14. Strong recommendations when the evidence is low quality

14.1 Background

To improve the guideline development process and facilitate the explicit, systematic and transparent formulation of recommendations from evidence, in 2007 WHO adopted GRADE (Grading of Recommendations Assessment, Development and Evaluation) as the method for assessing the quality of a body of evidence and for determining the direction and strength of the resulting recommendations (1) (see Chapters 9 and 10 of the WHO handbook for guideline development, 2014) (2).

According to the GRADE framework, the best estimates of the effects of an intervention come from systematic reviews of randomized controlled trials (RCTs) in which the intervention is tested against alternative management approaches. The certainty or level of confidence in an effect estimate depends on several factors, namely risk of bias, imprecision, indirectness, inconsistency and publication bias (3). GRADE rates certainty as high, moderate, low and very low based on a combination of these factors. Prior to consideration of these other factors, RCTs provide high-certainty estimates, whereas observational studies provide estimates having low certainty, except under unusual circumstances.

GRADE also provides guidance on how to formulate recommendations based on systematic reviews of the evidence (4). Recommendations can be strong or conditional, depending in part on the level of confidence (certainty) in the effects of a given intervention. When guideline development groups are confident that the desirable consequences (benefits) of an intervention outweigh its undesirable consequences (risks or harms), they will likely issue a strong recommendation in favour of the intervention; when they are confident that the opposite is true, they issue a strong recommendation against the intervention. In cases in which the balance between desirable and undesirable consequences is less certain, the guideline development group will issue a conditional recommendation.

In addition to the certainty surrounding effect estimates, several other factors influence the strength of a recommendation under the GRADE approach. These factors include the magnitude of the potential benefits and harms of alternative courses of action; value judgements on the trade-off between these harms and benefits; the level of uncertainty surrounding the
value judgements and preferences of the individuals affected by the recommendation; the extent to which these value judgements and preferences are estimated to vary across population groups; and considerations pertaining to the use of resources. In the context of public health guidelines, additional factors – i.e. the burden of illness; an intervention’s accessibility, feasibility and acceptability; social context; the extent of current suboptimal practice; and the intervention’s impact on health inequities – may be considered too.

The judgements regarding the relative value of the potential outcomes of an intervention and the preferences regarding the proposed intervention that are used to inform a recommendation should be those of the intended beneficiaries of the recommendation. To come to know these value judgements and preferences, the suggested approach under GRADE is to conduct a systematic review of studies in which they are explored. Although other methods can be employed for the purpose of developing guidelines, such methods are not well developed and often difficult to implement. Hence, guideline development groups should explicitly state the nature of the value judgements and preferences underpinning their recommendations and the source(s) of those data.

14.2 Cause for concern regarding WHO guidelines

According to a 2013 evaluation, the quality of WHO guidelines has improved markedly since the Guideline Review Committee (GRC) was formed in 2007 to promote and ensure consistent processes and standards in guideline development (5). Interviews with 20 WHO staff members revealed, however, that some technical units were purposely bypassing the established procedures. In addition, some staff said they were unsure of how GRADE should be applied, and members of the GRC expressed the concern that GRADE principles were not fully institutionalized. These results led the authors of the evaluation to conclude that the quality assurance standards formulated by the GRC were not yet fully embedded within the Organization.

Initial anecdotal evidence suggested a specific problem: WHO guideline panels frequently make strong recommendations despite evidence assessed as having low or very low certainty surrounding the effect estimates. Such recommendations are known as “discordant”. GRADE guidance warns against discordant recommendations because when either the benefits or harms of an intervention are uncertain, one cannot be confident that an intervention does more good than harm. Strong recommendations are directives that are meant to be followed by all or almost all guideline users and under all or almost all foreseeable circumstances. Because of this, discordant recommen-
dations may entrench practices whose benefit is uncertain. For instance, a discordant recommendation may lead the users of a WHO guideline to carry out interventions that are detrimental individually or collectively or to waste scarce resources on ineffective interventions.

Although in general discordant recommendations are not considered appropriate, the GRADE approach includes five situations in which such recommendations may be warranted (Table 1) (4). Researchers have explored the extent to which WHO guideline development groups are adhering to GRADE guidance – or failing to do so – when they issue discordant recommendations (6,7).

