National Institute for Health and Care Excellence

Draft for consultation

Depression in adults

Supplement 1: Methods

NICE guideline CG90 (update)

Development of the guideline and methods

November 2021

Draft for consultation

This supplement was developed by the National Guideline Alliance which is a part of the Royal College of Obstetricians and Gynaecologists

Disclaimer

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

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

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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN:

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to update the existing NICE guideline on Depression in adults: recognition and management (CG90) (NICE, 2009). As part of this update, this guideline has been renamed Depression in adults.

The following sections of the guideline were updated using the methods in this chapter:

1.3 Choice of treatments
1.4 General principles of care (partial – starting and stopping antidepressants)
1.5 Treatment for a new episode of less severe depression
1.6 Treatment for a new episode of more severe depression
1.7 Behavioural couples therapy for depression
1.8 Continuation of treatment for relapse prevention
1.9 Further-line treatment
1.10 Chronic depressive symptoms
1.11 Depression with a diagnosis of personality disorder
1.12 Psychotic depression
1.13 Electroconvulsive therapy for depression
1.15 Access, coordination and delivery of care

The following sections of the guideline were not included in the scope of this update:

1.1 Experience of care
1.2 Recognition, assessment and initial management
1.4 General principles of care (except sections highlighted above)
1.14 Transcranial magnetic stimulation for depression

Methods

Introduction

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

Declarations of interest were recorded and managed in accordance with NICE's 2018 Policy on declaring and managing interests for NICE advisory committees.

Developing the review questions and outcomes

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

The review questions were based on the following frameworks:

- intervention reviews – using population, intervention, comparison and outcome (PICO)
- qualitative reviews – using population, phenomenon of interest and context

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

Literature searches, critical appraisal and evidence reviews were completed for all review questions.

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

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

Evidence review	Review question	Type of review
[A] Service delivery	RQ1.1 For adults with depression, what are the relative benefits and harms associated with different models for the coordination and delivery of services? RQ1.2 For adults with depression, what are the relative benefits and harms	Intervention

Evidence review	Review question	Type of review
	associated with different settings for the delivery of care?	
[B] Treatment of a new episode of depression	RQ2.1 For adults with a new episode of less severe depression, what are the relative benefits and harms of psychological, psychosocial, pharmacological and physical interventions alone or in combination? RQ2.2 For adults with a new episode of more severe depression, what are the relative benefits and harms of psychological, psychosocial, pharmacological and physical interventions alone or in combination?	Intervention[1]
[C] Prevention of relapse	RQ2.3 For adults whose depression has responded to treatment, what are the relative benefits and harms of psychological, psychosocial, pharmacological and physical interventions for preventing relapse (including maintenance treatment)?	Intervention[1]
[D] Further-line treatment	RQ2.4/2.5 What are the relative benefits and harms of further-line psychological, psychosocial, pharmacological and physical interventions (alone or in combination), for adults with depression showing an inadequate response to at least one previous intervention for the current episode?	Intervention
[E] Chronic depression	RQ2.6 For adults with chronic depression or persistent subthreshold depression symptoms what are the relative benefits and harms of first-line treatment or relapse prevention with psychological, psychosocial, pharmacological and physical interventions (alone or in combination)?	Intervention

Evidence review	Review question	Type of review
[F] Depression with personality disorder	RQ2.7 For adults with depression and a coexisting personality disorder what are the relative benefits and harms of first-line treatment or relapse prevention with psychological, psychosocial, pharmacological and physical interventions alone or in combination?	Intervention
[G] Psychotic depression	RQ2.8 For adults with psychotic depression what are the relative benefits and harms of psychological, psychosocial, pharmacological and physical interventions alone or in combination (as first-line treatment or relapse prevention)?	Intervention
[H] Access	RQ3.0 For adults at risk of depression (or anxiety disorders) from particular vulnerable groups (older people, black minority ethnic groups, lesbian, gay bisexual, transgender groups and men) do service developments and interventions which are specifically designed to promote access, increase the proportion of people from the target group who access treatment, when compared with standard care?	Intervention
[I] Patient choice	RQ4.0 What are the facilitators and barriers that can enhance or inhibit choice of treatment for adults with depression?	Qualitative

[1]*Original health economic analysis conducted*

The outcomes were chosen based on committee discussions.

