National Institute for Health and Care Excellence

Draft for consultation

Fertility problems: assessment and treatment

NICE guideline: methods

NICE guideline NGxxx

Methods

September 2025

Draft for Consultation

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN:

Contents

Development of the guideline

Remit

This guideline will update the following National Institute for Health and Care Excellence (NICE) clinical guideline: Fertility problems: assessment and treatment (NICE CG156).

To see “What this guideline covers” and “What this guideline does not cover” please see the guideline scope.

We have developed sections of this guideline in stages, and the methods described in this document apply for evidence reviews covered in stage 1 and 2, see Table 1.

There are a number of potential future topics for update which are listed in the scope. These are:

- What is the clinical and cost effectiveness of ovulation induction strategies in people with polycystic ovary syndrome (PCOS)?
- What is the clinical and cost effectiveness of hysteroscopic septum resection compared to expectant management for people with fertility problems associated with a septate uterus?
- What is the effectiveness and safety of different embryo or blastocyst transfer strategies in relation to both: number of embryos, timing of transfer?
- What is the effectiveness and safety of different regimens of frozen embryo transfer?
- What is the effect of ART on obstetric risk in women undergoing fertility treatment?
- What is the long-term safety of in vitro fertilisation (IVF) with or without intracytoplasmic sperm injection (ICSI) in women with fertility problems and their children conceived through ART?

Methods

This guideline was developed using the methods described in Developing NICE guidelines: the manual as outlined in the table below.

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 frameworks:

- population, intervention, comparator and outcome (PICO) for reviews of interventions
- prognostic reviews and review of prediction model performance – using population, (presence or absence of) prognostic factors, and outcome
- single-arm proportional meta-analyses of case series – using population, intervention and outcome

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

We are updating this guideline in a staggered approach, and Table 1 indicates the review questions that have been allocated to stage 1 and stage 2. Stage 2 focuses on male factor fertility problems.

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

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

Evidence review	Review question	Type of review	Stage
A Ovarian reserve testing	What is the association between markers of ovarian reserve and: the likelihood of spontaneous conception, the response to fertility treatment, and the outcome of fertility treatment?	Prognostic	1
B Subclinical hypothyroidism	What is the clinical and cost effectiveness of treating subclinical hypothyroidism for female factor fertility problems?	Intervention	1
C Screening hysteroscopy	What is the effectiveness of hysteroscopy (with or without treatment of any detected uterine cavity abnormalities) on reproductive outcomes for people with female factor fertility problems?	Intervention	1

Evidence review	Review question	Type of review	Stage
D Endometrial receptivity testing	What is the clinical and cost effectiveness of tests for endometrial receptivity (including gene expression analysis and microbiological analysis) as a treatment add-on for people undergoing fertility treatment?	Intervention	1
E Ovulation induction strategies for hypogonadotropic hypogonadism	What is the clinical and cost effectiveness of ovulation induction strategies in people with hypogonadotropic hypogonadism?	Intervention	1
F Cabergoline for hyperprolactinaemia	What is the clinical and cost effectiveness of cabergoline for fertility problems associated with hyperprolactinaemic amenorrhoea or oligomenorrhea?	Intervention	1
G Tubal surgery	What is the clinical and cost effectiveness of tubal surgery (as a standalone treatment) compared to expectant management or in vitro fertilisation (IVF) for fertility problems associated with tubal disease?	Intervention	1
H Surgery for hydrosalpinges before IVF	What is the clinical and cost effectiveness of surgery for hydrosalpinges prior to assisted reproductive technology (ART), relative to standard ART without prior surgical optimisation, for people with tubal disease?	Intervention	1
I Tubal catheterisation	What is the likelihood of spontaneous conception when tubal catheterisation/cannulation is used for the treatment of proximal tubal obstruction?	Single-arm proportional meta-analysis of case series	1
J Fertility prediction models and IVF access	What is the predictive performance of clinical prediction models for assessing the chances of live birth for people with health-related fertility problems using: expectant management, intrauterine insemination (IUI), or IVF with or without intracytoplasmic sperm injection (ICSI)?	Prediction model performance[1]	1
K Assisted reproduction techniques for people with unexplained fertility	What is the clinical and cost effectiveness of ovarian stimulation, intrauterine insemination (IUI) with or without ovarian stimulation, IVF and expectant management for people	Intervention[1]	1

