CLINICAL PRACTICE AND PUBLIC HEALTH GUIDELINES

CLINICAL PRACTICE AND PUBLIC HEALTH GUIDELINES: THE MAKING OF APPROPRIATE STRONG RECOMMENDATIONS WHEN THE CONFIDENCE IN EFFECT ESTIMATES IS LOW OR VERY LOW (DISCORDANT)

By PAUL ELIAS ALEXANDER, MHSc, M.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2015)

Hamilton, Ontario (Health Research Methodology)

TITLE: Clinical practice and public health guidelines: The making of appropriate strong recommendations when the confidence in effect estimates is low

AUTHOR: Paul Elias Alexander, MHSc. (University of Toronto) M.Sc.

(University of Oxford (Oxon))

SUPERVISOR: Distinguished Professor Gordon Guyatt

NUMBER OF PAGES: xv, 216

Abstract

Clinical practice, public health, and policy guidelines should be developed based on a systematic approach that uses the best available evidence. The advent of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework has facilitated this, resulting in a transparent approach to guideline development.

GRADE suggests that guideline developers seldom make strong recommendations based on low or very low confidence in effect estimates (strong l/vl).

The World Health Organization (WHO) produces recommendations that guide public health policy and, in 2003, WHO adopted the GRADE approach to guideline development. Initial anecdotal evidence suggested that WHO issues a large number of strong recommendations and particularly strong l/vl.

Our research team evaluated the nature of WHO recommendations and conducted a qualitative study using interviews of guideline panel members. Key findings included: i) WHO makes a large proportion of recommendations as strong l/vl ii) many strong l/vl are inconsistent with GRADE guidance iii) reasons guideline panel members offered for strong l/vl included skepticism about the value of making conditional recommendations; political considerations; a high confidence in benefits despite formal

ratings of low confidence; and long-standing practices, funding, and policy; iv) methodologist interviewees indicated panelists' lack of commitment to conditional recommendations; a perceived tension between methodologists and panelists due to resistance to adhering to GRADE guidance; both financial and non-financial conflicts of interest among panel members as explanations of strong l/vl; and the need for greater clarity of, and support for, the role of methodologists as co-chairs of panels.

The understanding of when and why strong l/vl are formulated at WHO is an important methodological issue that has implications not just for WHO, but for a wide range of guideline developers elsewhere. Our findings offer insights that may guide interventions to enhance trustworthiness of practice guidelines.

Preface

This thesis has been conducted as a "sandwich thesis" and consists of five individual manuscripts/papers at various stages of publication (and associated protocols). These are:

1.) Chapter 1: Introduction of the thesis
2.) Chapter 2: Protocol for Phases I and II
3.) Chapter 3: Phase I study and results; manuscript
4.) Chapter 4: Phase II study and results; manuscript
5.) Chapter 5: Protocol for Phase III
6.) Chapter 6: Phase III study and results (WHO panel interview study); manuscript
7.) Chapter 7: Phase IV study and results (WHO methodologist interview study); manuscript
8.) Chapter 8: Recommendations guidance document produced for WHO; manuscript (accepted by WHO)
9.) Chapter 9: Discussion and conclusion/summary

At the time of writing (December 2014-January 2015) two of the five individual manuscripts (chapters three and four) have been accepted and published in peer reviewed journals, two of the remaining three (chapters six and seven) have been written and submitted for publication. The 5th manuscript (chapter eight, a WHO Guidance document) has been written and accepted by the World Health Organization as part of their publication process for February 2015 (Geneva, Switzerland). Our plan, with some modifications and permission from WHO, is to also publish the Chapter eight.

Mr. Alexander was the principal contributor to the conception and design of the study, with the aid of his supervisor Dr. Guyatt but with definite lead by Paul Alexander. Such a lead entailed developing the research questions, the writing of the respective protocols, designing all screening and data abstraction tools, performing all study tasks e.g. screenings, performing the data abstraction, performing the qualitative study interviews, conducting all descriptive, taxonomy (with three reviewers), and analytic steps, writing up of drafts of all manuscripts, submitting the manuscripts to his supervisor for guidance, working with various academic and research content experts for feedback and guidance, working with WHO for guidance and collaboration, and final submission of manuscripts for publication. This also involved editing the submitted manuscripts to allow for full acceptance and publications. This work is principally the undertaking of Mr. Alexander, with guidance from Dr. Guyatt and collaborative input from the larger research team. All studies were initiated in 2012 and research was conducted in 2012, 2013, and 2014.

Copyright permission:

Acknowledgments:

I wish to provide my deepest appreciation to my supervisor Dr. Gordon Guyatt for his never ending support and guidance of me. Working under Dr. Guyatt has been life altering for me for he has taken my very rough edges and helped polish me, improve me, and has pushed me to work harder and to learn and develop further. There are no words to express how I feel toward Dr. Guyatt as a mentor and the massive admiration I have for him as an advocate given I was able to witness and be part of his steep commitment to all students at McMaster.