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

- Supplement 2 (Glossary and abbreviations)
- Supplement 3 (Economic evidence)
- Supplement 4 (NGA staff and contributors)

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. All the searches were conducted in the following databases: Medline, Medline-in-Process, Cochrane Central Register of Controlled Trials (CCTR), Cochrane Database of Systematic Reviews (CDSR), Embase and PsycINFO.

Searches were run once for all reviews during development. Searches for the questions 2.1 to 2.7 were updated in June 2020 and for the remaining questions in March 2021.

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

Citation search

In order to identify follow-up studies, searches were undertaken for review questions 2.1 to 2.8 to identify published evidence that cited the original included references. The Science and Social Science Citation Indexes (Web of Science) were searched.

Economic systematic literature search

Systematic literature searches were also undertaken to identify published economic evidence and studies reporting utility data that could inform economic modelling. Databases were searched using subject headings, free-text terms and, economic evaluations and health utility search filters.

A single search, using the population search terms used in the evidence reviews, was conducted to identify economic evidence in the HTA database. Another single search, using the population search terms used in the evidence reviews combined with an economic evaluations and a health utility search filter, was conducted in Medline, Embase, PsycINFO and CINAHL. Where possible, searches were limited to studies published in English.

As with the general literature searches, the economic literature searches were updated in June 2020.

Details of the search strategies, including the study-design filter used and databases searched, are provided in Supplement 3 (Economic evidence).

Quality assurance

Search strategies were quality assured by cross-checking reference lists of relevant studies, analysing search strategies from published systematic reviews and asking members of the committee to highlight key studies. The principal search strategies for each search were also quality assured by a second information scientist using an adaptation of the PRESS 2015 Guideline Evidence-Based Checklist (McGowan 2016). In addition, all publications highlighted by stakeholders at the time of the consultation on the draft scope were considered for inclusion.

Reviewing evidence

Systematic review process

The evidence was reviewed in accordance with the following approach.

- Potentially relevant articles were identified from the search results for each review question by screening titles and abstracts. Full-text copies of the articles were then obtained.
- Full-text articles were reviewed against pre-specified inclusion and exclusion criteria in the review protocol (see appendix A of each evidence review).
- Key information was extracted from each article on study methods and results, in accordance with factors specified in the review protocol. The information was presented in a summary table in the corresponding evidence review and in a more detailed evidence table (see appendix D of each evidence review).
- Included studies were critically appraised using an appropriate checklist as specified in Developing NICE guidelines: the manual. Further detail on appraisal of the evidence is provided below.
- Summaries of evidence by outcome were presented in the corresponding evidence review and discussed by the committee.

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 or meta-syntheses were considered to be the highest quality evidence that could be selected for inclusion.

For intervention reviews, only randomised controlled trials (RCTs) were eligible for inclusion because they are considered to be the most robust type of study design that could produce an unbiased estimate of intervention effects.

For qualitative reviews, studies using focus groups, structured interviews or semi-structured interviews were considered for inclusion. Where qualitative evidence was sought, data from surveys or other types of questionnaire were considered for inclusion only if they provided data from open-ended questions, but not if they reported only quantitative data.

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 K 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 reviews

Pairwise meta-analysis

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

For dichotomous outcomes, such as remission, the Mantel–Haenszel method with a random effect model was used to calculate risk ratios (RRs). A random effect model was used due to assumed heterogeneity based on the clinical diversity of depression, differences between interventions that formed a class, and methodological variation between studies.

For continuous outcomes, measures of central tendency (mean) and variation (standard deviation; SD) are required for meta-analysis. Data for continuous outcomes, such as depression symptoms, were meta-analysed using random effects models of standardised mean differences (SMDs). A random effect model was used due to assumed heterogeneity based on the clinical diversity of depression, differences between interventions that formed a class, and methodological variation

between studies. SMD was used for all continuous outcome measures, even for comparisons that included only a single study, in order to ensure comparability between comparisons and timepoints.