### 14.3 Detailed study of WHO guidelines

An initial descriptive study of WHO’s recommendations conducted in 2012 confirmed that WHO frequently develops discordant recommendations (6). From 2007 to 2012, WHO published 43 guidelines containing 456 recommendations using the GRADE approach. Of the 456 recommendations, 289 (63.4%) were strong and 167 (36.6%) were conditional. Of the 289 strong recommendations, 160 (55.4%) were discordant.

Researchers examined the 160 discordant recommendations in depth (7). First, they assessed them for consistency with one of the five situations in Table 1 (i.e. for consistency with GRADE guidance). They then classified all recommendations judged to be inconsistent with those situations (and hence with GRADE guidance) using a taxonomy previously developed by the Endocrine Society (United States of America) to assess its own guidelines (Table 2) (4,8). This taxonomy classifies recommendations in one of three ways: (i) as good-practice statements (9), when the certainty surrounding effect estimates is high but the supportive evidence is indirect (and thus the GRADE approach cannot be employed); (ii) as based on a misclassification of the evidence, usually owing to the erroneous grading of the certainty surrounding effect estimates as low or very low when moderate or high was warranted; or (iii) as inappropriately formulated as a strong recommendation when a conditional one would have been more appropriate.

The researchers judged only 25 (15.6%) of the 160 discordant recommendations to be consistent with one of the five situations in which discordant recommendations are considered appropriate (Table 1). The most common situation was that in which there was potential for catastrophic harm although the certainty surrounding the effect estimates of harm was low (7). Of the remaining 135 (84.4%) recommendations, researchers reported that...
### Table 1. Situations in which strong recommendations may be indicated despite low or very low confidence in effect estimates

<table>
<thead>
<tr>
<th>Situation</th>
<th>Confidence in effect estimates (evidence quality)</th>
<th>Benefits versus harms</th>
<th>Value judgements and preferences</th>
<th>Resource considerations</th>
<th>Type of recommendation</th>
<th>Example of a discordant recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life-threatening situation</td>
<td>Low or very low</td>
<td>Immaterial (very low to high)</td>
<td>Intervention may save lives in a life-threatening situation. Adverse events not prohibitive.</td>
<td>A very high value is placed on an uncertain but potentially life-saving benefit.</td>
<td>Small incremental cost or use of resources relative to benefits justifies the intervention.</td>
<td>Strong recommendation in favor of the intervention</td>
</tr>
<tr>
<td>Uncertain benefit, certain harm</td>
<td>Low or very low</td>
<td>High or moderate</td>
<td>Possible but uncertain benefit. Substantial established harm.</td>
<td>A much higher value is placed on the harmful effects, which are certain, than on the benefits, which are uncertain.</td>
<td>Possible high incremental costs or use of resources in the face of uncertain benefits may dictate the need for a recommendation against the intervention.</td>
<td>Strong recommendation against the intervention (or in favor of a less harmful/costly comparator)</td>
</tr>
<tr>
<td>Potentially equivalent options, one clearly less risky or costly than the other</td>
<td>Low or very low</td>
<td>High or moderate</td>
<td>Both alternatives show similar though uncertain benefits, but one is certainly less harmful or expensive than the other.</td>
<td>A high value is placed on avoiding harm.</td>
<td>High incremental cost (or resource use) relative to benefits may justify recommending the comparator, if less harmful.</td>
<td>Strong recommendation in favor of the less harmful/costly comparator</td>
</tr>
</tbody>
</table>

*In the treatment of patients with MDR-tuberculosis, a fluoroquinolone should be used (10).*

*We recommend against screening for androgen deficiency in the general population (11).*

