
- Long-term benefits and risks of hormone replacement therapy for people with menopause at 45 years or older:
  - Venous thromboembolism
  - Type 2 diabetes
  - Loss of muscle mass and strength
- Diagnosing and managing premature ovarian insufficiency

Methods

This guideline was developed using the methods described in the 2018 NICE guidelines manual.

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

Developing the review questions and outcomes

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

The review questions were based on the following framework:

- population, intervention, comparator and outcome (PICO) for reviews of interventions

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

The review questions and evidence reviews corresponding to each question are summarised below.

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

Evidence review	Review question	Type of review
[A] Cognitive behavioural therapy	What is the effectiveness of cognitive behavioural therapy for managing symptoms associated with the menopause?	Intervention
[B1] Managing genitourinary symptoms (Network Meta Analysis)	What is the effectiveness of treatments such as local oestrogen, ospemifene, prasterone and transvaginal laser therapy for managing genitourinary symptoms associated with the menopause?	Intervention[1]
[B2] Managing genitourinary symptoms – breast cancer recurrence	Are treatments for managing genitourinary symptoms associated with the menopause safe for women with a personal history or high inherited risk of breast cancer?	Intervention
[G] Cardiovascular disease	What are the effects of hormone replacement therapy for menopausal symptoms on developing cardiovascular disease?	Intervention
[C] Breast cancer	What are the effects of hormone replacement therapy for menopausal symptoms on developing breast cancer?	Intervention
[D] Ovarian cancer	What are the effects of hormone replacement therapy for menopausal symptoms on developing ovarian cancer?	Intervention
[E] Endometrial cancer	What are the effects of hormone replacement therapy for menopausal symptoms on developing endometrial cancer?	Intervention
[F] Dementia	What are the effects of hormone replacement therapy for menopausal symptoms on developing dementia?	Intervention

Evidence review	Review question	Type of review
[H] All-cause mortality	What are the effects of hormone replacement therapy for menopausal symptoms on all-cause mortality?	Intervention
[I] Early menopause	What are the effects of hormone replacement therapy taken by women, non-binary and trans-masculine people with early menopause (aged 40 to 44) on all-cause mortality and developing: • venous thromboembolism • cardiovascular disease • type 2 diabetes • breast cancer • endometrial cancer • ovarian cancer • osteoporosis • dementia • loss of muscle mass and strength?	Intervention

[1]*Original health economic analysis conducted*

The COMET database was searched for core outcome sets relevant to this guideline. A core outcome set for genitourinary symptoms associated with menopause was identified (Lensen 2021) and used for evidence review B1. For the other reviews no core outcome sets were identified and therefore the outcomes were chosen based on committee discussions.

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

- Supplement 2 (Health economics)
- Supplement 3 (Acknowledgements)
- Supplements 4 to 16 (NMA supplementary files related to evidence review B1)
- Supplements 17 and 18 (data extractions related to evidence review C)
- Supplement 19 (absolute number calculations).

Searching for evidence

Scoping search

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

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. Limits to exclude animal studies, letters, editorials, news and conferences were applied where possible.

All the searches were conducted in the following databases: Medline, Embase, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Epistemonikos, International Network of Agencies for Health Technology Assessments (INAHTA) and HTA. For review questions related to CBT and Dementia, PsycInfo was also searched.

A multi-stranded approach was followed in the search for review questions G to I, by firstly combining the population, intervention and outcomes and secondly combining the population, intervention and applying study type filters, where appropriate. To manage the volume, the search for systematic reviews was limited to Epistemonikos and Cochrane Database of Systematic Reviews (CDSR).

Searches were run once for all reviews during development. 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), the International Network of Agencies for Health Technology Assessments (INAHTA), HTA and EconLit databases. Another single search, using the population search terms used in the evidence reviews combined with an economic evaluations search filter, was conducted in Medline, Embase, Cochrane Central Register of Controlled Trials (CENTRAL) and Cochrane Database of Systematic Reviews (CDSR). Where possible, searches were limited to studies published in English. A date limit was added to capture studies published from 2012 onwards. Limits to exclude animal studies, letters, editorials, news were applied where possible.

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

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 research 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. Further detail on appraisal of the evidence is provided below.
- Summaries of effectiveness evidence by outcome were presented in the corresponding evidence review and discussed by the committee.

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. Internal quality assurance processes included consideration of the outcomes of screening, study selection and data extraction and the committee reviewed the results of study selection and data extraction. The review protocol for each question specifies whether dual screening and study selection was undertaken for that particular question. Drafts of all evidence reviews were quality assured by a senior reviewer.