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 2011a; 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 pair-wise 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, 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.

Class models were used so that strength could be borrowed across treatments in the same class and to reconnect disconnected networks. Classes of treatments are groups of interventions which are thought to have similar modes of action and, consequently, similar effects. For all outcomes, both fixed and random class effects models were fitted. The random class effects model assumes the relative effects of treatments within a class are exchangeable. Treatment effects are shrunk towards a class mean and can borrow strength from other elements of the class. The fixed class effects model assumes treatments within a class have identical relative effects.

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. Non-informative prior distributions were used to maximise the weighting given to the data, in order 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 models were also estimated by using the posterior mean of the sum of the deviance contributions for each item by calculating the residual deviance and the deviance information criterion (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.

NMA was conducted for 2 topic areas:

- Treatment of a new episode of depression (evidence report B). NMA was conducted to inform the clinical analysis.
- Prevention of relapse (evidence report C). NMA was conducted to inform the economic analysis.

The NMA work around treatment of new episodes of depression was undertaken by the NICE Guidelines Technical Support Unit, University of Bristol (TSU). The NMA work around relapse prevention was undertaken by the NGA, and was subsequently quality assured by the NICE Guidelines TSU.

Overall methods and approaches adopted for the guideline NMA work were based on methodology described in the NICE Decision Support Unit (DSU) technical support document number 2 (Dias 2011a).

Details of the NMA methods employed in this guideline are provided in evidence reports B and C.

Data synthesis for qualitative reviews

Qualitative data extraction and synthesis was guided by a thematic analysis approach. This approach was selected as the relevant review question was explorative in nature. This was guided by the 6 phases outlined by Braun and Clarke (2006): familiarizing yourself with the data; generating initial codes; searching for themes; reviewing themes; defining and naming themes; producing the report. Thematic maps were used as an aid to think about the relationship between codes, between themes, and between different levels of themes (e.g. main overarching themes and subthemes within them), and to inductively identify, review and refine the themes and subthemes that describe the qualitative data. All data was double-coded.

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 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).

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

Quality element	Description
Risk of bias ('Study limitations')	This refers to limitations in study design or implementation that reduce the internal validity of the evidence
Inconsistency	This refers to unexplained heterogeneity in the results
Indirectness	This refers to differences in study populations, interventions, comparators or outcomes between the available evidence and inclusion criteria specified in the review protocol
Imprecision	This occurs when a study has few participants or few events of interest, resulting in wide confidence intervals that cross minimally important thresholds
Publication bias	This refers to systematic under- or over-estimation of the underlying benefit or harm resulting from selective publication of study results

Table 3: GRADE quality ratings (by quality element)

Quality issues	Description
None or not serious	No serious issues with the evidence for the quality element under consideration
Serious	Issues with the evidence sufficient to downgrade by 1 level for the quality element under consideration
Very serious	Issues with the evidence sufficient to downgrade by 2 levels for the quality element under consideration

Table 4: Overall quality of the evidence in GRADE (by outcome)

Overall quality grading	Description
High	Further research is very unlikely to change the level of confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on the level of confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on the level of confidence in the estimate of effect and is likely to change the estimate
Very low	The estimate of effect is very uncertain

Assessing risk of bias in intervention reviews

Bias is a systematic error, or consistent deviation from the truth in results obtained. When a risk of bias is present the true effect can be either under- or over-estimated.

Risk of bias in RCTs was assessed using the Cochrane risk of bias tool (see Appendix H in Developing NICE guidelines: the manual.

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

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

A study with a poor methodological design does not automatically imply high risk of bias; the bias is considered individually for each outcome and it is assessed whether the chosen design and methodology will impact on the estimation of the intervention effect.

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 2011).

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. In the case of unexplained heterogeneity, sensitivity analyses were planned based on the quality of studies, eliminating studies at high risk of bias (in relation to randomisation, allocation concealment and blinding, and/or missing outcome data).