*For management of post partum haemorrhage, oxytocin should be preferred over ergometrine alone, a fixed-dose combination of ergometrine and oxytocin, carbetocin, and prostaglandins (12).*
<table>
<thead>
<tr>
<th>Situation</th>
<th>Confidence in effect estimates (evidence quality)</th>
<th>Benefits versus harms</th>
<th>Value judgements and preferences</th>
<th>Resource considerations</th>
<th>Type of recommendation</th>
<th>Example of a discordant recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High confidence in benefits being similar, but one option potentially more risky or costly than the other</td>
<td>High or moderate</td>
<td>Low or very low</td>
<td>Have established that alternative management strategies afford similar benefits, but one of them may be more harmful than the other (low certainty).</td>
<td>High incremental cost (or resource use) of one intervention may justify recommending the comparator, if less harmful.</td>
<td>Strong recommendation against the potentially more harmful/costly comparator</td>
<td>In women requiring anticoagulation and planning conception or in pregnancy, the American College of Chest Physicians’ guidelines recommended against the use of certain anticoagulants (13). For example, high confidence estimates suggest similar effects of different anticoagulants. However, indirect evidence (low confidence in effect estimates) suggests potential harm to the unborn infant with oral direct thrombin (e.g. dabigatran) and factor Xa inhibitors (e.g. rivaroxaban, apixaban).</td>
</tr>
<tr>
<td>Potential catastrophic harm</td>
<td>Immaterial (very low to high)</td>
<td>Low or very low</td>
<td>Intervention potentially quite harmful, while its benefit varies in magnitude.</td>
<td>A high value is placed on avoiding greater harm.</td>
<td>High incremental cost (or resource use) of the potentially more harmful intervention may further justify recommending the less harmful comparator.</td>
<td>Strong recommendation against the intervention (or in favor of the less harmful/less expensive comparator). Children with suspected or confirmed pulmonary tuberculosis or tuberculous peripheral lymphadenitis living in settings with high HIV prevalence (or with confirmed HIV infection) should not be treated with intermittent regimens (14).</td>
</tr>
</tbody>
</table>

Adapted from Andrews et al. 2013 (4) and Alexander et al. (7).
29 (18% of the total 160) were good practice statements (9); 33 recommendations (20.6%) were based on certainty that had been misclassified (rated as low or very low quality instead of moderate or high); and 73 (45.6%) had been issued as strong when the recommendation should have been conditional.

Table 2 illustrates the various types of discordant recommendations not consistent with GRADE guidance. Of the three categories involved – i.e. good practice statements, recommendations based on misclassified evidence, and strong recommendations that should have been conditional – the third is of greatest concern. Good practice statements are sometimes appropriate. This is so when the supporting evidence is indirect and therefore difficult to collect and summarize but the certainty or confidence surrounding an intervention’s impact is high, and when the desirable consequences of an intervention clearly outweigh the undesirable ones. Good practice statements are, by nature, strong recommendations. This does not pose a problem as long as the certainty surrounding effect estimates is high.

Issuing strong recommendations when a conditional one would have been more appropriate does pose a problem, however, and is not trivial: 46% of the discordant recommendations and 16% of all recommendations published by WHO from 2007 to 2012 fell in this category.

14.4 Results of a qualitative study of discordant recommendations at WHO

Researchers conducted 13 open-ended, semi-structured, one-on-one interviews with the guideline development group chairs and technical staff who had developed one of the guidelines containing the highest proportion of discordant recommendations not consistent with GRADE guidance (Table 2). Four overarching themes emerged from these interviews: (i) the strengths of GRADE; (ii) the challenges and barriers involved in applying GRADE; (iii) ways to improve the use of GRADE; and (iv) the reasons for discordant recommendations.

More specifically, the interviews revealed the following:

Respondents identified and appreciated the strengths of GRADE, including its highly structured, open, analytical, standardized, evidence-focused approach and the attention paid to patient values and preferences in moving from evidence to recommendations.

One of the barriers or difficulties respondents encountered when applying GRADE was a lack of sufficient guidance on when and how to make conditional recommendations. They also had difficulty dealing with cases
### Table 2. WHO recommendations found to be inconsistent with GRADE guidance

<table>
<thead>
<tr>
<th>Condition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good practice statement (guideline development group should not apply GRADE methods). A large body of indirect evidence that is difficult to summarize indicates that the desirable consequences of the intervention far outweigh its undesirable consequences (i.e. confidence is high, but summarizing the evidence systematically would be a poor use of resources).</td>
<td>Triage people with tuberculosis symptoms (strong recommendation, low quality of evidence). These recommendations suggest that persons with a sufficiently high probability of having tuberculosis should be promptly separated from other patients and undergo the appropriate investigations (15).</td>
</tr>
<tr>
<td>Misclassification of the quality of the evidence as low or very low, rather than moderate or high.</td>
<td>Couples and partner voluntary HIV testing and counselling (CHTC) with support for mutual disclosure should be offered to individuals with known HIV status and their partners (strong recommendation, low-quality evidence for all people with HIV in all epidemic settings). In a randomized trial, couples who received CHTC versus health information reduced unprotected sex, providing moderate quality evidence supporting the recommendation (16).</td>
</tr>
<tr>
<td>Recommendation inconsistent with GRADE guidance which suggests the need for a conditional recommendation.</td>
<td>Uterine massage is recommended for the treatment of post-partum haemorrhage (strong recommendation, very low quality evidence). Because evidence supporting uterine massage is of very low quality and uterine massage might delay the institution of more effective interventions, a conditional recommendation would be optimal (17).</td>
</tr>
</tbody>
</table>

GRADE, Grading of Recommendations Assessment, Development and Evaluation. Adapted from Alexander et al. (7).
in which there was an absence of RCTs, and with the development of key questions in PICO (population, intervention, comparator, outcome) format.

