ELSEVIER

Journal of Clinical Epidemiology ■ (2013) ■

Journal of Clinical Epidemiology

BRIEF REPORT

World Health Organization recommendations are often strong based on low confidence in effect estimates

Paul E. Alexander[a,*], Lisa Bero[b], Victor M. Montori[c,d,e,f], Juan Pablo Brito[g], Rebecca Stoltzfus[h,i], Benjamin Djulbegovic[j], Ignacio Neumann[a,k], Supriya Rave[l], Gordon Guyatt[m,*]

[a]*Health Research Methods (HRM), Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[b]*University of California, San Francisco, Suite 420, Box 0613, 3333 California Street, San Francisco, CA 94118, USA*
[c]*Healthcare Delivery Research Program, Mayo Clinic, Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA*
[d]*Knowledge and Evaluation Research Unit, Mayo Clinic, Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA*
[e]*Division of Endocrinology, Diabetes, Metabolism, and Nutrition, Mayo Clinic, Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA*
[f]*Division of Health Care and Policy Research, Mayo Clinic, Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA*
[g]*Mayo Clinic, Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA*
[h]*Global Health Program, Division of Nutritional Sciences, 120 Savage Hall, Cornell University, Ithaca, NY 14853, USA*
[i]*Program in International Nutrition, Division of Nutritional Sciences, 120 Savage Hall, Cornell University, Ithaca, NY 14853, USA*
[j]*H Lee Moffitt Cancer Center, University of South Florida, Tampa, FL 33612, USA*
[k]*Department of Internal Medicine, School of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile*
[l]*Blood, Tissues, Organs and Xenografts Unit, Health Canada, Toronto, Ontario M1P4R7, Canada*
[m]*McMaster University Health Sciences Centre, 1200 Main Street West, Room 2C12, Hamilton, Ontario L8S 4K1, Canada*

Accepted 24 September 2013; Published online xxxx

Abstract

Objectives: Expert guideline panelists are sometimes reluctant to offer weak/conditional/contingent recommendations. Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidance warns against strong recommendations when confidence in effect estimates is low or very low, suggesting that such recommendations may seldom be justified. We aim to characterize the classification of strength of recommendations and confidence in estimates in World Health Organization (WHO) guidelines that used the GRADE approach and graded both strength and confidence (GRADEd).

Study Design and Setting: We reviewed all WHO guidelines (January 2007 to December 2012), identified those that were GRADEd, and, in these, examined the classifications of strong and weak and associated confidence in estimates (high, moderate, low, and very low).

Results: We identified 116 WHO guidelines in which 43 (37%) were GRADEd and had 456 recommendations, of which 289 (63.4%) were strong and 167 (36.6%) were conditional/weak. Of the 289 strong recommendations, 95 (33.0%) were based on evidence warranting low confidence in estimates and 65 (22.5%) on evidence warranting very low confidence in estimates (55.5% strong recommendations overall based on low or very low confidence in estimates).

Conclusion: Strong recommendations based on low or very low confidence estimates are very frequently made in WHO guidelines. Further study to determine the reasons for such high uncertainty recommendations is warranted.

Keywords: GRADE; Strength of recommendation; Confidence in effect estimates; Public health guidelines; Clinical practice guidelines; High uncertainty; World Health Organization

Conflict of interest: None declared. There has been no financial support or otherwise provided for the conduct of this study. All authors have contributed to the conduct of this study. All authors are connected in some manner to guidelines in terms of their area of research, review for WHO or another organization, or development. G.G. functioned as a developer of the GRADE methods. R.S. and L.B. functioned as a guideline developers for WHO in the past. S.R. (BSc in Microbiology and Human Biology from University of Toronto and MSc in EBHC from University of Oxford) has contributed to the double screening of the guidelines and recommendations as to strength and confidence classification. Dr. Susan Norris is a current employee of the WHO within the guideline development and review committee department and while not an author, did appreciably contribute to the initial development of the project and the provision of guideline information.

* Corresponding author. Tel.: 905-525-9140; fax: 905-524-3841.
E-mail address: pauleliasalexander@gmail.com (P.E. Alexander) or guyatt@mcmaster.ca (G. Guyatt).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.09.020

What is new?

Key findings

- Over 50% of WHO recommendations are strong and over 50% of those strong recommendations are based on low or very low confidence in effect estimates (study quality).
- Future study on why WHO panelists make such elevated numbers of high uncertainty strong recommendations is required, particularly an examination of factors that panelists may consider in making judgements as they arrive at such recommendations.

