Mental Health Gap Action Programme (mhGAP)
Guidelines on Interventions for Mental, Neurological and Substance use Disorders

WHO mhGAP Guideline Process 2009

This document describes the methodology followed for the process of guideline development and the standard procedures and methods that were followed for the evidence review and formulation of recommendations.
Background

Why this guideline was developed and the target audience

Mental, neurological, and substance use (MNS) disorders are prevalent throughout the world and are major contributors to morbidity and premature mortality. The treatment gap for these disorders is more than 75% in many low- and middle-income countries (LAMIC). Substantial initiatives have been made by in the last decade to bring mental health onto public health agenda; however, the task is far from complete. To address this challenge, WHO launched the mental health Gap Action Programme (mhGAP) to scale up mental health services, especially in LAMIC’s.

An essential component of mhGAP is to develop a model intervention guide for MNS disorders identified as conditions of high priority for LAMIC. The priority conditions were identified on the basis of high mortality and morbidity, high economic costs, or association with violation of human rights within the area of MNS disorders. These are depression, schizophrenia and other psychotic disorders (including bipolar disorder), suicide prevention, epilepsy, dementia, disorders due to use of alcohol and illicit drugs, and mental disorders in children.

Recommendations (i.e., guidelines) on interventions for the management of such high priority conditions form the basis of the mhGAP model intervention guide. Interventions are targeted to health care providers working at a first and second level facility in a health center at a
peripheral level or at district level. The first and second level facility includes the basic outpatient and inpatient services provided at these levels. The health care providers could be doctors, nurses, or other cadre of health workers.

WHO has recognized the need to apply more rigorous processes to ensure that the health care recommendations are informed by the best available research evidence and measures are taken towards more evidence based guidelines development. WHO has also stressed on the promotion of evidence base debate in its 11th General Programme of Work for 2006-2015 and other basic documents and has revised the process of developing guidelines.

**Guidelines Group and collaboration with external partners**

An international panel of individual experts and institutions with appropriate background experience - clinicians, researchers, programme managers, policy makers - who can provide inputs in the guidelines development process were identified and a global network of experts was developed as a first step. The multidisciplinary expertise that were sought included guideline development methodology, mental health, neurology, substance use, primary care, public health, epidemiology, policy making, health economics and caregivers.

A subset of experts from the international panel of experts with the multidisciplinary expertise and with adequate regional and gender representation were identified. This was the Guideline Development Group (GDG) which was convened to advise on the content and process, interpretation of evidence, to formulate and finalize the recommendations.

WHO funded the GDG meetings and processes for evidence retrieval, synthesis and making recommendations.

A conflict of interest declaration as per WHO rules was filled by all who were invited to participate in any way in the development of a guideline. The GDG members filled the WHO declaration of interest form before each GDG meeting. At each meeting of the guideline group,
all members provided a verbal summary of their written declaration of interest. None of the declared interest by the GDG members was considered significant conflict of interest by the GDG.

**Retrieving summarizing and presenting the evidence**

*Formulating questions and choosing outcomes*

For each of the priority conditions identified by mhGAP, a number of scoping questions and outcomes were initially developed by electronic consultation with the international panel of experts. In addition to the eight priority conditions, scoping questions and outcomes were also developed for "other common complaints". The scoping questions focused on areas of controversy that needed to be answered or topics where changes in policy or practice were needed. Scoping questions were formulated using the PICO framework (Population, Intervention, Comparator, Outcome, Time). In addition to scoping questions, identified outcomes were rated as critical, important or not important, for evidence review and making decisions and recommendations. The key scoping questions and outcomes and their rating were then finalized by the GDG in a meeting at WHO headquarters in Geneva organized in November 2008. The GDG divided its workload along clinically relevant lines to simplify the process, and GDG members formed smaller topic groups to undertake guideline work in that area of clinical practice. Each topic group was led by a GDG member with expert knowledge of the topic area. Topic groups refined the clinical questions, clinical definitions of treatment interventions, choosing and rating outcomes, planned the literature search strategy and helped the GDG to identify further expertise in the topic. A second GDG meeting was held in September 2009 to review the evidence profiles which also incorporated values, preferences and feasibility considerations, and to finalize the recommendations.
The process for identifying, appraising, synthesizing and grading the relevant evidence