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency.

Assessing indirectness in intervention reviews

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

Assessing imprecision and importance in intervention reviews

Imprecision in GRADE methodology refers to uncertainty around the effect estimate and whether or not there is an important difference between interventions (that is, whether the evidence clearly supports a particular recommendation or appears to be consistent with several candidate recommendations). Therefore, imprecision differs from other aspects of evidence quality because it is not concerned with whether the point estimate is accurate or correct (has internal or external validity). Instead, it is concerned with uncertainty about what the point estimate actually represents. This uncertainty is reflected in the width of the CI.

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

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

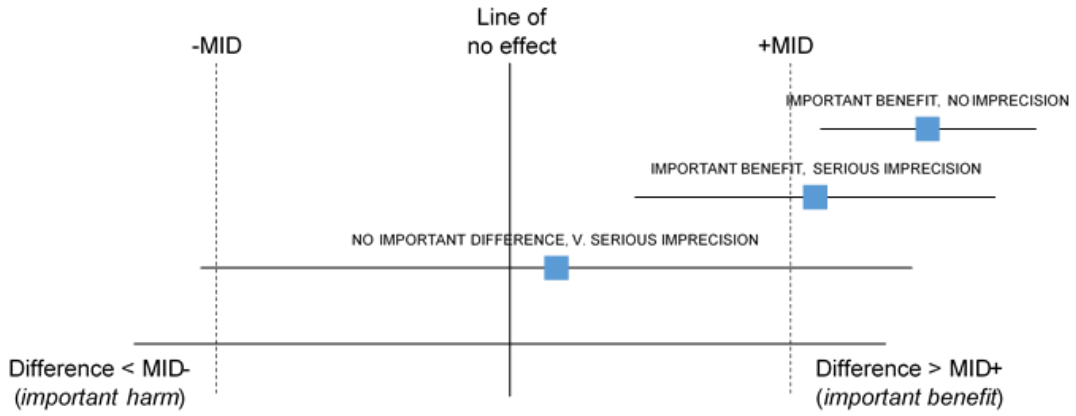
When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

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

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

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

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 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. For continuous outcomes minimally important thresholds for a SMD of -0.5 and 0.5 respectively were used as default MIDs in the guideline.

If risk difference was used for meta-analysis, for example if the majority of studies had zero events in either arm, imprecision was assessed based on sample size using 300 and 500 as cut-offs for very serious and serious imprecision respectively. The committee used these numbers based on commonly used optimal information size thresholds.

The same thresholds were used as default MIDs in the guideline for all outcomes considered in intervention evidence reviews.

Assessing publication bias in intervention reviews

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 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; van Valkenhoef 2016).

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.

However, tests of inconsistency are inherently underpowered, so they may fail to detect inconsistency even though this may be present in the network (Dias 2011b). Therefore, even if inconsistency is not detected, results of NMA should be interpreted following qualitative evaluation of the anticipated transitivity within the network and judgement of reasons for potential inconsistency (Linde 2016).

Bias adjustment models were fitted to down-weight trials at high or unclear risk of bias for domains of the Cohrane Risk of Bias tool that had sufficient variability in the ratings. Models that adjusted for small study bias were also fitted (Dias 2010, Welton 2009).

Threshold analysis was planned 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). After discussion with the committee, threshold

analysis was not undertaken as planned. Full details of the reasons for this decision are explained in evidence review B.