What this adds to what is known

- This is the second systematic study documenting that an organization using the GRADE approach to creating guidelines - in this case public health guidelines - has a proclivity to make strong recommendations based on effect estimates warranting low or very low confidence.

What is the implication and what should change now?

- It is possible that these recommendations represent the questionable application of GRADE which may inappropriately limit the discretion of clinicians and public health decision-makers. This may not, however, be the case. Further study of the reasons for this many strong recommendations based on low quality evidence is necessary.

1. Introduction

Clinical practice guidelines (CPG) and public health guidelines (PHG) are statements that are developed in a systematic manner intended to guide clinicians, patients, populations, and policy makers in making the most suitable decisions regarding health management. CPG focuses on individuals and PHG on populations. To produce credible recommendations, CPG or PHG must follow rigorous quality standards in their development [1]. Such standards [1] include use of an evidence-based approach with rating of the confidence in effect estimates (also known as quality of evidence). To encourage appropriate utilization of evidence in public health decision making, the World Health Organization (WHO) develops evidence-informed PHG using procedures outlined in the WHO handbook for guideline development [2]. The guideline development process at WHO involves strong support and guidance from the Guideline Review Committee (GRC) Secretariat who are also involved in the final approval of the guidelines [2].

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) [3,4] approach to guideline development is becoming widely used in WHO guidelines [5,6]. GRADE provides guidance in standardization of guideline development including rating confidence in effect estimates and moving from evidence to recommendations. The GRADE approach categorizes confidence in effect estimates as high, moderate, low, and very low. Randomized controlled trials (RCTs) start as high confidence and observational cohort studies as low confidence. High risk of bias, imprecision, indirectness, inconsistency of results, and likelihood of publication bias can lower confidence in effect estimates. Confidence can increase if effect estimates suggest large intervention effects or if there is evidence of a dose−response gradient. The GRADE approach also provides a framework to move from evidence to the recommendation, suggesting two categories of recommendations: strong and weak (the latter also labeled as conditional, discretionary, or contingent). The strength of recommendations depends on estimates of magnitude of effect, estimates of values and preference and their variability, confidence in each of these estimates, and resource use considerations [3,4]. In the context of public health decision making (and thus PHG development which is the focus of this article), other relevant factors include the burden of illness, accessibility, feasibility, the extent of current suboptimal practice, and the impact on health inequities.

Despite GRADE guidance warning against strong recommendations in the face of low or very low confidence in estimates for critical outcomes, such recommendations may not be uncommon [7]. A preliminary scoping exercise of WHO guidelines conducted in the fall of 2012 showed that approximately one-half of the recommendations were strong based on low or very low confidence (this initial exercise serving as a strong impetus for the present study). Given that strong recommendations are courses of actions recommended to all or almost all patients and circumstances, making strong recommendations in the face of low confidence in benefits or harms may be problematic.

If WHO guideline developers often make strong recommendations in the face of low or very confidence in estimates, then it raises concerns about whether GRADE is being optimally applied in the WHO guideline development process. Furthermore, it suggests that WHO guidelines are not optimally evidence based, do not give public health practitioners the optimal degree of discretion in their decision making, may entrench practices that ultimately prove harmful, and may inhibit needed research.

In an effort to help inform and improve the PHG development process at WHO, our objective was to examine all WHO guidelines developed with the GRADE approach and describe the classifications of strong and weak recommendations and their associated confidence in effect estimates (high, moderate, low, and very low).