Type of studies and inclusion/exclusion criteria

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

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

For intervention reviews, randomised controlled trials (RCTs) were prioritised for inclusion because they are considered to be the most robust type of study design that could produce an unbiased estimate of intervention effects. Where there was insufficient evidence from RCTs to inform guideline decision making, non-randomised studies (NRS) were considered for inclusion. Sufficiency was judged taking into account the number, quality and sample size of RCTs, as well as outcomes reported and availability of data from subgroups of interest. When NRS were considered for inclusion, priority was given to controlled studies, with separate control groups that were not allocated on the basis of the outcome, that adjusted for relevant confounders or matched participants on important confounding domains.

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

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

Methods of combining evidence

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

Data synthesis for intervention studies

Pairwise meta-analysis

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

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

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

If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into RevMan5. If the control event rate was reported this was used to generate the absolute risk difference in GRADEpro. If multivariable analysis was used to derive the summary statistic but no adjusted control event rate was reported, no absolute risk difference was calculated. Where a study reported multiple adjusted estimates for the same outcome, the one that minimised the risk of bias due to confounding was chosen.

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

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

Data from RCTs and NRS, or from NRS with substantially different designs (i.e., cohort studies and case-control studies), that were theoretically possible to pool were entered into RevMan5 as subgroups based on study design. This was to take into account the likelihood of increased heterogeneity from studies with different design features and different approaches to appraising the quality of evidence based on study design (see appraising the quality of evidence: intervention studies below).

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

When case series were included, descriptive data from the studies were included and no further analysis was 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.

Treatment effects were modelled at the class level, grouping treatments of the same class together. Both fixed and random class models were generated. The fixed class model assumes the effect is the same for all treatments in the same class. The random class model: assumes there is heterogeneity between treatment effects within a class.

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 prior distribution was used to maximise the weighting given to the data and to generate the posterior distribution of the results.

For the analyses, a series of burn-in simulations were run to allow the posterior distributions to converge and then further simulations were run to produce the posterior 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.

Treatment specific posterior effects were generated for every possible pair of comparisons by combining direct and indirect evidence in each network. The probability that each treatment is best, 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 intervention compared to the reference treatment and counting the proportion of simulations of the Markov chain in which each intervention had the highest treatment effect.

The effect of potential treatment modifiers was also investigated. Models were developed to quantify the impact of the duration of the study, the method of analysis (per-protocol or intention-to-treat) and study-level bias. For the bias-adjusted models it was assumed that bias was proportional to the study size (with smaller studies tending to be more biased) and that in biased studies active interventions would tend to have larger effect sizes than inactive interventions.

The NMA work was undertaken by the NICE Guidelines Technical Support Unit, University of Bristol (TSU). They adapted standard fixed and random effects models available from NICE Decision Support Unit (DSU) technical support document number 2: http://nicedsu.org.uk/wpcontent/uploads/2017/05/TSD2-General-meta-analysis-corrected-2Sep2016v2.pdf

To determine if there is evidence of inconsistency, the selected consistency model (fixed or random effects) was compared to an "inconsistency", or unrelated mean effects, model. We performed further checks for evidence of inconsistency through node-splitting.

Appraising the quality of evidence

Intervention studies

Pairwise meta-analysis

GRADE methodology for intervention reviews

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

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

The selection of outcomes for each review question was agreed during development of the associated review protocol in discussion with the committee. The evidence for each outcome was examined separately for the quality elements summarised in Table 2. Criteria considered in the rating of these elements are discussed below. Each element was graded using the quality ratings summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a 'serious' or 'very serious' quality issue. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

The initial quality rating was based on the study design: RCTs and NRS assessed by ROBINS-I start as 'high' quality evidence, other non-randomised studies (such as case-control studies) start 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 one or two 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 (RoB 2; see Appendix H in Developing NICE guidelines: the manual).

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

- risk of bias arising from the randomisation process
- risk of bias due to deviations from the intended interventions
- risk of bias due to missing outcome data
  - risk of bias due to measurement of the outcome
  - risk of bias in selection of the reported result.

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.

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 2022).

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

For non-randomised controlled studies or cohort studies the ROBINS-I checklist was used and for case-control studies the CASP case control checklist was used (see Appendix H in Developing NICE guidelines: the manual). For studies reporting individual participant data (meta-analyses of randomised controlled trials), the checklist published by Tierney et al was used for assessing risk of bias (Tierney 2015).