I must also thank Dr. Holger Schünemann and Dr. Lehana Thabane for their willingness to be part of my thesis committee and to be a knowledge resource and support avenue along the way. My gratitude to all three named persons. Finally, there really are no words suitable enough to express what McMaster has meant to me and my family and I will always be indebted to these persons and this fine institution. I also thank all graduate office and support staff who work tirelessly behind the scenes to make our lives as students easier and to help facilitate our success. I/we owe a substantial portion of our success to them. Thank you all.

Table of contents

Lists of Figures and Tables

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

Chapter 8:

List of Abbreviations

CPG: Clinical practice guidelines

Discordant: Strong recommendation based on low or very low confidence

GRADE: Grading of Recommendations Assessment, Development and Evaluation

GRADEd: Recommendations produced via use of the GRADE methods

GRC: Guideline review committee

HIV: Human Immunodeficiency virus

IPT: Isoniazid preventive therapy

MA: Meta-analysis (or meta-analyses)

PHG: Public health guidelines

PICO: Patients/population, intervention, comparator, outcome

RCT: Randomized controlled trial

SR: Systematic review

Strong 1/vl: Strong recommendations based on low or very low confidence

SVS: Society for Vascular Surgery

TB: Tuberculosis

TES: The Endocrine Society Guidelines Study

V/Ps: Values and preferences of patients or populations

WHO: World Health Organization

Declaration of Academic Achievement

Mr. Alexander was the main contributor and first author of all the manuscripts/documents contained in this thesis. The details of the contributions of coauthors are described at the end of each manuscript.

Chapter 1: Introduction

Clinical practice guidelines (CPG) and public health guidelines (PHG) are statements that are developed in a systematic manner in order to guide clinicians, patients, and policy makers in making the most suitable decisions regarding health management. CPG focuses on individuals with particular treatments and care for particular illnesses. PHG provide guidance on the ways of helping populations improve their health and reduce the risk of illness. In offering guidance, CPG or PHG must follow rigorous quality standards in their development to produce credible evidence-informed recommendations.

The World Health Organization (WHO) is a global health leader that develops PHG for all nations. The guidance that WHO produces, particularly important to lower income nations, is geared toward public healthcare practitioners, policy developers, and consumers.[1] WHO strives to develop PHG that are of the highest methodological quality and supported by systematic reviews of the underlying evidence. The aim by WHO is to adopt standards using a transparent, systematic, and evidence-based decision making process that allows for an in-depth analysis of the desirable and undesirable outcomes of healthcare options.[2-4] However, such high level guideline development quality remains a challenge for those responsible for ensuring trustworthy guidelines, as articulated by the Institute of Medicine[4] in their 2011 standards.

The guideline development process at WHO received a scathing critique in 2007[5] that prompted WHO to initiate the Guidelines Review Committee (GRC) Secretariat. The

critique included a failure by WHO guideline panels to use a systematic evidence-informed approach to guideline development.[5,6] As part of the critique, authors called for a more acute focus on the best evidence and appropriate utilization of such evidence in public health decision making.[5] WHO has responded with standardized guideline procedures outlined in the WHO handbook for guideline development.[7] A more recent 2013 review of WHO guidelines found that while there was room for improvement, guideline quality had improved markedly since the GRC was initiated in 2007.[6]

A central feature of the guideline development process at WHO has been the adoption and utilization of a more structured, explicit, systematic, and transparent framework to evaluate and summarize the evidence, and move from evidence to recommendation. Such an evaluation and transition from evidence to recommendation allows for trustworthy and reliable recommendations built on standardized ratings of the quality of underlying evidence and the grading of their strength. This is known as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach or framework.[8-12] GRADE was adopted by WHO in 2003,[11] becoming more mainstreamed at roughly the same time that the GRC was initiated in 2007. A core mandate of the GRC was guideline development quality assurance improvement (improved standards) along with the institutional use of GRADE. GRADE is becoming widely used across WHO in guideline development (a list of over 80 organizations using GRADE can be found at www.gradeworkinggroup.org).[13-17]

Following the GRADE approach,[8-12, 18,19] best estimates of intervention effects come from systematic reviews of randomized controlled trials (RCT) of the impact of alternative treatment approaches. Factors including risk of bias (consideration of randomization to treatment arms, allocation concealment of the randomization sequence, blinding of study participants, personnel, lead researchers, outcome assessors, baseline similarity of treatment groups, losses to follow-up/attrition, incomplete outcome reporting, selective outcome reporting etc.), imprecision (95% confidence interval; number of events; size of sample), indirectness (relating the uncovered evidence to the particular research question e.g. patients/population, interventions, comparators, outcomes), inconsistency of results (heterogeneity of effect estimates), and publication bias (small sample size, negative, or non-significant studies not being published) determine confidence in estimates of effect (rated as high, moderate, low and very low).