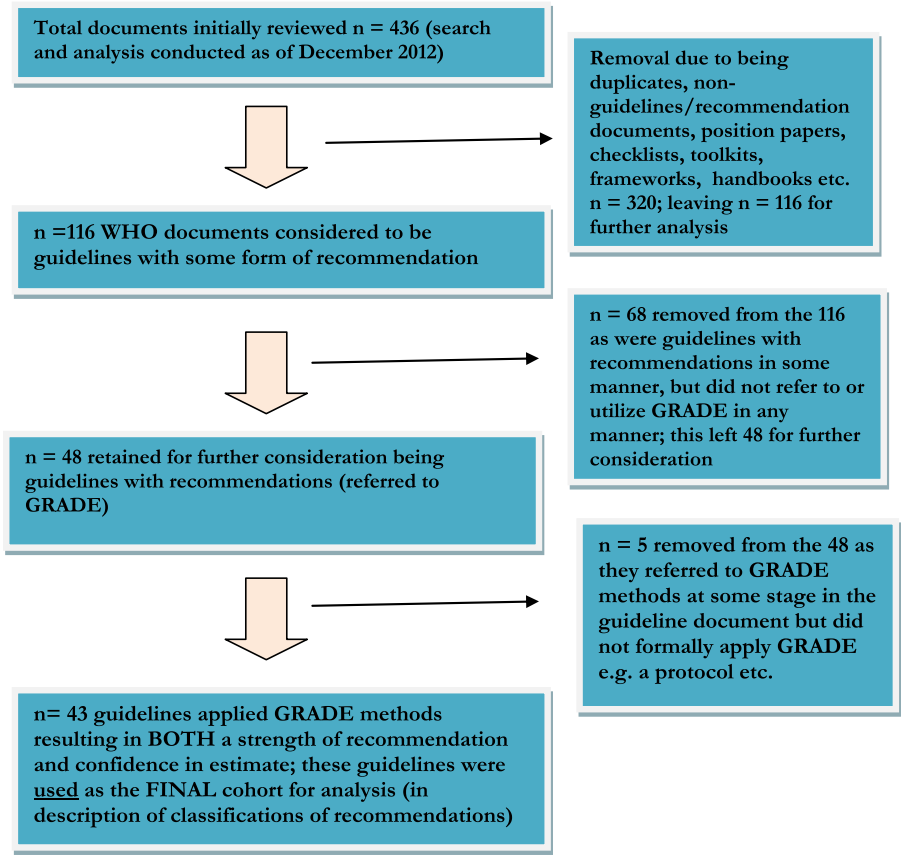

Fig. 1. Flow Diagram of WHO Guidelines (documents) used for the WHO GRADE Guidelines Project (strong recommendations based on low/very low confidence in estimates). GRADE, Grading of Recommendations Assessment, Development and Evaluation.

2. Methods

2.1. Data source

In December 2012, we retrieved all the available WHO PHG. This included WHO guidelines that are available on the main WHO Web site [5] (under "WHO Guidelines: a selection of evidence-based guidelines"; http://www.who.int/publications/guidelines/en/index.html) and those retrieved from an internal WHO database, which includes all final guidelines approved by the WHO GRC that covered 2007 to 2012. For this study, a guideline was defined as a document produced by WHO and available publicly on their Web site or as part of the GRC-retrieved data set file and which culminated in a recommendation(s) or guidance. WHO guidelines eligible for this study applied GRADE methods and included both a rating of confidence in effect estimates and a grading of strength of recommendations. Documents examined for eligibility had a date of publication from January 2007 to December 2012. For our exercise, we adopted the nine guideline categorizations as delineated by WHO [child health, chronic diseases, injuries and disability, environmental health, HIV and AIDS, maternal and reproductive health, mental health and substance abuse, nutrition, patient safety, and tuberculosis (TB)].

2.2. Data abstraction

For each eligible guideline, two reviewers (P.E.A. and S.R.; independently and in duplicate) abstracted the recommendations and noted the confidence in estimates/quality of evidence (high, moderate, low, and very low) and strength of recommendation (strong and weak/conditional). Reviewers

Table 1. Confidence in estimates by strength of recommendation (as of December 2012)

Strong recommendations (n = 289)	n (%)	Weak recommendations (n = 167)	n (%)	Total (%)
High confidence in estimates	50 (17.3)	High confidence in estimates	4 (2.4)	54 (11.8)
Moderate confidence in estimates	79 (27.3)	Moderate confidence in estimates	21 (12.6)	100 (22.0)
Low confidence in estimates	95 (33.0)	Low confidence in estimates	63 (37.7)	158 (34.6)
Very low confidence in estimates	65 (22.5)	Very low confidence in estimates	79 (47.3)	144 (31.6)
Total (%)	289 (100)	Total (%)	167 (100)	456 (100)

Table 3. Paradigmatic situations in which a strong recommendation may be warranted despite low or very low confidence in effect estimates