For each scoping question, individual experts or institutions familiar with systematic reviews and evidence retrieval, evaluation and synthesis were identified and contacted. The experts and WHO focal points formed review teams that as a first step searched for systematic reviews from databases and existing evidence-based guidelines which have tried to answer the same or similar scoping question. In case no evidence in a synthesized form such as systematic review or evidence based guideline was available, new systematic reviews were commissioned. The review teams summarized the evidence base using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology and the GRADE profiler software, a tool developed to help transparently summarize the evidence and grade its quality. In case, it was not possible to GRADE the evidence, a study by study was developed to summarise the evidence.

Considering that a high number of scoping questions were being addressed and a large number of experts were involved, in order to increase consistency among reviewers in how the evidence was retrieved, synthesized, summarized and reported, templates were developed. These are enclosed as Annex 1 and include:

- 1i) Instructions for evidence retrieval
- 1ii) Identification of systematic review
- 1iii) Instructions for making recommendations
- 1iv) Checklist for values, preferences and feasibility issues

Template for recording evidence profile: The template required to include a PICO table; a list of systematic reviews identified by the search process with rationale for inclusion and exclusion (for each outcome); the GRADE tables (produced by the GRADE software); and a narrative description of evidence to accompany GRADE tables. The narrative description included, for
example, long-term safety and tolerability data, or results of observational or qualitative
studies or any other information that was considered relevant with respect to the outcome.

A few pragmatic instructions to guide the process of applying GRADE were developed to
increase consistency by ensuring that the same background logic was employed in assessing the evidence
for each scoping question. These instructions were used to rate the five aspects considered in the
quality assessment: (1) limitations (risk of bias); (2) inconsistency; (3) indirectness; (4) imprecision;
(5) reporting bias. The criteria for GRADE are enclosed as Annex

2. After agreement was achieved on the logic behind the grading, the review teams graded the
evidence. One rater graded the quality of evidence for each outcome and ratings were checked for
consistency by a second person. Agreement between raters was reached after discussion (a third
rater was sometimes involved).

Formulation and grading of the recommendations (including how resource use/costs, values and
preferences, and benefits and harms were considered)

A two-step process was followed for drafting the recommendation. Firstly, for each scoping
question, the review teams summarized the evidence included in the GRADE tables or study be study
tables and its quality ratings, and drafted a narrative description of the balance between desirable
and undesirable effects. This formed the basis for the zero draft recommendation. Secondly, the
review teams were required to take into consideration values, preferences and feasibility issues. A
checklist of aspects that deserve consideration was developed to increase consistency across
scoping questions. These aspects ranged from inclusion in the WHO list of essential medicines (for
drugs), and likely availability of medication in LAMIC, to number of sessions and number of
minutes per sessions required (for non-pharmacological interventions). An explicit judgment, to
be reported in the same table,
described whether each intervention poses a concern, or represents a positive effect, in terms of values and preferences that are considered priority issues in the field of mental health: promotion of social inclusion, protection of human rights and dignity, prevention of discrimination (and stigma), prevention of medicalization of social problems, promotion of individual and family members’ capacity and skills. A modified recommendation, based on the evidence base and on the considerations on values, preferences and feasibility issues, was then drafted and reported in the table.

Each step of the process from evidence to recommendation was drafted as preliminary material to be submitted to the global network of experts for peer review, and to the GDG for comments, revision and approval. On the basis of considerations on the balance between desirable and undesirable effects, quality of evidence, values, preferences and feasibility issues, GDG was additionally asked to provide a judgment on the strength of each recommendation, to be categorized as strong or standard. We anticipated that, in some circumstances, recommendations may apply only if specific conditions are met.

Consensus, external review and updating

The scoping questions and outcomes for the questions were circulated for peer-review to experts from different regions of the world. An audit was kept for all the comments received from the external reviewers and these were discussed and agreed upon with the GDG members during the first GDG meeting. The evidence profiles which included selected systematic reviews, synthesized evidence in form of evidence or GRADE table was circulated to selected peer-reviewers and their comments were incorporated. The evidence profiles and draft recommendations were also circulated to selected peer-reviewers and their comments discussed during second GDG meeting.