GRADE then provides guidance in using the evidence to determine a direction and strength of the recommendation and suggests a dichotomy for strength: a guideline panel may be confident that desirable consequences do or do not outweigh undesirable consequences (a strong recommendation), or that the balance of desirable and undesirable consequences is less certain (resulting in a weak, conditional, discretionary, qualified, or contingent recommendation).

Using the GRADE approach in clinical practice guidelines (CPG), determinants of the strength of recommendations include confidence in estimates of treatment effects

(also known as study quality), magnitude of the desirable and undesirable consequences of alternative courses of action, value and preference (V/Ps) judgements required in trading off desirable and undesirable consequences, uncertainty regarding patients' V/Ps, variability in these V/Ps, and resource use considerations. In the context of public health decision-making (and thus PHG development), determinants of the strength is the same as for CPG, but other factors such as (but not an exhaustive list) the burden of illness, accessibility, feasibility, acceptability, social context, the extent of current suboptimal practice, and the impact on health inequities, may also require consideration.

A strong impression and initial anecdotal evidence and for a variety of reasons not yet fully understood, indicated that WHO was producing a large volume of their guideline recommendations and across all WHO health topic areas, as strong recommendations and based on high uncertainty (low or very low study quality or confidence in effect estimates). Such a large proportion of strong recommendations based on low or very low confidence in effect estimates (strong l/vl) is challenging and a concern as strong recommendations may be questionable in the presence of low quality evidence that implies uncertainty regarding effects of the recommended course of action. The issue of strong l/vl being made by WHO represents the central thrust or theme of this thesis.

GRADE guidance warns against strong l/vl and suggests that such recommendations be made sparingly and with precaution. This is because strong recommendations are "just do it", readily adopted, and unquestioned courses of actions

recommended to all or almost all guideline users and circumstances, and thus these recommendations must generally be undergirded by the high quality evidence. There may, however, be circumstances in which a strong l/vl is warranted. Indeed through many years of research and practical experience, GRADE has identified five paradigmatic situations in which a strong l/vl is warranted (Table 1).[10] These paradigmatic situations are a life threatening situation, uncertain benefit/certain harm, potential equivalence with one option clearly less risky or costly, high confidence in similar benefits where one option is potentially more risky or costly, and potential catastrophic harm.

As such, if our study could definitively show that WHO is making a large proportion of their recommendations as strong l/vl, then, as it lays the framework for deeper study, it raises concerns about whether GRADE is being optimally applied in the WHO guideline development process. More important, it suggests the possibility 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. When strong recommendations are made that should have been weak, public health practitioners may feel constrained to, and against their better judgement, follow the strong recommendation while if the recommendation were weak in the first place, this would have provided the appropriate flexibility. The result could be global practitioners adopting recommended actions that could be detrimental to the patients or populations.

Table 1: Paradigmatic situations in which panels may reasonably offer (optimally made) strong recommendations on the basis of low or very low confidence in effect estimate.

Paradigmatic situation	Confidence in effect-estimates for health outcomes (Quality of evidence)		Balance of benefits and harms	Values and Preferences	Resource considerations	Recommendation
	Benefits	Harms				
Life threatening situation	Low or very low confidence	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation. Adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits	Strong recommendation in favour
Uncertain benefit, certain harm	Low or very low	High or Moderate	Possible but uncertain benefit. Substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	Possible high incremental cost (or resource use) may further mandate a recommendation against the intervention	Strong recommendation against (or in favor of the less harmful/less expensive alternative when two are compared)
Potential equivalence, one option clearly less risky or costly	Low or very low	High or Moderate	Magnitude of benefit apparently similar - though uncertain - for alternatives. We are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits, may further support recommendation for less harmful alternative	Strong recommendation for less harmful/less expensive
High confidence in similar benefits, one option potentially more risky or costly	High or Moderate	Low or very low	Established that magnitude of benefit similar for alternative management strategies. Best (though uncertain) estimate is that one alternative has appreciably greater harm.	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) of one alternative	Strong recommendation against the intervention with possible greater harm or cost
Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) of potentially harmful intervention may further justify recommendation for less harmful.	Strong recommendation against the intervention (or in favor of the less harmful/less expensive alternative when two are compared)

The concern therefore was, and based on our initial anecdotal impression (and an informal snap-shot of WHO guidelines), whether WHO guideline panelists are overly keen to offer strong recommendations when their confidence in effect estimates are low. Are WHO guideline panels making strong recommendations irrespective of the underlying confidence in effect estimates? Questions also emerge regarding whether GRADE is being optimally applied in the WHO guideline development process. This could indicate that WHO guidelines are not optimally evidence-based. On the other hand while these concerning questions arose, we were cognizant to the fact that WHO panelists may be entirely reasonable in their judgements regarding strong recommendations, but are neglecting to sufficiently make their rationale explicit for the guideline user.