Evidence review	Review question	Type of review	Stage
problems, mild endometriosis, mild male factor fertility problems	with unexplained health-related fertility problems, mild endometriosis, and people with a single abnormal semen parameter? (NMA)		
L Intracytoplasmic sperm injection for non-male factor fertility problems	What is the effectiveness of intracytoplasmic sperm injection (ICSI) compared to standard in vitro fertilisation (IVF) in non-male factor fertility problems?	Intervention	1
M Advanced sperm selection techniques as a treatment add-on	What is the clinical and cost effectiveness of alternatives to standard sperm selection techniques as a treatment add-on for people undergoing fertility treatment?	Intervention	1
N Pre-implantation genetic testing for aneuploidy as a treatment add-on	What is the clinical and cost effectiveness of pre-implantation genetic testing for aneuploidy (PGT-A; with blastocyst stage biopsy and genome-wide analysis) as a treatment add-on for people undergoing fertility treatment?	Intervention	1
O Embryo selection guided by continuous time-lapse sequence as a treatment add-on	What is the clinical and cost effectiveness of embryo selection guided by continuous time-lapse monitoring (with or without artificial intelligence algorithms) as a treatment add-on for people undergoing fertility treatment?	Intervention	1
P Endometrial scratch as a treatment add-on	What is the clinical and cost effectiveness of endometrial scratch as a treatment add-on for people undergoing fertility treatment?	Intervention	1
Q Immune therapies as a treatment add-on	What is the clinical and cost effectiveness of immune therapies as a treatment add-on for people undergoing fertility treatment?	Intervention	1
R Fertility preservation	What is the success rate, and which factors affect the outcome, of fertility preservation for children and adults undergoing treatment for cancer and other conditions or situations which are likely to impair their fertility?	Single-arm proportional meta-analysis of case series	1
S Sperm DNA fragmentation	What is the clinical and cost effectiveness of treating DNA fragmentation (identified from screening ejaculated sperm) on	Intervention	2

Evidence review	Review question	Type of review	Stage
	reproductive outcomes for people with male factor fertility problems?		
T Y chromosome microdeletion	What is the association between Y chromosome microdeletions (positive AZF a, b, and c) and successful sperm retrieval in people with non-obstructive azoospermia?	Prognostic	2
U Hormone treatment for male factor fertility problems	What is the effectiveness of hormone treatment in male factor fertility problems?	Intervention	2
V Treatments for ejaculatory failure	What is the clinical and cost effectiveness of treatments for ejaculatory failure?	Intervention	2
W Surgical interventions for obstructive azoospermia	What is the clinical and cost effectiveness of surgical interventions for fertility problems associated with obstructive azoospermia?	Intervention	2
X Treatments for varicocele	What is the effectiveness of treatments for fertility problems associated with varicocele (including radiological embolisation and surgery)?	Intervention	2
Y Surgical sperm retrieval techniques	What is the clinical and cost effectiveness of surgical sperm retrieval (SSR) techniques for fertility problems associated with non-obstructive azoospermia, or obstructive azoospermia?	Intervention	2

[1]*Original health economic analysis conducted*

The COMET database was searched for core outcome sets relevant to this guideline. A core outcome set was identified, including live birth, clinical pregnancy (viable intrauterine pregnancy confirmed by ultrasound), pregnancy loss, gestational age at delivery, birth weight, neonatal mortality, and major congenital anomaly, as core clinical outcomes (Duffy 2020). Additional outcomes were chosen based on committee discussions.

Additional information related to development of the guideline is contained in: Supplement 2 (Acknowledgements).

Searching for evidence

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 articles and conference reports were applied where possible. Searches were conducted in the following databases: MEDLINE ALL (Ovid); Embase (Ovid); Cochrane Central Register of Controlled Trials (CENTRAL) (Wiley); Cochrane Database of Systematic Reviews (CDSR) (Wiley) and Epistemonikos.

In cases where a previous systematic review was available from the Cochrane Collaboration, we relied upon the existing reviews as an evidence source. The committee were asked about their knowledge of any new evidence published subsequent to the Cochrane searches and where appropriate the searches were updated to include any more recent evidence for both effectiveness and cost-effectiveness review questions.

During the development of the guideline, the fertility treatment add-ons rating system developed by the Human Fertilisation and Embryology Authority (HFEA) was identified as relevant to a number of review questions (evidence reviews M to R). Given the potential for efficiencies to the guideline development process and the applicability of the HFEA's work to the UK setting, the committee took the pragmatic decision to draft recommendations relevant to these review questions based on the evidence identified by the HFEA, and the HFEA ratings and as such no new systematic review of evidence were conducted for these review question.