Qualitative reviews

GRADE-CERQual methodology for qualitative reviews

For qualitative reviews an adapted GRADE Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) approach (Lewin 2015) was used. In this approach the quality of evidence is considered according to themes in the evidence. The themes may have been identified in the primary studies or they may have been identified by considering the reports of a number of studies. Quality elements assessed using GRADE-CERQual are listed and defined in Table 5. Each element was graded using the levels of concern summarised in Table 6. The ratings for each component were combined (as with other types of evidence) to obtain an overall confidence in the evidence for each theme as described in Table 7. 'Confidence' in this context refers to the extent to which the review finding is a reasonable representation of the phenomenon of interest set out in the protocol. Similar to other types of evidence all review findings start off with 'high confidence' and are rated down by one or more levels if there are concerns about any of the individual CERQual components. In line with advice from the CERQual developers, the overall assessment does not involve numerical scoring for each component but in order to ensure consistency across and between guidelines, the NGA established some guiding principles for overall ratings. For example, a review finding would not be downgraded (and therefore would be assessed with 'high' confidence) if all 4 components had 'no or very minor' concerns or 3 'no or very minor' and 1 'minor'. At the other extreme, a review finding would be downgraded 3 times (to 'very low') if at least 2 components had serious concerns or at least 3 had moderate concerns. A basic principle was that if any components had serious concerns then overall confidence in the review finding would be downgraded at least once (potentially more depending on the other ratings). Transparency about overall judgements is provided in the CERQual tables, including a brief reference to components for which there were concerns in the 'overall confidence' cell.

Table 5: Adaptation of GRADE quality elements for qualitative reviews

Quality element	Description
Risk of bias ('Methodological limitations')	Limitations in study design and implementation may bias interpretation of qualitative themes identified. High risk of bias for the majority of the evidence reduces confidence in review findings. Qualitative studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Relevance (or applicability) of evidence	This refers to the extent to which the evidence supporting the review findings is applicable to the context specified in the review question
Coherence of findings	This refers to the extent to which review findings are well grounded in data from the contributing primary studies and provide a credible explanation for patterns identified in the evidence

Quality element	Description
Adequacy of data (theme saturation or sufficiency)	This corresponds to a similar concept in primary qualitative research, that is, whether a theoretical point of theme saturation was achieved, at which point no further citations or observations would provide more insight or suggest a different interpretation of the particular theme. Individual studies that may have contributed to a theme or sub-theme may have been conducted in a manner that by design would have not reached theoretical saturation at an individual study level

Table 6: CERQual levels of concern (by quality element)

Level of concern	Definition
None or very minor concerns	Unlikely to reduce confidence in the review finding
Minor concerns	May reduce confidence in the review finding
Moderate concerns	Will probably reduce confidence in the review finding
Serious concerns	Very likely to reduce confidence in the review finding

Table 7: Overall confidence in the evidence in CERQual (by review finding)

Overall confidence level	Definition
High	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest
Moderate	It is likely that the review finding is a reasonable representation of the phenomenon of interest
Low	It is possible that the review finding is a reasonable representation of the phenomenon of interest
Very low	It is unclear whether the review finding is a reasonable representation of the phenomenon of interest

Assessing methodological limitations in qualitative reviews

Methodological limitations in qualitative studies were assessed using the Critical Appraisal Skills Programme (CASP) checklist for qualitative studies (see appendix H in Developing NICE guidelines: the manual). Overall methodological limitations were derived by assessing the methodological limitations across the 6 domains summarised in Table 8.

Table 8: Methodological limitations in qualitative studies

Aim and appropriateness of qualitative evidence	This domain assesses whether the aims and relevance of the study were described clearly and whether qualitative research methods were appropriate for investigating the research question
Rigour in study design or validity of theoretical approach	This domain assesses whether the study approach was documented clearly and whether it was based on a theoretical framework (such as ethnography or grounded theory). This does not necessarily mean that the framework has to be stated explicitly, but a detailed description ensuring transparency and reproducibility should be provided
Sample selection	This domain assesses the background, the procedure and reasons for the method of selecting participants. The assessment should include consideration of any relationship between the researcher and the participants, and how this might have influenced the findings
Data collection	This domain assesses the documentation of the method of data collection (in-depth interviews, semi-structured interviews, focus groups or observations). It also assesses who conducted any interviews, how long they lasted and where they took place
Data analysis	This domain assesses whether sufficient detail was documented for the analytical process and whether it was in accordance with the theoretical approach. For example, if a thematic analysis was used, the assessment would focus on the description of the approach used to generate themes. Consideration of data saturation would also form part of this assessment (it could be reported directly or it might be inferred from the citations documented that more themes could be found)
Results	This domain assesses any reasoning accompanying reporting of results (for example, whether a theoretical proposal or framework is provided)

Assessing relevance of evidence in qualitative reviews

Relevance (applicability) of findings in qualitative research is the equivalent of indirectness for quantitative outcomes, and refers to how closely the aims and context of studies contributing to a theme reflect the objectives outlined in the guideline review protocol.