To improve the use of GRADE, respondents suggested training sessions, particularly at the start of guideline development, with the use of web-based interactive videos showing how to develop questions in PICO format and match them to the evidence. They also expressed the need for greater leadership on the part of the technical units at WHO, and for more organizational support.

Respondents gave several reasons for discordant recommendations. They expressed general scepticism with respect to conditional recommendations and, more specifically, a reluctance to issue them for established or long-standing practices or where personal experiences suggested significant benefit. They also blamed funding, budget and policy issues and the fear of an unfavourable response from Member States’ ministries of health to conditional recommendations. Respondents were further concerned that conditional recommendations would be ignored particularly in resource-poor settings, and that only strong recommendations would ensure people’s access to treatment. Finally, respondents feared that conditional recommendations would send confusing signals to the public health community and be misinterpreted by health practitioners and policy-makers.

During interviews, respondents indicated that they generally felt very confident about the balance between an intervention’s benefits and harms related to the recommendation being formulated, even when the supporting evidence was rated as low or very low in quality. This revealed a clear misunderstanding of the GRADE system. In some instances, of course, the quality rating was an error and the respondents were right in feeling confident. This occurred most often when guideline development group members intuitively understood the importance of indirect evidence but were unaware of GRADE’s explicit guidance in this area. In other cases, respondents’ certainty was not supported by compelling evidence, direct or indirect, but was based on personal experience or anecdotal evidence instead.

14.5 Recommendations

Although many guideline development groups and their chairs strongly support the use of GRADE, correct and consistent application of the method continues to pose challenges. Staff involved in developing recommendations and members of guideline development groups at WHO need to learn how to apply the GRADE method appropriately. The remainder of this chapter explains key points involved in using GRADE correctly.
14.5.1 Approaches to improve the use of GRADE in WHO guidelines

14.5.1.1 Make sure comparators are explicit
When making discordant recommendations (or any recommendation using GRADE), guideline development groups must ensure that the comparator intervention is explicitly stated in the text of the recommendation or in the contiguous text.

14.5.1.2 Know when to issue a good practice statement
WHO guideline development groups often issue discordant recommendations when good practice statements would have been more appropriate. Good practice statements typically represent situations in which a large body of indirect evidence, often composed of several bodies of evidence linked together in the causal pathway, including indirect comparisons, unequivocally demonstrates the net benefit of the recommended action. GRADE provides guidance on good practice statements (9) and warns against their overuse. A good practice statement should be labelled as such, not as a recommendation, and clearly distinguished from a recommendation formulated using GRADE.

14.5.1.3 Rate confidence in effect estimates correctly
As shown by the study described earlier, members of the guideline development group often feel confident or certain that an intervention produces certain effects, even when the quality of the supporting evidence is rated as low or very low. Since under the GRADE approach the quality of the evidence is synonymous with the degree of confidence in effect estimates, this situation shows that GRADE is not well understood. In many instances, however, group members were correct in being confident in the effect estimates owing to an intuitive assessment of indirect evidence. In some of these cases, good practice statements were appropriately issued. In others, however, guideline development groups should have conducted a formal GRADE assessment and rated the confidence in effect estimates as moderate or high using GRADE’s explicit guidance on the use of indirect evidence. It is essential for WHO guideline development groups to understand the importance of indirect evidence.

14.5.1.4 Engage an experienced guideline methodologist
To ensure rigorous methods and adherence to GRADE, WHO guideline development groups need to engage a guideline methodologist with knowledge of GRADE and skills in group process. The role of the
methodologist should be well defined and should be clarified early in the process and well before any meeting of the guideline development group. Will this individual be a consultant or a co-chair? If the latter, how will his/her role and that of the other co-chair, usually a content expert, be divided? It may be useful to have training materials and hold educational sessions for methodologists too, as well as to develop strategies for dealing with the type of situations that lead guideline development groups to make discordant recommendations inconsistent with GRADE guidance.