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, economic evaluations and quality of life filters. Limits to exclude animal studies, letters, editorials, news articles and conference reports were applied where possible. Searches were conducted in the following databases: MEDLINE ALL (Ovid); Embase (Ovid); the International Network of Agencies for Health Technology Assessments (INAHTA) database and Health Technology Assessments (HTA) database (CRD).

In addition, searches were undertaken for evidence on quality-of-life. Databases were searched using population subject headings and free-text terms combined with a quality-of-life search filter. Limits to exclude animal studies, letters, editorials, news articles and conference reports were applied where possible. Searches were conducted in the following databases: MEDLINE ALL (Ovid); Embase (Ovid); the International Network of Agencies for Health Technology Assessments (INAHTA) database and Health Technology Assessments (HTA) database (CRD).

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

Quality assurance

A NICE senior information specialist (SIS) conducted the searches. The MEDLINE strategy was quality assured by another NICE SIS. All translated search strategies were peer reviewed to ensure their accuracy. Both procedures were adapted from the 2015 PRESS Guideline Statement. Further details and full search strategies for each database are provided in Appendix B.

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 quantitative evidence by outcome were presented in the corresponding evidence review and discussed by the committee.

For all review questions, titles and abstracts of identified studies were dual screened until a good inter-rater reliability had been observed (at least 90% agreement). Initially 10% of references were double-screened, and if inter-rater agreement was satisfactory then the remaining references were screened by one reviewer. Any discrepancies were resolved by discussion between the first and second reviewers or by reference to a third (senior) reviewer. At least 10% of the data extraction was double-coded.

Drafts of all evidence reviews were checked by a senior reviewer.

Type of studies and inclusion/exclusion criteria

Inclusion and exclusion of studies was based on criteria specified in the corresponding review protocol. A general rule across reviews was that if some, but not all, of a study's participants were eligible for the review, then the study would be included if at least 80% of its participants were eligible for the review.

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 no

RCT evidence to inform guideline decision making, quasi-randomised controlled trials (experimental studies using a non-randomly assigned control group design with match comparison or another method of controlling for confounding variables) were considered. For reviews where the guideline committee had identified that there would be no comparative studies due to ethical issues, case series studies were considered for inclusion, and prospective case series were prioritised.

For prognostic reviews, prospective and retrospective cohort studies were considered for inclusion. Prospective studies, and studies that included multivariable analysis, were prioritised.

For reviews of clinical prediction models, model development and validation studies were considered for inclusion. External validation studies were prioritised.

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.

Conference abstracts, dissertations, unpublished data and studies published in languages other than English were generally not considered for inclusion, unless the data could be extracted (and risk of bias assessed) from elsewhere (for instance, from an existing systematic review).

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 live birth, the Mantel–Haenszel method was used to calculate risk ratios (RRs) or odds ratios (ORs). 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 gestational age at delivery, were meta-analysed using an inverse-variance method for pooling weighted mean differences (WMDs).

For the pairwise meta-analysis reviews, it was generally considered likely that a random-effects model would be used (based on assumptions about methodological diversity of the studies). The initial choice of a fixed-choice or random-effects model was pre-defined at the protocol stage (see the protocols for each review for further detail). Funnel plot asymmetry (relationship between the magnitude of the effect estimate and study size) was considered for meta-analyses that included at least 10 studies, and where asymmetry was indicated a fixed-effects model was conducted (and both random-effects and fixed-effects analyses were presented) or sensitivity analyses excluding small studies was considered.

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

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

Network meta-analysis

Network meta-analysis (NMA) is a generalization of standard pairwise meta-analysis for A versus B trials, to data structures that include, for example, A versus B, B versus C, and A versus C trials (Dias 2011, Lu 2004). A basic assumption of NMA methods is that direct and indirect evidence estimate the same parameter, that is, the relative effect between A and B measured directly from an A versus B trial, is the same with the relative effect between A and B estimated indirectly from A versus C and B versus C trials. NMA techniques strengthen inference concerning the relative effect of two treatments by including both direct and indirect comparisons between treatments, and, at the same time, allow simultaneous inference on all treatments examined in the pairwise trial comparisons, which is essential for consideration of treatment in economic analysis (Caldwell 2005, Lu 2004). Simultaneous inference on the relative effect of a number of treatments is possible provided that treatments participate in a single “network of evidence”, that is, every treatment is linked to at least one of the other treatments under assessment through direct or indirect comparisons. NMA takes all trial information into consideration, without ignoring part of the evidence and without introducing bias by breaking the rules of randomisation.