Assessing coherence of findings in qualitative reviews

For qualitative research, a similar concept to inconsistency is coherence, which refers to the way findings within themes are described and whether they make sense. This concept was used in the quality assessment across studies for individual themes. This does not mean that contradictory evidence was automatically downgraded, but that it was highlighted and presented, and that reasoning was provided. Provided the themes, or components of themes, from individual studies fit into a theoretical framework, they do not necessarily have to reflect the same perspective. It should, however, be possible to explain these by differences in context (for example, the views of healthcare professionals might not be the same as those of family members, but they could contribute to the same overarching themes).

Assessing adequacy of data in qualitative reviews

Adequacy of data (theme saturation or sufficiency) corresponds to a similar concept in primary qualitative research in which consideration is made of whether a theoretical point of theme saturation was achieved, meaning that no further citations or observations would provide more insight or suggest a different interpretation of the theme concerned. As noted above, it is not equivalent to the number of studies contributing to a theme, but rather to the depth of evidence and whether sufficient quotations or observations were provided to underpin the findings.

Reviewing economic evidence

Inclusion and exclusion of economic studies

Systematic reviews of economic literature were conducted in all areas covered in the guideline. Titles and abstracts of articles identified through the economic literature searches were assessed for inclusion using the predefined eligibility criteria listed in Table 9.

Table 9: Inclusion and exclusion criteria for systematic reviews of economic evaluations

Inclusion criteria
Only studies from Organisation for Economic Co-operation and Development member countries were included, as the aim of the review was to identify economic information transferable to the UK context. For each review question and each strategy (intervention or service delivery model/setting), the focus of the economic literature review was on UK evidence.

Inclusion criteria
• For review questions that were supported by guideline economic modelling, only UK economic studies were included in the review. • For the remaining review questions that were not supported by economic modelling, UK evidence on each strategy was sought first; if no UK economic evidence was identified or the UK evidence was very thin (i.e. if it came from a single UK study or was characterised by very serious limitations), then a hierarchy of criteria were used to include studies in the economic review according to the country of origin, considering the similarities of each country's health system to the UK NHS, as follows: ○ Economic studies from Europe, Canada, Australia and New Zealand ○ Economic studies from the US ○ Economic studies from the remaining OECD countries (Chile, Mexico, Turkey, Israel, Japan, Korea) The described hierarchy for identification of eligible studies was agreed by the GC and the Health Economist and was followed until at least 2 economic studies were identified for each intervention or model of care considered in every review question; if less than 2 studies were identified, then studies meeting the next criterion in the hierarchy were sought.
Only studies published from 2002 onwards were included in the review. This date restriction was imposed so that retrieved economic evidence was relevant to current healthcare settings and costs.
Study population, interventions and comparators in accordance with the guideline scope and review protocols for each review question
Studies were included provided that sufficient details regarding methods and results were available to enable the methodological quality of the study to be assessed, and provided that the study's data and results were extractable.
Full economic evaluations that compared two or more relevant options and considered both costs and consequences were included in the review (i.e. cost-utility, cost-effectiveness, cost-benefit or cost-consequence analyses)
Economic studies were included if they used clinical effectiveness data from a randomised or non-randomised clinical trial, a prospective cohort study, or a review and meta-analysis of clinical studies. Economic analyses that utilised data from studies with a mirror-image design and studies that recruited participants retrospectively were not considered in the review, due to their lower methodological quality.
The outcome measure of the economic analyses should be the Quality Adjusted Life Year (QALY) or one of the measures considered in the clinical review.
Studies should be reporting separately costs from a healthcare (and, if available, personal social services) perspective.
Exclusion criteria
Poster presentations, conference or dissertation abstracts and letters containing insufficient methodological details
Non-English language papers
Cost-of-illness type studies
Non-comparative studies
Before-and-after studies and studies based on retrospective analyses of administrative healthcare data, due to associated methodological limitations and overall low quality characterising these study designs.