Our initial strong anecdotal impression and cursory examinations of WHO guidelines in 2012 was given credence by the Endocrine Society Guidelines (TES)[20] Study that found that 58% of TES recommendations were strong, and of those 59% were based on low confidence. A further examination of the 59% of strong l/vl revealed that only 29% could be judged as reasonable, defensible, and consistent with one of the five paradigmatic situations that warrant a strong l/vl.[10]

Similarly, other researchers conducted a study on the American Association of Blood Banking clinical practice guidelines (CPG) for the use of fresh-frozen plasma (FFP) by analyzing the relationship between confidence in effect estimates and the strength of recommendations.[21] Researchers found that when the underlying confidence is higher, the probability of making a strong recommendation for or against an intervention increases

considerably. They also found that 64% of the examined strong recommendations were based on low or very low confidence.[21]

In an examination of clinical practice guidelines produced by the Society for Vascular Surgery (SVS),[22] researchers found that 65% of recommendations were also strong and supported by low or very confidence respectively.[22] SVS researchers made value judgements that superseded the underlying evidence. For example, they described one strong l/vl against carotid artery stenting in asymptomatic patients with carotid stenosis. Improved data exists today but at the time of the guideline's development, there were no randomized trials comparing stenting to medical management. Also, the comparative evidence of stenting to endarterectomy was thin and imprecise. Panelists therefore placed a relatively high value on avoiding the possible downsides (undesirable outcomes) of an invasive procedure for a low-risk patient, and despite the availability of low-quality evidence, the panel issued a strong recommendation.

Such large numbers of strong l/vl may also be driven by concerns that a weak recommendation, while also a credible recommendation, may not be understood, taken seriously, or followed by users of the guidelines. These types of reasons may be a concern not just for WHO but for guideline developers elsewhere and suggested that while dedicated to WHO's making of strong l/vl, this thesis study could be informative and instructive for guideline development globally.

To address this tendency to make strong l/vl at WHO and whether they are being made consistent or inconsistent with GRADE guidance, we initially designed a three part

research project with a fourth phase developed and added on in order to strengthen our findings. This has resulted in five manuscripts (four studies and one guidance report) that have been published or currently are in the publication process. The individual studies have different methodologies while building upon each other, and the reader is cautioned that there is some overlap in the background literature in several papers.

The second chapter outlines the initial study protocol that lays out the conduct of Phases I (descriptive study) and II (taxonomy study). In the third chapter (JCE published Phase I study) of this thesis, "World Health Organization recommendations are often strong based on low confidence in effect estimates",[23] we wanted to first gain a clear understanding of the distribution of recommendations being made by WHO in terms of confidence and strength. We thus provide a descriptive characterization of all GRC approved publicly available WHO guidelines and recommendations made from 2007 to 2012 and which utilized GRADE methods. In doing so, we separate recommendations by guideline topic areas, strength and confidence, and with a focus on delineating strong l/vl. Our Phase I descriptive analysis[23] findings supported initial anecdotal impressions.

Our aim was to access the strong l/vl for a follow-up in-depth examination of consistency or inconsistency of strong l/vl with GRADE guidance.[10] Thus, once we separated out the strong l/vl and by the predominant WHO guideline topic areas, in the fourth chapter or Phase II (in press), "World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance",[24] we expand on the third chapter (Phase I) with a

taxonomy exercise (Phase II) to assess whether the uncovered strong l/vl are consistent with the five identified paradigms in which it is reasonable to make such recommendations.[10] Through actual explicit WHO recommendations emerging from the Phase I project, we provide our taxonomy reviewer judgements in Phase II[24] in order to document the extent to which WHO strong l/vl are being optimally (or sub-optimally) made. Our taxonomy judgements are based on the use of the taxonomy/paradigms for appropriately made strong l/vl shown in Table 1.[10, 20]

The fifth thesis chapter represents the Phase III panel interview study protocol. Following the taxonomy exercise (Phase II[24]) whereby we categorized strong l/vl as either consistent or inconsistent with GRADE guidance for strong l/vl, we then embarked on the Phase III (the results of which are described in the sixth chapter of this thesis), "WHO guideline panelist experience with GRADE methods when making strong recommendations based on low or very low confidence in effect estimates: A qualitative descriptive study". For this qualitative descriptive interview study,[25] we focus only on the strong l/vl recommendations made by WHO panelists that we judged to be inconsistent with GRADE guidance from Phase II.[24] Through interviews with members of the guideline development panels involved in selected guideline recommendations, we sought to gain insight into the process of making strong l/vl. Gathering this information via direct one-on-one interviews with WHO panel members became imperative so that if GRADE was being employed sub-optimally e.g. due to a lack of GRADE guidance or

training etc., then this would require renewed training focus by WHO guideline developers seeking to use GRADE. As well, we were seeking to assess what modifications to GRADE, if any, were needed and felt that interviews with panel members making strong l/vl would take us beyond the written guideline document and allow an understanding of what panelists consider in making strong l/vl especially given the paucity of explicit strong l/vl rationales. WHO panels may have also been correct in their judgements and our Phase II judgements[24] may have been erroneous.