During the second GDG meeting to review the evidence profiles and recommendations, the
GDG divided its workload along clinically relevant lines to simplify the process, and GDG members formed smaller work groups. A moderator for each of the work groups was appointed by GDG. During the work group process, the moderator in coordination with WHO focal point presented the evidence profiles and draft recommendation and the prior comments received from GDG. Following discussion within each working group, relevant changes were made on the master e-copy by WHO focal point and final recommendation along with its strength was noted. For each of the section of the evidence profiles - summary of the evidence base, quality of evidence, balance of benefits and harms, values and preferences, and resource use and feasibility issues, it was noted whether there was full agreement or agreement with additional non-agreed suggestions or dissent or no agreement with summary of reasons. During the plenary session of the GDG, small group work was presented by the moderator and WHO focal person. Agreement was reached by consensus. In case of no consensus and if further information was required, that particular point was deferred to last day of the meeting. With additional information by the Moderator and WHO focal point, further discussions were held on the last day and agreement was reached by consensus. Consensus was not reached in case of one scoping question, in which case voting was done by the GDG members and agreement recorded based on the majority decision.

In addition to the 9 areas, WHO Secretariat had also prepared background material on overall management issues based on well recognized international standards. The GDG also discussed the recommendations on these overall management issues and unanimous consensus was achieved.

Detailed evidence profiles were not felt necessary by GDG.

The "review by" date was discussed by GDG members and it was decided that Department of Mental Health and Substance Abuse will review the update for the guidelines after 5 years.
members formed smaller work groups. A moderator for each of the work groups was appointed by GDG.

During the work group process, the moderator in coordination with WHO focal point presented the evidence profiles and draft recommendation and the prior comments received from GDG. Following discussion within each working group, relevant changes were made on the master e-copy by WHO focal point and final recommendation along with its strength was noted. For each of the section of the evidence profiles - summary of the evidence base, quality of evidence, balance of benefits and harms, values and preferences, and resource use and feasibility issues, it was noted whether there was full agreement or agreement with additional non-agreed suggestions or dissent or no agreement with summary of reasons.

During the plenary session of the GDG, small group work was presented by the moderator and WHO focal person. Agreement was reached by consensus. In case of no consensus and if further information was required, that particular point was deferred to last day of the meeting. With additional information by the Moderator and WHO focal point, further discussions were held on the last day and agreement was reached by consensus. Consensus was not reached in case of one scoping question, in which case voting was done by the GDG members and agreement recorded based on the majority decision.

In addition to the 9 areas, WHO Secretariat had also prepared background material on overall management issues based on well recognized international standards. The GDG also discussed the recommendations on these overall management issues and unanimous consensus was achieved. Detailed evidence profiles were not felt necessary by GDG.

The "review by" date was discussed by GDG members and it was decided that Department of Mental Health and Substance Abuse will review the update for the guidelines after 5 years.
Annex 1

1i) Instructions for evidence retrieval

1. From the scoping question

- Identify intervention/s and comparator/s

- Identify critical and important outcomes for each of the intervention/comparator

2. Develop the PICO table (see the template, annex 1) by patient population (P), Intervention (I), Comparator (C), and outcome (O).

3. Perform a search to identify existing systematic reviews for the different intervention/comparators for the patient population

- systematic search of evidence including Cochrane database, BMJ clinical evidence, existing evidence based guidelines (NICE, SIGN, Health Technology Assessment etc.), and scientific databases

4. Identify the systematic review which will be used for GRADE for each of the outcome for the different intervention/comparators, and explain the rationale (see annex 2 for general suggestions for selecting the systematic review). Complete this information in the PICO table.

5. Discuss and agree with WHO focal point on the PICO table including the systematic review to be GRADEd.

6. GRADE the evidence following the practical instructions (see attached, annex 3) for each of the outcome for the different intervention/comparators.