A key assumption when conducting an NMA is that the populations included in all randomised controlled trials (RCTs) considered in the NMA are similar so that the treatment effects are exchangeable across all populations (Mavridis 2015). This assumption of ‘transitivity’ of the effect may not hold if there are different potential effect modifiers that are not equally distributed across the different comparisons (Jansen 2014).

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.

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.

NMAs were conducted for 1 topic area:

- Assisted reproduction techniques for people with unexplained fertility problems, mild endometriosis, mild male factor fertility problems (evidence report K)

We 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.

The Guidelines Technical Support Unit (TSU), at University of Bristol, provided advice, models, inconsistency checking and quality assurance for the network meta-analyses included in this review.

Single-arm proportional meta-analysis

For some review questions, where there would not be relevant comparative studies, for example, due to ethical issues around not treating an identified abnormality where screening and treatment are part of the same procedure, proportional (single-arm) meta-analyses were conducted to address the key clinical question.

Meta-analyses of proportions were conducted using the metaprop function in the R software package. Data analysis was conducted using the generalized linear mixed model (GLMM) (Lin 2020). The outcomes were reported as proportions with corresponding 95% confidence intervals, as well as statistical heterogeneity data (I^2, τ^2). Heterogeneity was explored using planned subgroup analyses (outlined in the corresponding evidence review protocol).

Data synthesis for prognostic reviews

Where multiple studies reported on the same prognostic factor(s) and the definitions used and approach to analysis in the primary papers was sufficiently consistent, the evidence was meta-analysed using Cochrane Review Manager software. Adjusted and unadjusted estimates were considered in separate analyses, and adjusted estimates were prioritised.

For meta-analyses of adjusted estimates, only estimates that adjusted for a core/minimal set of prognostic factors (female age and duration of infertility) were included (although studies may also adjust for other prognostic factors in addition to this core set).

Random effects meta-analyses were conducted (to allow for unexplained heterogeneity across prognosis studies) and data were presented as risk ratios if possible or odds ratios when required (for example, if only an adjusted odds ratio was reported). In addition to separate analyses for adjusted and unadjusted estimates, separate meta-analyses were conducted for risk ratios and odds ratios, and for different thresholds or cut-offs (where relevant). Heterogeneity in the effect estimates of the individual studies was assessed by visual inspection of the forest plots and consideration of the I^2 statistic. Heterogeneity was explored as appropriate using sensitivity analyses and pre-specified subgroup analyses (outlined in the corresponding evidence review protocol). If heterogeneity could not be explained through subgroup analysis then the data were not pooled.

Data synthesis for prediction model reviews

Predictive performance data from clinical prediction model studies was quantitatively summarised, including summaries of discrimination (ability of the model to distinguish between people who had a live birth and those who did not) and calibration (agreement between predicted outcomes and observed outcomes).

Where there was discrimination and calibration data from multiple external validation studies for the same clinical prediction model, estimates were meta-analysed with a random-effects model (to allow for the presence of heterogeneity due to variability in design and population [case mix] of validation studies) using the metamisc package in R (see Debray 2019).

Thresholds for interpreting discrimination and calibration were pre-specified in the corresponding evidence review protocol. ‘Good’ discrimination was defined as a C statistic >0.75 and ‘good’ calibration was defined as an observed:expected (O:E) ratio between 0.8 and 1.2 (based on Debray 2017).

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

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

Quality element	Description
Risk of bias ('Study limitations')	This refers to limitations in study design or implementation that reduce the internal validity of the evidence
Inconsistency	This refers to unexplained heterogeneity in the results
Indirectness	This refers to differences in study populations, interventions, comparators or outcomes between the available evidence and inclusion criteria specified in the review protocol

Quality element	Description
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

Risk of bias in RCTs (and quasi-RCTs if RCTs could not be identified) was assessed using the Cochrane risk of bias 2.0 tool (see Appendix H in Developing NICE guidelines: the manual).

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

- randomisation process
- deviations from the intended interventions
- missing outcome data
- measurement of the outcome
- 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 2.0 tool can be found in Section 8 of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins 2020).