We were unable to recruit methodologists trained in GRADE and who sat on the guideline panels for the selected guidelines as part of the Phase III interview study[25] (Chapter six). Chapter seven of this thesis titled "Experiences of senior GRADE methodologists as part of WHO guideline development panels: an inductive content analysis" sought to rectify this limitation. Chapter seven reflects an additional interview study (that can be regarded as a Phase IV) with senior GRADE methodologists[26] as we felt that GRADE methodologists who worked on WHO guidelines in the past that made strong l/vl, could provide further insight into the process of making strong l/vl. For the Phase III and this Phase IV methodologist interview study, the term "strong l/vl" is used interchangeably with "discordant recommendations"). The methodologists' interview study protocol (in terms of methodological conduct) is the same as used for the panelists' interviews (Phase III, Chapter six) except that the methodologist interviewees were not part of the guidelines used as the basis for interviews in Phase III. We felt that the findings from the methodologists' interviews[26] could substantiate or refute the emergent

findings from the Phase III interview study.[25] We were given strong impetus from the Phase III interview study that methodologists held a critical leadership and guidance role and would therefore be key informants as we sought to expand our understanding of factors/drivers of strong l/vl being made by WHO panels.

With these findings from the four studies (Chapters three (Phase I), four (Phase II), six (Phase III/guideline panel interviews), and seven (senior GRADE methodologist interviews)), we summarize this thesis by offering Chapter eight which is a guidance document requested by WHO for their guideline development handbook, 2015 publication, for when panelists confront the making of strong l/vl. This guidance is written for WHO panels specifically (but potentially guideline development panels elsewhere that use GRADE) titled "Strong recommendations based on low quality evidence (discordant recommendations): guidance for WHO guideline developers".

We have written the Chapter eight guidance based on the findings from all Phases of the WHO GRADE Guidelines Project with an assumption that WHO will continue commitment to GRADE principles and will provide leadership in fostering the guidance. We offer what we think will benefit WHO panels (and guideline developers globally) when seeking to make strong l/vl and we bring the thesis to a close with Chapter nine which is a discussion of what we principally found and what it means, with a final conclusion.

References:

1. World Health Organization (WHO). WHO guidelines approved by the Guidelines Review Committee. Accessed from url: http://www.who.int/publications/guidelines/en/; May 29th 2014.

2. Graham R, Mancher M, Wolman DM, Greenfield S, Steingerg E. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Institute of Medicine. Clinical Practice Guidelines We Can Trust. 1st ed. Washington, DC: National Academies Press; 2011.

3. Institute of Medicine. Of the national academies. url: http://www.iom.edu/?ID=68004 (Accessed on April 25th 2013).

4. Institute of Medicine (IOM). Standards for developing trustworthy clinical practice guidelines. March 2011. url: http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust/Standards.aspx (Accessed on January 6th 2015).

5. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet* 2007; 369: 1883-9.**

6. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. Guideline development at the World Health Organization: an evaluation. ***PLoS One* 2013; 8(5):e63715.**

7. WHO Handbook for Guideline Development. 2012 url: http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf (Accessed on April 25th 2013).

8. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. ***J Clin Epidemiol* 2011; 64(4):380-2.**

9. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.*** **2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

10. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M, Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G.
GRADE guidelines 15: Going from evidence to recommendation-determinants of a recommendation's directionand strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

11. Schünemann HJ, Fretheim A, and Oxman AD. WHO Advisory Committee on Health Research. Improving the use of research evidence in guideline development: 1. Guidelines for Guidelines. ***Health Res Policy Syst.*** **2006; 4:13.**

12. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008; 336(7650):924-6.**

13. WHO. Guidelines for the screening, care and treatment of persons with hepatitis C infection. url: http://apps.who.int/iris/bitstream/10665/111747/1/9789241548755_eng.pdf?ua=1&ua=1 (Accessed on November 5th 2014).

14. WHO. Emergency response to antimalarial drug resistance. url: http://www.who.int/en/ (Accessed on April 24th 2013).

15. WHO. WHO guidelines approved by the Guidelines Review Committee. url: http://www.who.int/publications/guidelines/en/index.html (Accessed on May 20th 2014).