Inclusion criteria
Studies that considered exclusively intervention costs, e.g. drug acquisition costs, without considering wider healthcare costs associated with the management of depression. In addition, studies that considered an employer's perspective and included only productivity losses and/or benefit payments.
Studies that compared costs of branded vs generic forms of the same drug

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, summaries of economic evidence, economic evidence tables and health economic evidence profiles for each review question are presented in respective evidence reports (appendix G, H and I, respectively). Full lists of included and excluded economic studies and studies reporting utility data are provided in Supplement 3.

Appraising the quality of economic evidence

The applicability and quality of economic evidence, including economic evidence derived from primary economic modelling conducted for the guideline, was assessed using the economic evaluations checklist specified in Developing NICE guidelines: the manual, Appendix H, for all studies that met the inclusion criteria.

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

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

Inclusion and exclusion of health state utility studies

Literature on the health-related quality of life of adults with depression was systematically searched to identify studies reporting appropriate utility scores that could be utilised in a primary economic modelling. The titles and abstracts of papers identified through the searches were independently assessed for inclusion using predefined eligibility criteria defined in Table 10.

Table 10: Inclusion and exclusion criteria for the systematic review of health state utility values

Inclusion criteria
Studies from Organisation for Economic Co-operation and Development member countries

Only studies published from 2002 onwards were included in the review, so that evidence were relevant to current healthcare settings and preferences.
Studies were included provided that sufficient details regarding methods and results were available to enable the methodological quality of the study to be assessed, and provided that the study's data and results were extractable.
To be included, studies should report utility data for specific health states associated with depression through the care pathway.
HRQoL should have been rated directly from adults with depression using the EQ-5D valued by the general UK population, according to NICE recommendations (NICE 2014). If no such studies were available, then a hierarchy of criteria were used to include studies in the review, as follows: • use of SF-6D utility data, derived using the UK algorithm for valuation (Brazier 2002) • use of EQ-5D valued by a population of another country • use of another validated generic PBM (e.g. SF-6D valued by a non-UK population, HUI-3) • use of a condition-specific PBM valued by general population (UK data prioritised over non-UK ones) using TTO or SG techniques • use of vignettes valued by the general population (UK data prioritised over non-UK ones) using TTO or SG • use of condition-specific PBM valued by service users (UK data prioritised over non-UK ones) using TTO or SG • use of vignettes valued by service users using TTO or SG, or direct service user valuations of their own HRQoL (UK data prioritised over non-UK ones).
Exclusion criteria
Poster presentations, dissertation abstracts, abstracts in conference proceedings, letters
Non-English language papers
Studies reporting an overall utility score for people with depression (and/or people without depression), who might have a mixture of depression-related health states or a range of symptom severity

HRQoL: health-related quality of life; PBM: preference-based measure; SG: standard gamble; TTO: time trade-off

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

Utility studies that met inclusion criteria and those that were excluded after full text was obtained are listed in supplement 3.

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, as recommendations on these areas need to be supported by robust evidence on cost effectiveness.

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

- cost effectiveness of pharmacological, psychological, physical and combined interventions for adults with a new episode of less severe depression. The methods and results of the de novo economic analysis are fully reported in appendix J of evidence review B.
- cost effectiveness of pharmacological, psychological, physical and combined interventions for adults with a new episode of more severe depression. The methods and results of the de novo economic analysis are fully reported in appendix J of evidence review B.
- cost effectiveness of pharmacological, psychological and combined pharmacological and psychological interventions for preventing relapse in adults whose depression has responded to treatment. The methods and results of the de novo economic analysis are fully reported in appendix J of evidence review C.