7. Share the GRADE table/s with the WHO focal point who will check for consistency. Agreement would need to be reached between the rater and the WHO focal point.

8. Provide additional information which was not GRADEd for the outcomes, and any other relevant information which might be of use for making recommendations.

OUTPUTS

A. Filled in template (see Annex 1)
B. Search strategy and outputs of the systematic search (refer to point number 3 of instructions)
C. GRADE files
D. Paper/electronic copies of systematic reviews that were used for GRADE

Link for downloading the GRADE software

http://www.cc-ims.net/revman/gradepro/download

Additional references and reading material - WHO handbook for guidelines
1ii) Identification of systematic review

**IDENTIFICATION OF SYSTEMATIC REVIEW**

In case of >1 systematic review for the intervention/comparator and outcome combination according to PICO table, identify the ones being used for GRADE, and explain the rationale

- Comment on search strategy
- Comment on whether the interventions and outcomes defined in scoping question match those defined by the systematic review
- Check criteria for inclusion and exclusion, and comment on whether they are appropriate
  - Comment on how quality-validity of included studies was assessed and method of data abstraction
  - Comment on heterogeneity of results, including whether or not a test for it was done
  - Comment on statistical analyses performed (e.g. meta-analysis)
  - Comment on applicability in terms of: 1) LAMIC, 2) community care/first and second levels of care, 3) nonspecialist healthcare providers
  - Comment on limitations of systematic review itself

Retrieve all the relevant articles

---

**PICO framework:**

1. Population: what group of people are we studying?
2. Intervention: what specific intervention are we testing?
3. Comparator: what are we comparing the intervention to? Another intervention? Standard care?
4. Outcome: what are

**GRADE criteria:**

1. limitations/risk of bias: this includes limitations of individual studies included
2. inconsistency: heterogeneity
3. indirectness: includes appropriateness of the study populations in terms of age, gender, etc.
4. imprecision: related to size of the sample population
5. reporting bias: relates to both bias found in individual studies, as well as potential bias resulting from the search strategy or quality assessment used, such as

---

**When is GRADEing unnecessary?**

1. For epidemiologic and prognostic studies, use study by study table
2. If the systematic review does not directly apply to the scoping question and/or not used for
1iii) Instructions for making recommendations

From evidence to recommendations

<table>
<thead>
<tr>
<th>Factor</th>
<th>Explanation</th>
<th>GDG comments and agreement (agree, agree with modifications, revise and resubmit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative summary of the evidence base</td>
<td>Provide a narrative summary of the evidence that is included in the GRADE tables and report the overall summary estimates (including confidence intervals) for all critical outcomes. Additional information on critical outcomes (number of included trials, number of included patients, absolute differences, baseline risks), or outcomes rated as important but not critical, should not be reported on a routine basis (refer to GRADE tables). If in the GRADE tables one critical outcome is measured with more than one parameter (for example adverse effects may be measured using the total number of dropouts and the total number of subjects experiencing at least one adverse event) then the parameter closest to the outcome of interest should be reported. If there are many interventions or patient populations or outcomes, a tabular format may be chosen.</td>
<td></td>
</tr>
<tr>
<td>Summary of the quality of evidence</td>
<td>From GRADE tables.</td>
<td></td>
</tr>
<tr>
<td>Balance of benefits versus harms</td>
<td>Provide a narrative description that takes into consideration the evidence included in the GRADE tables plus any other evidence considered in the narrative part.</td>
<td></td>
</tr>
<tr>
<td>Zero draft recommendation</td>
<td>This draft recommendation should be based on the evidence that is reported above.</td>
<td></td>
</tr>
<tr>
<td>Define the values and preferences including any variability and human rights issues.</td>
<td>Use the checklist of values and preferences to make a judgment on whether the intervention may be expected to be in line or pose a problem with the values and preferences. Only report the most salient issues of concern in this box.</td>
<td></td>
</tr>
<tr>
<td>Define the costs and resource use and any other relevant feasibility issues.</td>
<td>Use the checklist of feasibility issues to make a judgment on whether the intervention may be expected to be doable in most nonspecialized health care settings. Only report the most salient issues of concern in this box.</td>
<td></td>
</tr>
</tbody>
</table>

**Modified recommendation**

This modified recommendation should be based on the evidence base and on the considerations on values, preferences and feasibility issues reported above.