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

For case series studies, the JBI checklist for case series was used (see Appendix H in Developing NICE guidelines: the manual).

For prognostic studies, the Quality in Prognosis Studies (QUIPS) tool was used (see Appendix H in Developing NICE guidelines: the manual).

For studies that included the development and/or validation (internal or external) of clinical prediction models, the Prediction model Risk Of Bias Assessment Tool (PROBAST) was used (see Appendix H in Developing NICE guidelines: the manual).

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. If heterogeneity was very serious and could not be accounted for by sub-group analyses the data was not pooled.

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.

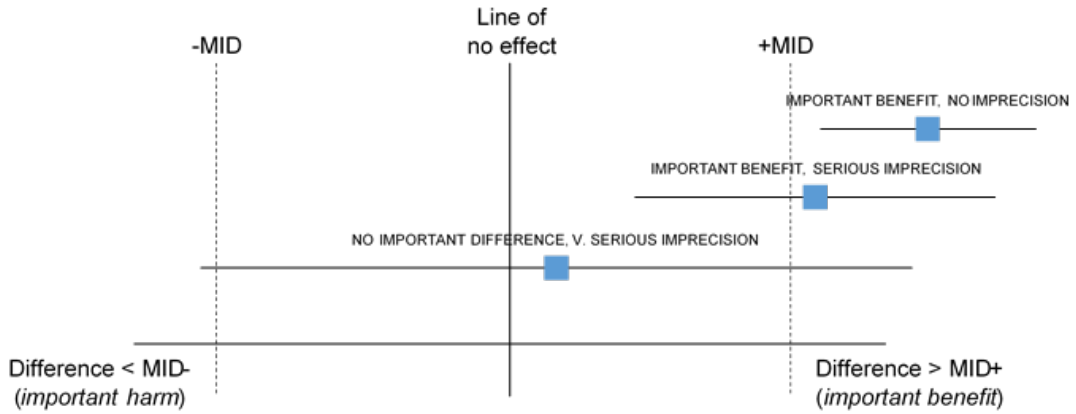
When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

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 agreed that live birth was a sufficiently critical outcome that any statistically significant difference would be considered important.

For the remaining outcomes, 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 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.

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

For live birth, where the MID was statistical significance, imprecision was judged based on optimal information size (OIS) criteria. Evidence was considered seriously imprecise if there were less than 300 events, based on the rule-of-thumb specified in version 3.2 of the GRADE handbook (Schünemann 2009), and very seriously imprecise if there were less than 150 events. The threshold for very serious imprecision was a pragmatic decision, in the absence of a rule-of-thumb being available, based on the fact that this is half the number required for serious imprecision, which would be consistent with the approach suggested for continuous outcomes. Imprecision was also judged based on OIS criteria for proportional single-arm meta-analyses due to the absence of minimally important differences for these reviews.

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 2.0 tool, as well as heterogeneity and consistency (also called coherence). Heterogeneity concerns the differences in treatment effects between trials within each treatment contrast (measured by the posterior median between-study standard deviation and compared with treatment posterior mean effects), while consistency concerns the differences between the direct and indirect evidence informing the treatment contrasts. Direct and indirect comparisons measure the same underlying true effect, and therefore, in principle they should be consistent. However, this is not the case if effect modifiers and heterogeneity across studies, populations and comparisons are present. Inconsistency arises when there is a conflict between direct evidence (from an A vs. B trial) and indirect evidence (gained from A vs. C and B vs. C trials) and can only be assessed when there are closed loops of evidence on three treatments that are informed by at least three distinct trials (Caldwell 2014).

Checking for inconsistency between direct and indirect evidence can reveal whether the transitivity assumption holds. To determine if there was evidence of

inconsistency, in each analysis, the selected consistency model (fixed or random effects) was compared to an “inconsistency”, or unrelated mean effects, model (Dias 2013). When evidence of inconsistency was found, studies contributing to between-trial heterogeneity were checked for data accuracy and analyses were repeated if corrections in the data extraction were made. However, following any data corrections and if inconsistency persisted, no studies were excluded from the analysis, as their results could not be considered as less valid than those of other studies solely because of the inconsistency findings. Nevertheless, the presence of inconsistency in the network was highlighted and results were interpreted accordingly by the committee.

Threshold analysis was conducted to test the robustness of treatment recommendations based on the NMA, to potential biases or sampling variation in the included evidence. Threshold analysis has been developed as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analysis (Phillippo 2019).