16. GRADE Working Group. url: http://www.gradeworkinggroup.org/ (Accessed on June 17th 2014).

17. WHO. Guidelines for the identification and management of substance use and substance use disorders in pregnancy. ISBN 978 92 4 154873 1. url:

http://apps.who.int/iris/bitstream/10665/107130/1/9789241548731_eng.pdf?ua=1 (Accessed on November 12th 2014).

18. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. ***J Clin Epidemiol.*** **2011 ;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026. Epub 2010 Dec 31.**

19. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y, Schünemann HJ. GRADE guidelines: 2. Framing the question and deciding on important outcomes. ***J Clin Epidemiol.*** **2011; 64(4):395-400. doi: 10.1016/j.jclinepi.2010.09.012. Epub 2010 Dec 30.**

20. 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

21. Djulbegovic B, Trikalinos TA, Roback J, Chen R, Guyatt G. Impact of quality of evidence on the strength of recommendations: an empirical study. ***BMC Health Serv Res.*** **2009; 9:120.**

22. Murad MH, Montori VM, Sidawy AN. Guideline methodology of the Society for Vascular Surgery including the experience with the GRADE framework. ***J Vasc Surg.*** **2011; 53:1375–1380.**

23. Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014; 67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

24. Paul E Alexander, Juan P. Brito, Ignacio Neumann, Michael R. Gionfriddo, Lisa Bero, Benjamin Djulbegovic, Rebecca Stoltzfus, Victor M. Montori, Susan L. Norris, Holger J Schünemann, Gordon H Guyatt. World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. ***J Clin Epidemiol.*** **2014; In Press.**

25. PHASE III panel interview study; JCE has agreed to accept submission.

26. Phase IV methodologist interview study; JCE has agreed to accept submission.

CHAPTER 2: Protocol for Phases I and II

The use of GRADE methods in WHO guidelines: a protocol focus on strong recommendations based on low and very low confidence in estimates

Paul Alexander, Gordon Guyatt, Susan Norris, Victor Montori, Juan Pablo Brito, Lisa Bero, Benjamin Djulbegovic, Ignacio Neumann, Rebecca Stoltzfus

This chapter represents the protocol used to direct the conduct of Phases I and II (Chapters three and four) of the thesis and spells out the steps taken.

Project background

Clinical practice guidelines (CPG) and public health guidelines (PHG) are statements that are developed in a systematic manner in order to guide clinicians, patients, populations and policy makers in making the most suitable decisions regarding health management. While CPG focuses on individuals with particular treatments and care for particular illnesses, PHG provides guidance on the ways of helping populations improve their health and reduce the risk of illness.

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach was developed to standardize guideline development.[1-3] The GRADE working group[1-3] has suggested a structured approach to rating confidence in estimates of effect (quality of evidence) and moving from evidence to recommendations in CPG and PHG. The GRADE approach is increasingly being used and has been widely adopted, including endorsement by over 80 organizations worldwide.[1-3]

Following the GRADE approach, best estimates of intervention effects come from systematic reviews of randomized controlled trials. Factors including risk of bias, precision, and consistency of results determine confidence in estimates (rated high, moderate, low and very low). GRADE then provides guidance in using the evidence to determine a direction and strength of recommendation and suggests a dichotomy for strength: a guideline panel may be confident that desirable consequences do or do not

outweigh undesirable consequences (a strong recommendation), or that the balance of desirable and undesirable consequences is less certain (resulting in a weak, discretionary, or contingent recommendation).

Using the GRADE approach in CPG, determinants of the strength of recommendations include confidence in estimates of treatment effects, magnitude of the desirable and undesirable consequences of alternative courses of action, value and preference judgements required in trading off desirable and undesirable consequences, uncertainty regarding patients' values and preferences (V/P), variability in these V/P, and resource use considerations. In the context of public health decision-making (and thus PHG development), other factors including the burden of illness, accessibility, the extent of current suboptimal practice, and the impact on health inequities, may also require consideration.

What happens when guideline panels use GRADE as part of guideline development?

Questions of interest in how panels use GRADE could include: How often are recommendations strong versus weak/conditional/discretionary? How often are strong versus weak recommendations supported by high, moderate, low and very low quality evidence? To what extent do panels make V/P underlying their guidelines explicit? To what extent do they specify the basis of their estimates of V/P?

One particularly relevant question may be the extent to which panels make strong recommendations on the basis of low or very low confidence in effect estimates. The strong anecdotal impression is that, for a variety of reasons, many guideline panels prefer to make strong recommendations. However, because it is problematic to make a strong recommendation when one is not clear about the balance between desirable and undesirable consequences (which is the case if evidence warrants low or very low confidence in estimates), GRADE discourages such recommendations, and anticipates they will seldom be made.

Through practical experience with guideline development, the GRADE working group[1-3] has suggested a taxonomy of situations (paradigms) in which it may indeed be appropriate to make strong recommendations on the basis of low or very low confidence in effect estimates. The extent to which this taxonomy is comprehensive, particularly in application to PHG, and that it can be reproducibly applied, has not yet been explored.