**Final recommendation**

**Strength of recommendation (Strong/standard/conditional)**

**Context specific issues**

**Research priorities**
1iv) Checklist for values, preferences and feasibility issues

In the recommendation table mention those that apply (in favor or against the intervention).

FEASIBILITY ISSUES

Inclusion in the WHO list of essential medicines and likely availability of medication in LAMIC
Acquisition cost (also relation between cost and effectiveness)
Current treatment skills availability in LAMIC for this intervention
Specific training requirements (comment if > 1.0 day just for this specific intervention)
Number of sessions, number of minutes per sessions required
Specific laboratory requirements
Other equipment requirements
Continuous supply of medication (comment if sudden disruption of supply could have harmful consequences, e.g. for anti-epileptics)
Specific supervision requirements (comment if more than 1 supervisory discussion is needed per 3 months)
Any other feasibility issues

VALUES/PREFERENCES

Promotion of social inclusion
Protection of human rights and dignity (e.g. interventions that are sometimes provided on a non-voluntary basis)
Prevention of discrimination (and stigma)
Prevention of medicalization of social problems
Promotion of individual and family members’ capacity and skills
Any other values

NOTE: Many of these aspects are not absolute concepts, and their relevance may vary according to local context characteristics.
Annex 2

Practical instructions for assessing the quality of evidence included in systematic reviews

GENERAL PRINCIPLES

In order to assess the quality of evidence using the GRADE template, it is essential that raters agree on basic criteria to be used to downgrade or upgrade the evidence. This will enhance the consistency and reliability of ratings.

General principles:

(1) A first rater will grade the quality of evidence for each outcome, and will summarize findings using the GRADE template for each outcome. Ratings will be checked for consistency by a second member of the review group. Agreement between raters should be reached (a third rater might be involved in case of disagreement).

(2) When assessing quality of evidence you may follow this diagram.

(3) When assessing the quality of evidence you may want to follow the GRADE general approach:
- GRADE is not a quantitative system for grading the quality of evidence. Each factor for downgrading or upgrading reflects not discrete categories but a continuum within each category and among the categories. When the body of evidence is intermediate with respect to a particular factor, the decision about whether a study falls above or below the threshold for up- or downgrading the quality (by one or more factors) depends on judgment.

- Despite the limitations of breaking continua into categories, treating each criterion for rating quality up or down as discrete categories enhances transparency. Indeed, the great merit of GRADE is not that it ensures reproducible judgments but that it requires explicit judgment that is made transparent to users.

**NOTE:** Observational studies that have been downgraded to very low quality for any reason should not be upgraded.

(4) To achieve transparency and implicility, the GRADE system classifies the quality of evidence in one of four grades:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Further research is very unlikely to change our confidence in the estimate of effect.</td>
</tr>
<tr>
<td>Moderate</td>
<td>Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.</td>
</tr>
<tr>
<td>Low</td>
<td>Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.</td>
</tr>
<tr>
<td>Very low</td>
<td>Any estimate of effect is very uncertain.</td>
</tr>
</tbody>
</table>

(4) According to the GRADE system, raters are required to make a judgement on studies included in systematic reviews with respect to the following criteria:

(1) LIMITATIONS (RISK OF BIAS)
(2) INCONSISTENCY
(3) INDIRECTNESS
(4) IMPRECISION
(5) REPORTING BIAS

(1) **LIMITATIONS** (risk of bias)

Definition. Limitations in the study design and implementation may bias the estimates of the treatment effect. Our confidence in the estimate of the effect and in the following recommendation decreases if studies suffer from major limitations. The more serious limitations are, the more likely it is that the quality of evidence will be downgraded. Our confidence in an estimate of effect decreases if studies suffer from major limitations that are likely to result in a biased assessment of the intervention effect. For randomized trials, the following limitations are likely to result in biased results: lack of allocation concealment, lack of blinding, incomplete accounting of patients and outcome events, selective outcome reporting, other (for further details see the GRADEprofiler instructions).