Prognostic studies

Adapted GRADE methodology for prognostic reviews

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

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

Table 5: Adaptation of GRADE quality elements for prognostic reviews

Quality element	Description
Risk of bias (‘Study limitations’)	Limitations in study design and implementation may bias estimates and interpretation of the effect of the prognostic/risk factor. High risk of bias for the majority of the evidence reduces confidence in the estimated effect. Prognostic studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Inconsistency	This refers to unexplained heterogeneity between studies looking at the same prognostic/risk factor, resulting in wide variability in estimates of association (such as RRs or ORs), with little or no overlap in confidence intervals
Indirectness	This refers to any departure from inclusion criteria listed in the review protocol (such as differences in study populations or prognostic/risk factors), that may affect the generalisability of results
Imprecision	This occurs when a study has relatively few participants and also when the number of participants is too small for a multivariable

Quality element	Description
	analysis (as a rule of thumb, 10 participants are needed per variable). This was assessed by considering the confidence interval in relation to the point estimate for each outcome reported in the included studies

RR, relative risk; OR, odds ratio

Assessing risk of bias in prognostic reviews

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

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

Assessing inconsistency in prognostic reviews

Where multiple results were deemed appropriate to meta-analyse (that is, there was sufficient similarity between risk factor and outcome under investigation) inconsistency was assessed by visually inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis. This was assessed by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating 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.

Assessing indirectness in prognostic reviews

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

Assessing imprecision and importance in prognostic reviews

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

Clinical prediction model studies

Adapted GRADE methodology for clinical prediction models

For clinical prediction models, an adapted GRADE approach was used. GRADE methodology is designed for intervention reviews but the quality assessment elements and outcome presentation were adapted by the guideline developers for prediction models. For example, GRADE tables were modified to include model performance estimates (C-statistic and observed:expected event ratios).

The evidence for each outcome in the prediction models was examined separately for the quality elements listed and defined in Table 6. The criteria considered in the rating of these elements are discussed below. Each element was graded using the quality levels summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having 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.

Table 6: Adaptation of GRADE quality elements for prediction model reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in study design and implementation may bias estimates of model performance. High risk of bias for the majority of the evidence reduces confidence in the estimated effect. Clinical prediction model studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Inconsistency	This refers to unexplained heterogeneity in model performance estimates (such as C-statistic and O:E ratio) between studies
Indirectness	This refers to differences in study populations, models, or outcomes between the available evidence and inclusion criteria specified in the review protocol
Imprecision	Imprecision was considered in line with the pre-specified thresholds for good discrimination and calibration

Assessing risk of bias in clinical prediction model studies

Risk of bias in clinical prediction model studies was assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST) checklist (see Appendix H in Developing NICE guidelines: the manual).

Risk of bias in prediction models in PROBAST consists of 4 domains:

- participants
- predictors
- outcome
- analysis.

More details about the PROBAST tool can be found on the developer's website.

Assessing inconsistency in clinical prediction model studies

Inconsistency refers to the unexplained heterogeneity of the results in meta-analysis. When estimates of model performance parameters vary widely across studies (that

is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only truly applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled).

Inconsistency for meta-analyses of external validation studies of prediction models was assessed based on the prediction interval, which gives an estimate of an interval in which a model performance estimate from a future study would be expected to fall.

Assessing indirectness in prediction model studies

Indirectness in clinical prediction model studies was assessed using the PROBAST checklist by assessing the applicability of the studies in relation to the review question in the following domains:

- participants
- predictors
- outcome.

More details about the PROBAST tool can be found on the developer's website.

Assessing imprecision in clinical prediction model studies

The judgement of precision for clinical prediction model evidence was based on the CIs of the C-statistic and O:E ratios. The minimally important differences of 0.75 for the C-statistic and a range of 0.8-1.2 for the observed:expected (O:E) ratio were used (based on definitions of 'good' discrimination and calibration in Debray et al. 2017).

The following cut-offs were used when judging the imprecision of the model performance data in terms of the C-statistic, and outcomes were downgraded once or twice based on the number of these thresholds that were crossed by the CI:

- C-statistic <0.6 poor discrimination
- C-statistic 0.6-0.75 possibly helpful discrimination
- C-statistic >0.75 clearly useful discrimination.