Assessing inconsistency in intervention reviews

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

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

When no plausible explanation for the serious or very serious heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency and the meta-analysis was re-run using the Der-Simonian and Laird method with a random effects model and this was used for the final analysis. Where very serious heterogeneity was observed (I-squared value > 80%), the single-study point estimates were additionally reported in the GRADE table.

Assessing indirectness in intervention reviews

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

Assessing imprecision and importance in intervention reviews

Imprecision in GRADE methodology refers to uncertainty around the effect estimate and whether or not there is an important difference between interventions (that is,

whether the evidence clearly supports a particular recommendation or appears to be consistent with several candidate recommendations). Therefore, imprecision differs from other aspects of evidence quality because it is not concerned with whether the point estimate is accurate or correct (has internal or external validity). Instead, it is concerned with uncertainty about what the point estimate actually represents. This uncertainty is reflected in the width of the CI.

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

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

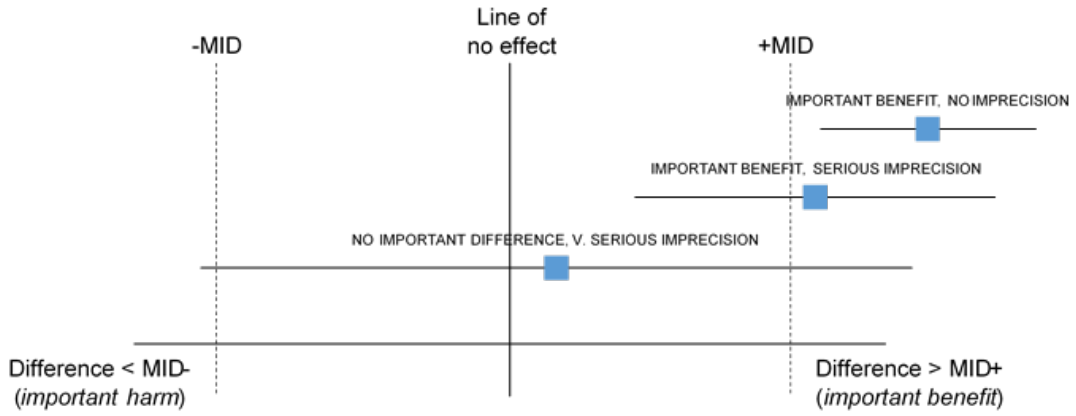
When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

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

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

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

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



MID, minimally important difference

Defining minimally important differences for intervention reviews

The committee was asked whether there were any recognised or acceptable MIDs in the published literature and community relevant to the review questions under consideration. The committee was aware of one published MID for the outcome 'vasomotor symptoms, distress, or bother'. This outcome was measured with the Hot Flush Rating Scale (HFRS), with a MID of 2 (Ayers 2012), and was applicable to the cognitive behavioural therapy evidence review. Otherwise, the committee was not aware of any MIDs that could be used for the guideline.

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

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

The same thresholds were used as default MIDs for decision making in the guideline for the dichotomous outcomes in reviews A and B1.

For the dichotomous outcomes in reviews B2, C, D, E, F, G, H and I, statistical significance was used instead of the default MIDs for dichotomous outcomes when assessing clinical importance and therefore decision making. This was because the outcomes were considered very serious and the committee agreed that any significant difference was meaningful enough to influence decision making. Statistical significance was determined when the 95% confidence interval for the effect excluded the null value.

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

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.

Network meta-analysis

For the NMAs, quality was assessed by looking at risk of bias across the included evidence using the Cochrane Risk of Bias Tool for Randomized Controlled Trials, as well as heterogeneity and consistency (also called incoherence).

The following limits of the upper 95% credible interval (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

The consistency 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.

To determine if there is evidence of inconsistency, the selected consistency model (fixed or random effects) was compared to an "inconsistency", or unrelated mean effects, model. The latter is equivalent to having separate, unrelated, meta-analyses for every pairwise contrast, with a common variance parameter assumed in the case of random effects models. Further checks for evidence of inconsistency either through Bucher's method or node-splitting were undertaken. Bucher's method compares the direct and indirect estimates for a contrast in a loop (e.g., A-B-C) where the direct estimate of contrast B vs. C is compared to its corresponding indirect estimate, which is informed from the direct estimates of the other contrasts in the loop (A vs. B and A vs. C). This method was used to assess consistency in networks, where there was a single loop and the network contained sparse evidence with zero events, limiting the stability of the results of more sophisticated methods such as the node-splitting method. The node-splitting method allowed the direct and indirect evidence contributing to an estimate of a relative effect to be split and compared. The consistency checks were undertaken by the TSU.