World Health Organization's use of GRADE in PHG

The World Health Organization (WHO) has been developing PHG to ensure the appropriate utilization of evidence in public health decision making.[4] WHO departments develop evidence-informed recommendations that are geared toward populations using procedures outlined in the WHO handbook for guideline development.[5] The steps in the WHO guideline development process include: (i) identification of priority questions and outcomes; (ii) retrieval of the evidence; (iii) assessment and synthesis of the evidence; (iv)

formulation of recommendations, including research priorities; and (v) planning for dissemination, implementation, impact evaluation and updating of the guideline. The GRADE methodology[1-3] is increasingly being used by WHO to prepare evidence profiles based on up-to date systematic reviews (SRs).

Commencing in July 2012, we conducted a preliminary examination of the WHO PHG. This included WHO guidelines that are available on the main WHO website under 'Guidelines' as well as those retrieved from an internal WHO database which includes all final guidelines approved by the WHO Guideline Review Committee (GRC). Initial results suggested that 35-40% of WHO guidelines offer strong recommendations based on low and/or very low confidence in estimates. While preliminary, these results raise concerns about whether GRADE is being optimally applied in WHO guideline development. Are the WHO guideline panelists overly keen to offer strong recommendations when their confidence in effect estimates are low? Can differences in guideline panel composition account for the distributions of strong versus weak and the level of quality of evidence used? To what extent can the reasons for strong recommendations in the face of low or very low confidence in estimates be reproducibly characterized?

A preliminary consideration of these issues has led to further development of GRADE's prior approach to strong recommendations based on low or very low confidence in estimates of effect (Table 1a). In addition to the five paradigmatic

situations in which such recommendations may be appropriate, we have thus far formulated 5 paradigmatic situations (Table 1 b) in which strong recommendations based on low/very low confidence in estimates are made inappropriately by authors/panelists.

Study goals

The study goals are a) to document the distributions of PHG recommendations and confidence in estimates and b) to explore the factors that may contribute to panelists making strong recommendations in the face of low or very low quality evidence.

Study objectives

To address the study goals, we seek to answer several questions based on examination of the WHO guidelines that employ GRADE methods, and culminate in both a strength of recommendation and a confidence in estimates of effect:

1. How often are recommendations strong versus weak/conditional/discretionary?

2. How often are strong versus weak recommendations supported by high, moderate, low and very low quality evidence?

3. What are the reasons for strong recommendations based on low and/or very low confidence estimates?

Overview of study methodology

A group of nine researchers have developed this protocol and will carry out the project. There are two PHASES to this project. In PHASE I we will collect and describe all documents that have been proposed by WHO as guidelines, and to determine which utilize GRADE and result in a strength of recommendation and confidence in estimates grading. PHASE II focuses on the strong recommendations based on low/very low confidence in estimates and the possible reasons for the strong recommendations in the face of low/very low confidence. PHASE I collection and categorization of guidelines will be conducted by a single reviewer. PHASES I and PHASE II data extraction will be conducted by pairs of reviewers with 3rd party adjudication as needed. In PHASE II we will classify strong recommendations based on low/very low confidence according to the taxonomy in Table 1 a and b, with revisions as yet to be determined. Both PHASES I and II will result in the documentation of abstraction findings into a master MS EXCEL spreadsheet database for final analysis and summarization. The data abstraction for both phases will utilize a data abstraction tool in MS WORD format which will mirror the variables contained in the MS EXCEL database (and fully described in the relevant tables of this protocol). A user manual will accompany the MS WORD data abstraction tool providing further guidance on the variables sought, what they refer to, and response options.

PHASE I

1.) A single researcher will examine all WHO documents listed as guidelines in the publicly available WHO guideline website http://www.who.int/publications/guidelines/en/index.html,[4] as well as any supplementary documents provided by WHO that are not yet publicly available but are part of the internal GRC vetting process, and determine potential eligibility. This will ensure that a complete set of WHO guidelines (date of publication by WHO from January 2007 to December 2012) will form the basis and data set for this project.

2.) Guidelines that are structured as position statements, policy statements, best practices, emergency updates, strategies, field manuals, guidance documents, checklists, toolkits, frameworks, technical notes and papers, reference guides, interim policy frameworks, handbooks, or recommendations without the use of GRADE, will not be eligible.

3.) `GRADEd` guidelines, meaning guidelines with the application of the GRADE methods with BOTH a strength of recommendation and confidence in estimate grading, will be eligible.