When relevant economic evidence was not available and 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 The NICE Principles sets out the principles that committees should consider when judging whether an intervention offers good value for money. In general, an intervention was considered to be cost effective if any of the following criteria applied (provided that the estimate was considered plausible):

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more effective compared with all the other relevant alternative strategies)
- the intervention cost less than £20,000 per QALY gained compared with the next best strategy
- the 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 'The committee's discussion of the evidence' under subheading 'Cost effectiveness and resource use' in the relevant evidence reviews.

Additional sources of evidence

In addition to the evidence obtained from the systematic review process, the committee was also made aware of another guideline in development at the same time as the depression guideline. This guideline was called 'Medicines associated with dependence or withdrawal symptoms: safe prescribing and withdrawal management for adults' and further details can be found on the NICE website page for this guideline. The scope of this guideline included antidepressants. In order to update the recommendations in the depression guideline on starting and stopping antidepressants and to ensure that the 2 guidelines did not produce conflicting recommendations the committee discussed the completed evidence reviews produced for the safe prescribing guideline and take them into consideration.

The safe prescribing evidence reviews presented to the depression guideline committee were as follows:

Evidence review A: patient information and support

Evidence review B: prescribing strategies

Evidence review C: safe withdrawal

Evidence review D: withdrawal interventions

Evidence review F: monitoring

A further evidence review (Evidence review E: risk factors) was not presented to the committee as it did not include any evidence relating to antidepressants.

Developing recommendations

Guideline recommendations

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

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

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

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.

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.

Funding

The NGA was commissioned by NICE to develop this guideline.

References

Brazier 2002

Brazier, J., Roberts, J., Deverill, M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics, 21, 271-92, 2002.

Braun 2006

Braun, V., Clarke, V. Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101, 2006.

Caldwell 2005

Caldwell, D. M., Ades, A. E., & Higgins, J. P. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ, 331(7521), 897-900., 2005.

Caldwell 2014

Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. Syst Rev, 3:109, 2014.

Dias 2010

Dias, S., Welton, N. J., Marinho, V. C. C., Salanti, G., Higgins, J. P. T., Ades, A. E. Estimation and adjustment of bias in randomised evidence by using Mixed Treatment Comparison Meta-analysis. Journal of the Royal Statistical Society (A), 173(3), 613-629, 2010.

Dias 2011a

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

Dias 2011b

Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials, 2011b (last updated 2014) http://www.nicedsu.org.uk

Dias 2013

Dias, D., Welton, N. J., Sutton, A. J., Caldwell, D. M., Lu, G., Ades, A.E. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Medical Decision Making, 33, 641-56, 2017.

Higgins 2011

Higgins JPT, Green S (editors) (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated 2019] The Cochrane Collaboration. Available from www.handbook.cochrane.org (accessed 15 June 2018)

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 Health, 17(2), 157-73., 2014.

Lewin 2015

Lewin S, Glenton C, Munthe-Kaas H et al. (2015) Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 12(10), e1001895

Linde 2016

Linde K, Rücker G, Schneider A, Kriston L. Questionable assumptions hampered interpretation of a network meta-analysis of primary care depression treatments. J Clin Epidemiol, 71, 86-96., 2016.

Lu 2004

Lu, G., & Ades, A. E. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med, 23(20), 3105-3124, 2004.

Mavridis 2015

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

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

National Institute for Health and Care Excellence (NICE) Developing NICE guidelines: the manual. Available from https://www.nice.org.uk/process/pmg20/chapter/introduction (accessed 14 September 2021)

NICE

National Institute for Health and Care Excellence (NICE) NICE Policy on conflicts of interest. Available from https://www.nice.org.uk/about/who-we-are/policies-and-procedures (accessed 14 September 2021)

Phillippo 2019

Phillippo, D. M., Dias, S., Welton, N. J., Caldwell, D. M., Taske, N., Ades, A. E. Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. Ann Intern Med, 170(8), 538-46, 2019.

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

Welton 2009

Welton, N.J., Ades, A. E., Carlin, J. B., Altman, D. G., & Sterne J. A. C. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society (A), 172(1), 119-136, 2009.