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

Reviewing economic evidence

Titles and abstracts of articles identified through the economic literature searches were independently assessed for inclusion using the predefined inclusion and exclusion criteria

Inclusion and exclusion criteria for systematic reviews of economic evaluations

Inclusion and exclusion criteria

- If a study is rated as both 'Directly applicable' and with 'Minor limitations' then it was included in the guideline. A health economic evidence table was completed and it was included in the health economic evidence profile.
- If a study is rated as either 'Not applicable' or with 'Very serious limitations' then it was excluded from the guideline. If it is excluded then a health economic evidence table was not be completed and it was not be included in the health economic evidence profile.
- If a study is rated as 'Partially applicable', with 'Potentially serious limitations' or both then discretion was used over whether it should be included.

Where there is discretion

The health economist made a decision based on the relative applicability and quality of the available evidence for that question, in discussion with the guideline committee if required. The ultimate aim was to include health economic studies that are helpful for decision-making in the context of the guideline and the current NHS setting.

The health economist was guided by the following hierarchies.

Setting:

- UK NHS (most applicable).
- OECD countries with predominantly public health insurance systems (for example, France, Germany, Sweden).
- OECD countries with predominantly private health insurance systems (for example, Switzerland).
- Studies set in non-OECD countries or in the USA were excluded before being assessed for applicability and methodological limitations.

Health economic study type:

- Cost utility analysis (most applicable).
- Other type of full economic evaluation (cost benefit analysis, cost effectiveness analysis, cost–consequences analysis).
- Comparative cost analysis.
- Non-comparative cost analyses including cost-of-illness studies were excluded before being assessed for applicability and methodological limitations.

Year of analysis:

- The more recent the study, the more applicable it will be.
- Studies published in 2005 or later (including any such studies included in the previous guideline(s)) but that depend on unit costs and resource data entirely or predominantly from before 2005 was rated as 'Not applicable'.
- Studies published before 2005 (including any such studies included in the previous guideline(s)) was excluded before being assessed for applicability and methodological limitations.

Appraising the quality of economic evidence

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

Economic modelling

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

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

B1 - What is the effectiveness of treatments such as local oestrogen, ospemifene, prasterone and transvaginal laser therapy for managing genitourinary symptoms associated with the menopause?

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

Cost effectiveness criteria

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

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

The committee's considerations of cost effectiveness are discussed explicitly in the section 'The committee's discussion and interpretation of the evidence' and heading 'Cost effectiveness and resource use'.

Details of the cost effectiveness analyses undertaken for the guideline are presented in Supplement 2 (Health economics).

Developing recommendations

Guideline recommendations

Recommendations were drafted on the basis of the committee's interpretation of the available evidence, taking account of the quality of the evidence, 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 committee also considered statistical significance as an important aspect of decision making for all outcomes alongside the GRADE quality rating.

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.

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 and NICE's Research recommendations process and methods guide.

Validation process

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

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.

References

Ayers 2012

Ayers B, Smith M, Hellier J et al. (2012) Effectiveness of group and self-help cognitive behavior therapy in reducing problematic menopausal hot flushes and night sweats (MENOS 2): a randomized controlled trial. Menopause (New York, N.Y.) 19(7): 749-759

Bradburn 2007

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

Hayden 2013

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

Higgins 2022

Higgins JPT, Thomas J, Chandler J, et al. (editors; updated 2022). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 Cochrane. Available from www.training.cochrane.org/handbook (accessed 15 June 2023)

Lensen 2021

Lensen S, Bell RJ, Carpenter JS et al. A core outcome set for genitourinary symptoms associated with menopause: the COMMA (Core Outcomes in Menopause) global initiative. Menopause 28(8):p 859-866, August 2021.

McGowan 2016

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

NICE 2018

National Institute for Health and Care Excellence (NICE) (2014) NICE Policy on conflicts of interest (updated 2017). Available from https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/declaration-of-interests-policy.pdf (accessed 15 June 2023)

Santesso 2016

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

Tierney 2015

Tierney JF, Vale C, Riley R, Smith CT, Stewart L, Clarke M, et al. (2015) Individual Participant Data (IPD) Meta-analyses of Randomised Controlled Trials: Guidance on Their Use. PLoS Med 12(7): e1001855)