Situation	Condition	Example
1	When low-quality evidence suggests benefit in a life-threatening situation (evidence regarding harms can be low or high)	Fresh frozen plasma or vitamin K in a patient receiving warfarin with elevated INR and an intracranial bleed. Only low-quality evidence supports the benefits of limiting the extent of the bleeding.
2	When low-quality evidence suggests benefit and high-quality evidence suggests harm or a very high cost	Head-to-toe CT/MRI screening for cancer. Low-quality evidence of benefit of early detection but high-quality evidence of possible harm and/or high cost (strong recommendation against this strategy)
3	When low-quality evidence suggests equivalence of two alternatives but high-quality evidence of less harm for one of the competing alternatives	*Helicobacter pylori* eradication in patients with early stage gastric MALT lymphoma with *H pylori* positive. Low-quality evidence suggests that initial *H pylori* eradication results in similar rates of complete response compared with the alternatives of radiation therapy or gastrectomy; high-quality evidence suggests less harm/morbidity.
4	When high-quality evidence suggests equivalence of two alternatives and low-quality evidence suggests harm in one alternative	Hypertension in women planning conception and in pregnancy. Strong recommendations for labetalol and nifedipine and strong recommendations against angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARB); all agents have high-quality evidence of equivalent beneficial outcomes, with low-quality evidence for greater adverse effects with ACE inhibitors and ARBs
5	When high-quality evidence suggests modest benefits and low/very low quality evidence suggests possibility of catastrophic harm	Testosterone in males with or at risk of prostate cancer. High-quality evidence for moderate benefits of testosterone treatment in men with symptomatic androgen deficiency to improve bone mineral density and muscle strength. Low-quality evidence for harm in patients with or at risk of prostate cancer

Abbreviations: INR, international normalized ratio; CT, computed tomography; MRI, magnetic resonance imaging; MALT, mucosa-associated lymphoid tissue.

also noted whether each recommendation was *for* or *against* an action. WHO generally uses the term "conditional" but, in some instances, uses the term "weak" to describe recommendations that are not strong. Recommendations were further classified by topics designated by WHO: maternal and reproductive health, child health, HIV/AIDS, TB related, chronic diseases, injuries and disabilities, environmental health, patient safety, and nutrition guidelines, and so forth. All abstracted guideline data were entered into a master MS Excel spreadsheet for management and cleaning (assessment of eligibility, removal of duplicates, etc.) as well as during the duplicate data abstraction.

To resolve differences following abstraction, both reviewers consulted and referred to their individual data abstraction results and identified those guidelines and classifications for which there were discrepancies. Each discrepancy was discussed in terms of the strength and confidence in estimates. Reviewers reexamined the classifications,

Table 2. WHO GRADE guideline recommendations that are strong and based on low or very low confidence in estimates by subcategories of all strong recommendations

WHO guideline subcategory topic	Number strong low/very low/total	% Strong low/very low
WHO six of nine guideline subcategories examination period 2007–2012		
Maternal and reproductive health[a]	28/54	52
Child health[a]	49/93	53
HIV/AIDS[a]	32/66	49
TB[a]	24/42	57
Nutrition	2/2	100
Mental health and substance abuse	6/8	75
WHO additional guidelines with no clear subcategory topic area examination period 2007–2012		
Pandemic (H1N1) 2009 influenza and other influenza viruses	11/11	100
Malaria	1/5	20
Increasing access to health workers in remote and rural areas	7/8	87

Abbreviations: WHO, World Health Organization; GRADE, Grading of Recommendations Assessment, Development and Evaluation; TB, tuberculosis.

The three guideline topic areas, (1) chronic diseases, injuries, and disabilities; (2) patient safety; and (3) environmental health, reported no GRADEd strong recommendations with low or very low confidence.

[a] These four subcategories account for 88% of all GRADEd strong recommendations.

background information, and information contained in the annexes (GRADE tables). Third-party adjudication was not needed as consensus was reached given the clarity of the data. Kappa agreement score was computed. MS Excel was used for basic descriptive analysis of data classifications (frequencies and percentages).

3. Results

We reviewed all 436 WHO documents (69 from the public WHO Web site and 367 from the internal GRC data set) for eligibility, of which 320 (73.3%) proved either duplicates or not guidelines, including position articles, checklists, tool kits, frameworks, and handbooks. The remaining 116 (26.6%) documents were guidelines with some form of recommendation. Of these 116 documents, 68 (58.6%) proved ineligible because they did not use GRADE methods. Another five were ineligible because although they used GRADE methods, they did not provide both a rating of confidence in estimates and a grading of strength of recommendation. Thus, 43 of the initial 436 WHO documents (9.8%; 37% of the 116 guideline documents) were included in the final analysis (Fig. 1, flow diagram).