The following criteria will be followed by WHO raters on LIMITATIONS:

If one or more of the three criteria reported below is not met in up to 10% of trials included in the systematic review = no downgrading (negligible limitations)

If one or more of the three criteria reported below is not met in 10-30% of trials included in the systematic review = - 1 (serious limitations)

If one or more of the three criteria reported below is not met in more than 30% of trials included in the systematic review = - 2 (very serious limitations)

The three criteria are:
(1) trials are described as randomised;
(2) outcome assessment is described as masked;
(3) dropout rate (both treatment arms) is below or equal to 30% (and dropouts are similarly distributed between treatment arms).

A different criterion may be followed by WHO raters in exceptional situations. Explanations should be reported as footnote in the corresponding GRADE table.

(2) INCONSISTENCY

Definition. Inconsistency refers to an unexplained heterogeneity of results across studies. Widely differing estimates of the treatment effect (i.e. heterogeneity or variability in results) across studies suggest true differences in underlying treatment effect. When heterogeneity exists, but investigators fail to identify a plausible explanation, the quality of evidence should be downgraded by one or two levels, depending on the magnitude of the inconsistency in the results (for further details see the GRADE profiler instructions). Inconsistency may arise from differences in:

- populations (e.g. drugs may have larger relative effects in sicker populations)
- interventions (e.g. larger effects with higher drug doses)
- outcomes (e.g. diminishing treatment effect with time).

Guideline panels or authors of systematic reviews should also consider the extent to which they are uncertain about the underlying effect due to the inconsistency in results and they may downgrade the quality rating by one or even two levels.

The following criteria will be followed by WHO raters on INCONSISTENCY:

If visual investigation of forest plots suggests some degree of heterogeneity (supported by a formal test of heterogeneity indicating some degree of heterogeneity, for example I-squared between 50% and 75%) = - 1 (serious inconsistency)

If visual investigation of forest plots suggests high degree of heterogeneity (supported by a formal test of heterogeneity indicating high heterogeneity, for example I-squared higher than 75%) = - 2 (very serious inconsistency)

A different criterion may be followed by WHO raters in exceptional situations. Explanations should be reported as footnote in the corresponding GRADE table.
NOTE: Raters will not downgrade for inconsistency when only one study contributes to the evidence base.

(3) INDIRECTNESS

Definition. There are two types of indirectness.
1. Indirect comparison – occurs when a comparisons of intervention A versus B is not available, but A was compared with C and B was compared with C. Such studies allow indirect comparisons of the magnitude of effect of A versus B. Such evidence is of lower quality than head-to-head comparisons of A and B would provide.
2. Indirect population, intervention, comparator, or outcome – the question being addressed by the guideline panel or by the authors of a systematic review is different from the available evidence regarding the population, intervention, comparator, or an outcome.

Indirectness may additionally refer to the extent to which the characteristics of those who will deliver the intervention in the real-world (including context characteristics) match with the characteristics of those who actually delivered the intervention under experimental conditions (in terms of background education, training, referral possibilities, context, other features).

Those making recommendations or authors of systematic reviews should consider the extent to which they are uncertain about the applicability of the evidence to their relevant question and downgrade the quality rating by one or even two levels.

The following criteria will be followed by WHO raters on INDIRECTNESS:

The question being addressed by the guideline panel is different from the available evidence regarding the population, intervention, comparator, outcome or regarding the characteristics of those who will deliver the intervention = - 1 (serious doubts about directness)

The question being addressed by the guideline panel is markedly different from the available evidence regarding the population, intervention, comparator, outcome or regarding the characteristics of those who will deliver the intervention = - 2 (very serious doubts about directness)

A different criterion may be followed by WHO raters in exceptional situations, explanations should be reported as footnote in the corresponding GRADE table.

NOTE: If only one study contributes to the evidence base, raters may consider if this affects directness and, if yes, downgrading may be appropriate.

(4) IMPRECISION

Definition. Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of the effect. In this case guideline panel will judge the quality of the evidence lower than
it otherwise would because of resulting uncertainty in the results (for further details see the GRADEprofiler instructions).