14.5.1.5 **Educate guideline development group members**

Many guideline development groups have a poor understanding of GRADE and would benefit from education and training on the GRADE approach and its underlying principles. The training should focus on the development of questions in the PICO format, the meaning of certainty in effect estimates, the difference between strong and conditional recommendations, and the nature and value of indirect or observational evidence. The educational materials used should provide plenty of examples. WHO staff, as well as former members of WHO guideline development groups, should be involved in the development of these educational materials. Guideline development group members should have training before any meeting where the scope and key questions of the proposed guideline are defined and, particularly, before recommendations are formulated.

14.5.1.6 **Rely on established situations to formulate discordant recommendations**

Considerable thought has gone into characterizing situations in which discordant recommendations are appropriate, and five such situations have been defined (Table 1). When guideline development groups are considering drafting a discordant recommendation, they should determine if any of the five situations applies. WHO staff must make these situations known and help the guideline development group to understand them. These measures will probably result in fewer discordant recommendations, greater discussion of recommendations among guideline development group members, and the formulation of a clear rationale for any discordant recommendations.

14.5.1.7 **Use evidence-to-recommendation (decision) tables**

Evidence-to-recommendation (decision) tables have recently been developed, as described in Chapter 10 of the *WHO handbook for guideline development* (2). Experience with these tables has been positive so far, particularly in organizations making public health recommendations. Explicitly acknowl-
edging the factors that influence the direction and strength of each recommendation through the use of evidence-to-recommendation templates and in the remarks section of each guideline will result in more trustworthy guidelines.

14.5.2 Additional approaches to help ensure the development of valid recommendations

14.5.2.1 Carefully select the guideline development group chair
The choice of members of the guideline development group is important, but the choice of chair(s) is particularly so. Chairs with content area expertise should be skilled in managing a consensus-based group process and facilitating interaction among group members. Most importantly, they should understand GRADE thoroughly and be willing to work closely with the guideline methodologist to ensure that GRADE is correctly applied in rating the certainty of effect estimates and the strength of the recommendations.

14.5.2.1 Clearly identify the target audience
WHO recommendations are typically addressed to health workers, programme managers and policy-makers. WHO-sponsored guideline development groups should know the target audience of each recommendation and make this audience clear to the end-users of the guideline. If a recommendation has more than one target audience, guideline development groups may consider issuing a separate recommendation for each one. However, this should generally be avoided because it may prove confusing to both target audiences and if it is done, an explicit rationale must be given.

14.5.2.3 Avoid prejudicial influences
As noted earlier, study respondents named the following reasons for the formulation of discordant recommendations inconsistent with GRADE guidance: excessive reliance on personal experience; fear of health ministries’ reactions to a conditional recommendation; reluctance to conditionally recommend long-established practices; and the possibility that conditional recommendations might be ignored. Such influences must be identified and minimized to make optimal recommendations possible.

14.5.2.4 Identify and manage conflicts of interest
For recommendations and guidelines to be valid, it is critically important to try to identify financial and nonfinancial conflicts of interest, avoid them
where possible, and appropriately manage any existing conflicts. WHO staff must study and carefully implement all aspects of WHO’s policy on conflicts of interest for expert groups (18) and be familiar with Chapter 6 of the *WHO handbook for guideline development* (2).

### 14.5.2.5 Match time available to the magnitude of the task

The time and resources available for developing recommendations need to be sufficient to develop trustworthy guidelines. If this is not the case, it may be wise to decrease the scope of the guideline and recommendations to be discussed at a meeting, or to prolong the meeting or timeline.

### 14.6 Conclusions

WHO has frequently issued strong recommendations based on low-quality evidence. In essence, this means that WHO is advising Member States to implement, in most situations, an intervention with an uncertain balance of benefits and harms. Although in some cases it is entirely reasonable to formulate strong recommendations based on low-quality evidence, doing so is rarely appropriate. WHO staff who develop guidelines must be aware of this and help guideline development groups to formulate recommendations whose strength is consistent with the level of uncertainty surrounding the underlying evidence and are thus more consistent with GRADE.

### 14.7 Acknowledgements

Gordon Guyatt and Paul Alexander (McMaster University, Canada) prepared the report upon which this chapter is based. Susan L. Norris revised and finalized the text. Maria Luisa Clark copy-edited the document, Myriam Felber provided technical support and Martin Mendelson proofread.

### 14.8 References