The 43 guidelines included 456 recommendations, of which 289 (63.4%) were strong recommendations and 167 (36.6%) were weak (Table 1). Of the 289 strong recommendations, 160 (55.5%) were based on low and very low confidence in estimates. Of the 289 strong recommendations, 95 (33.0%) were based on low confidence in effect estimates and 65 (22.5%) were based on very low confidence in effect estimates (Table 1). Of the 167 weak recommendations, 63 (37.7%) were based on low confidence in effect estimates and 79 (47.3%) were based on very low confidence in effect estimates (Table 1). The kappa for agreement was 0.81. There were 258 strong recommendations *for* the intervention (89.3%) and 31 *against* (10.7%). Of the 167 weak recommendations, 142 (85.0%) were *for* the intervention and 25 (15.0%) were *against*. When focusing only on strong recommendations that had at least one low or very low confidence rating, 91.0% of these were for the intervention or action being recommended.

The classification of strong and weak recommendations, and ratings of confidence, differed across categories. Maternal and reproductive health, child health, HIV/AIDS, and TB guidelines report $>50\%$ of their recommendations as being strong, of these $>50\%$ were based on low or very low confidence in estimates (Table 2). Pandemic influenza and nutrition guideline recommendations were strong and based on low or very low confidence in estimates (Table 2), but it should be noted that this was based on a small number of recommendations. In contrast, guidelines for chronic diseases, injuries and disabilities, patient safety, and environmental health did not report any GRADEd strong recommendations based on low or very low confidence in estimates (Table 2).

4. Discussion

The predominant finding was that WHO guideline panels often make strong recommendations (63%); more than half (55%) of these strong recommendations are based on low or very low confidence in effect estimates (Table 1). Strong recommendations based on low or very low confidence were particularly frequent in certain content areas including maternal and reproductive health, child health, HIV/AIDS, and TB (Table 2). These findings raise questions as to whether GRADE is being applied appropriately and the extent to which WHO panelists neglect uncertainties in the evidence when they consider strength of recommendations.

When guideline panelists make strong recommendations, they are suggesting that frontline decision makers need not consider the issue any further. Public health officials who view WHO guidelines as authoritative may feel that, in the face of such recommendations, they should put aside concern that the recommendation may not be optimal in their setting. If they do respond in this way, these strong recommendations may be inappropriately restricting the discretion of public health decision makers.

There may be circumstances in which a strong recommendation is warranted despite low or very low confidence in effect estimates. Indeed, the GRADE working group has identified five paradigmatic situations in which this may be the case (Table 3) [8]. Further study of WHO GRADEd recommendations (and other organizations, institutions, or relevant entities that produce CPG and PHG and recommendations using GRADE) to determine the extent to which strong recommendations based on low or very low confidence estimates meet these conditions would be helpful.

Additional work on other determinants of strong recommendations that might be enlightening includes the role of intellectual or financial conflicts of interest (balancing the need to use expert input into guideline development while mitigating the impact of intellectual and financial conflicts) [9] and panel composition. Additional inquiry into the appropriateness of WHO strong recommendations based on low or very low confidence estimates and the reasons why guideline panelists are making these recommendations (whether appropriate or inappropriate) are warranted.

Acknowledgments

The authors acknowledge Susan L. Norris for her important contributions to this work, including conceptualization, provision of data, and comments on the draft manuscript.

References

[1] Institute of Medicine. Of the national academies. Available at http://www.iom.edu/?ID=68004. Accessed April 25, 2013.

[2] WHO handbook for guideline development. 2012. Available at http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf. Accessed April 25, 2013.

[3] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94. Available at http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract. Accessed March 10, 2013.

[4] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6. http://dx.doi.org/10.1016/j.jclinepi.2010.07.015.

[5] WHO. WHO guidelines approved by the Guidelines Review Committee. Available at http://www.who.int/publications/guidelines/en/index.html. Accessed April 24, 2013.

[6] WHO. Prevention and control of NCDs: guidelines for primary health care in low-resource settings. 2013. Available at http://www.who.int/nmh/publications/phc2012/en/index.html. Accessed August 20, 2013.

[7] Brito JP, Domecq JP, Murad MH, Guyatt GH, Montori VM. The endocrine society guidelines: when the confidence cart goes before the evidence horse. J Clin Endocrinol Metab 2013;98:3246−52. http://dx.doi.org/10.1210/jc.2013-1814.

[8] Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines 15: going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35. pii: S0895-4356(13)00054-1 http://dx.doi.org/10.1016/j.jclinepi.2013.02.003.

[9] Guyatt G, Akl EA, Hirsh J, Kearon C, Crowther M, Gutterman D, et al. The vexing problem of guidelines and conflict of interest: a potential solution. Ann Intern Med 2010;152:738−41. http://dx.doi.org/10.1059/0003-4819-152-11-201006010-00254.