The following criteria will be followed by WHO raters on IMPRECISION:

If (a) the overall number of individuals included in trials is low (between 100 and 200 individuals, both treatment arms) or (b) the 95% confidence interval includes both 1) no effect and 2) appreciable benefit or appreciable harm = - 1 (serious imprecision)

If (a) the overall number of individuals included in trials is very low (less than 100 individuals, both treatment arms) and (b) the 95% confidence interval includes both 1) no effect and 2) appreciable benefit or appreciable harm = - 2 (very serious imprecision)

NOTE: For continuous outcomes “no effect” means a SMD with a confidence interval that crosses zero; appreciable benefit or appreciable harm means that the upper or lower confidence limit crosses an effect size of 0.5 in either direction. For dichotomous outcomes “no effect” means an estimate with a confidence interval that crosses one; appreciable benefit or appreciable harm means that the upper or lower confidence limit crosses a risk of 2.0 or 0.5.

AdifferentcriteriannaybefollowedbyWHO raters in exceptional situations. Explanations should be reported as footnote in the corresponding GRADE table.

(5) REPORTING BIAS

Definition. Publication bias is a systematic underestimate or an overestimate of the underlying beneficial or harmful effect due to the selective publication of studies. Publication bias arises when investigators fail to report studies they have undertaken (typically those that show no effect). Methods to detect the possibility of publication bias in systematic reviews exist, although authors of the reviews and guideline panels must often guess about the likelihood of publication bias. A prototypical situation that should elicit suspicion of publication bias occurs when published evidence is limited to a small number of trials, all of which are showing benefits of the studied intervention.

The following criteria will be followed by WHO raters on REPORTING BIAS:

If the graphical inspection of the funnel plot suggests some asymmetry, or if any other reasons (to be recorded as footnote) suggest that reporting bias might have had an impact on the overall summary estimate (for example: unpublished grey literature was not included) = - 1

If the graphical inspection of the funnel plot suggests high asymmetry, or if any other reasons (to be recorded as footnote) suggest that reporting bias might have had a high impact on the overall summary estimate (for example: unpublished grey literature was not included) = - 2
Different criterion may be followed by WHO raters in exceptional situations. Explanations should be reported as footnote in the corresponding GRADE table.

Upgrading the evidence

(6) DOSE-RESPONSE GRADIENT

Please note: In randomized trials and observational studies downgraded for any reason, donot rate the presence of dose-response gradient and choose no. You should assess if there was a dose-response gradient only in observational studies not downgraded for any reason.

The presence of a dose-response gradient may increase our confidence in the findings of observational studies and thereby increase the quality of evidence. Only observational studies with no threats to validity (not downgraded for any reason) can be upgraded.

(7) LARGE MAGNITUDE OF EFFECT

You should assess if the effect was large or very large and, if so, upgrade the quality of evidence accordingly for this outcome. For observational studies, only studies with no important threats to validity (not downgraded for any reasons) should be upgraded.

To rate magnitude of the effect:
- If the effect was not large (RR between 0.5 and 2.0) choose no
- If the effect was large (RR either >2.0 or <0.5 based on consistent evidence from at least 2 studies, with no plausible confounders) choose RR >2 or <0.5 «this will upgrade the quality of evidence for this outcome by 1 level»
- If the effect was very large (RR either >5.0 or <0.2 based on direct evidence with no major threats to validity) choose RR >5 or <0.2 «this will upgrade the quality of evidence for this outcome by 2 levels»
- Explain your choice in a footnote whenever you upgrade the quality of evidence for any reason, because it is important for others to understand your choice.

(8) EFFECT OF ALL PLAUSIBLE CONFOUNDING

On occasion, all plausible confounding from observational studies or randomized trials may be working to reduce the demonstrated effect or increase the effect if no effect was observed. For example, if only sicker patients receive an experimental intervention or exposure, yet they still fare better, it is likely that the actual intervention or exposure effect is larger than the data suggest. For observational studies, only studies with no important threats to validity (not downgraded for any reasons) should be upgraded.