Reviewing economic evidence

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

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

Inclusion criteria
Intervention or comparators in accordance with the guideline scope
Study population in accordance with the guideline scope
Full economic evaluations (cost-utility, cost effectiveness, cost-benefit or cost-consequence analyses) assessing both costs and outcomes associated with interventions of interest
Exclusion criteria
Abstracts containing insufficient methodological details
Cost-of-illness type studies

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

Details of economic evidence study selection, lists of excluded studies, economic evidence tables, the results of quality assessment of economic evidence (see below) and health economic evidence profiles are presented in the evidence reviews.

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 Integrated Care Boards (ICB) 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.

- What is the clinical and cost effectiveness of ovarian stimulation, intrauterine insemination (IUI) with or without ovarian stimulation, in vitro fertilisation (IVF) and expectant management for people with unexplained health-related fertility problems, mild endometriosis, and people with a single abnormal semen parameter?
- What is the predictive performance of clinical prediction models for assessing the chances of live birth for people with health-related fertility problems using:
  - expectant management,
  - intrauterine insemination (IUI),
  - IVF with or without intracytoplasmic sperm injection (ICSI)?

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

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

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

Other sources of evidence

Human Fertilisation and Embryology Authority (HFEA) fertility treatment add-on ratings

During the development of the guideline, the fertility treatment add-ons rating system developed by the Human Fertilisation and Embryology Authority (HFEA) was identified as relevant to the effectiveness of treatment add-ons (Evidence reviews M-Q). Given the potential for efficiencies to the guideline development process and the applicability of the HFEA's work to the UK setting, it was agreed that the committee would draft recommendations relevant to these review questions based on the evidence identified by the HFEA, and the HFEA ratings. The HFEA ratings are available here: https://www.hfea.gov.uk/treatments/treatment-add-ons/

For the reviews on fertility treatment add-ons, the committee also considered the European Society of Human Reproduction and Embryology (ESHRE) Good practice recommendations on add-ons in reproductive medicine, and any relevant Cochrane reviews on the treatment add-on.

Developing recommendations

Guideline recommendations

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

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

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

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.

References

Bradburn 2007

Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Statistics in Medicine. 2007 26(1): 53-77.

Caldwell 2005

Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005 331(7521): 897-900.

Caldwell 2014

Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. Systematic Reviews. 2014 3(1): 109.

Debray 2017

Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KG. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017: 356.

Debray 2019

Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, Collins GS, Moons KG. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Statistical Methods in Medical Research. 2019 28(9): 2768-2786.

Dias 2011

Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials, 2011 (last updated 2016). http://www.nicedsu.org.uk

Dias 2013

Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades A. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Medical Decision Making. 2013 33(5): 641-656.

Duffy 2020

Duffy JM, AlAhwany H, Bhattacharya S, Collura B, Curtis C, Evers JL, Farquharson RG, Franik S, Giudice LC, Khalaf Y, Knijnenburg JM. Developing a core outcome set for future infertility research: an international consensus development study. Human Reproduction. 2020 35(12): 2725-2734.

Hayden 2013

Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013 158(4): 280-286.

Higgins 2020

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions Version 6.1 [updated September 2020] The Cochrane Collaboration. Available from www.training.cochrane.org/handbook (accessed November 2024)

Jansen 2014

Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, Salanti G. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value in Health. 2014 17(2): 157-173.

Lin 2020

Lin L, Chu H. Meta-analysis of proportions using generalized linear mixed models. Epidemiology. 2020 31(5): 713-717.

Lu 2004

Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine. 2004 23(20): 3105-3124.

Mavridis 2015

Mavridis D, Giannatsi M, Cipriani A, Salanti G. A primer on network meta-analysis with emphasis on mental health. BMJ Mental Health. 2015 18(2): 40-46.

McGowan 2016

McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. Journal of Clinical Epidemiology. 2016 75: 40-46.

Phillippo 2019

Phillippo DM, Dias S, Welton NJ, Caldwell DM, Taske N, Ades AE. Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses. Annals of Internal Medicine. 2019 170(8): 538-546.

Santesso 2016

Santesso N, Carrasco-Labra A, Langendam M, Brignardello-Petersen R, Mustafa RA, Heus P, Lasserson T, Opiyo N, Kunnamo I, Sinclair D, 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. 2016 74: 28-39.

Schünemann 2009

Schünemann H., Brożek J., Oxman A., editors. (2009). GRADE handbook Handbook for grading Grading quality Quality of evidence Evidence and strength Strength of recommendationRecommendation. Version 3.2. 2009 [updated March 2009]