4.) The guidelines will be classified according to a system of 9 established WHO guideline sub-categories (consideration is being given to an updated classification system) as outlined on the publicly available WHO guideline

website.[4] Some of the additional documents shared by the WHO Secretariat to complete the full set of guidelines do not fit into the established 9 sub-categories, and for these we have created an additional sub-category termed 'no clear category'. An independent researcher from the research team will examine the guidelines to determine that those which do not fit any of the existing 9 sub-categories, do in fact warrant being placed into the sub-group labelled 'no clear category'. Existing 9 WHO guideline sub-categories:

- **Child Health**
- **Chronic diseases, injuries and disabilities**
- **Environmental health**
- **HIV/AIDS**
- **Maternal and reproductive health**
- **Mental health and substance abuse guidelines**
- **Nutrition**
- **Patient safety**
- **Tuberculosis**

Additional WHO guideline sub-category:

- **No clear category**

The guidelines that pertain to pandemic influenza, malaria, aging, measles, dengue, anthrax, and immunization, vaccines and biological fall within the newly created 'no clear category'.

5.) We will characterize the extent to which the same authors participate in more than one guideline.

6.) For some questions, the unit of analysis in this project is the guideline; for other questions, the individual recommendation (s) and not the guideline from which the recommendation (s) originated is the unit of analysis.

7.) The master MS EXCEL repository database, which will include information obtained from the data abstraction phases via the MS WORD abstraction tool, will comprise 3 spreadsheets:

Sheet 1- will include characteristics of all WHO guidelines irrespective of reference to GRADE methods. The publicly available guidelines[4] will be cross-checked against the internally held GRC guidelines (that were shared by WHO in MS EXCEL format) for duplicates and final eligibility. Once checked and eligibility affirmed, all documents/guidelines will be included in sheet 1. This sheet will categorize guidelines according to use of GRADE, and other variables as per Table 2 in this protocol.

Sheet 2- will list all WHO guidelines that utilized GRADE fully and resulted in both a strength of recommendation and confidence in estimates grading (either overall or for each outcome), and the distributions of strength and confidence grading. This step will initially be conducted by the lead researcher

(PA). A research team member will then conduct a verification of the distribution data.

Following data abstraction via the MS WORD tool, sheet 2 (Table 3) of the final MS EXCEL repository, will also document variables such as burden of illness, consideration of equity, and affordability. These judgments will be made in duplicate.

8.) A guideline may present recommendations with confidence in estimates ratings for each outcome, but without an overall confidence in estimate rating. GRADE provides the guidance that the overall confidence in estimates is the lowest confidence associated with any of the critical outcomes. The guideline should ideally specify which are critical outcomes. If the guideline rates the evidence for every outcome but does not specify which are critical, the pair of reviewers who were assigned to that recommendation in question, will discuss and make an inference on what is critical and on that basis specify the overall confidence.

Sheet 3- will include recommendations that are strong and based on low or very low confidence in estimates. The variables (reflected in Table 4 of this protocol) abstracted via the MS WORD abstraction tool will be those that could help clarify why guideline panelists make strong recommendations based on low or very low confidence in estimates. Reviewers will copy and

paste sections of the guideline into the abstraction tool, that they consider relevant to particular items.

9.) To complete Tables 1a and 1 b of the existing paradigm taxonomy (appropriate and inappropriate reasons for a strong recommendation based on low/very low confidence), each group member will review a proportion of the strong recommendations based on low or very low confidence in estimates to a) identify additional paradigmatic situations in which strong recommendations based on low/very low quality evidence are appropriate (in addition to the five already identified) and b.) identify any paradigmatic situations in which there is a clear explanation of why the recommendations have been inappropriately made (adding to the four reasons identified thus far: best practice recommendations, influencing decision-makers to ensure funding is available for a procedure, existing long-standing practice (s), will not be considered if it is a weak recommendation, and misclassified as strong based on low/very low confidence in estimates when it was moderate or high confidence in the first place). Any new possible paradigmatic situations will be brought to the entire group for discussion and either accepted as valid, modified and accepted as valid, or rejected. Group decisions will be based on consensus. Tables 1 a and/or 1 b will be amended accordingly. As this process proceeds, we may develop a more theoretically grounded conceptual

framework to understand the reasons for strong recommendations based on low or very low confidence.

10.) Prior to beginning data abstraction, we will conduct a calibration exercise. Two reviewers will be assigned the same 3 guidelines and will use the MS WORD abstraction tool to collect the data which they will do independently. The reviewers and other members of the group who are interested will review the responses together. Disagreements will be discussed in detail and reasons established and necessary clarifications developed.

11.) All primary response options will be binary or categorical. It is likely to be useful to have information from the guidelines that substantiates or explicates categories chosen for responses (this may be particularly the case for V/P). Reviewers will copy relevant sections from the guideline text and paste them directly into the MS WORD abstraction tool.

PHASE II - evaluation of strong recommendations based on low or very low quality evidence

12.) Data abstraction for PHASE I and for PHASE II (will be assessed in duplicate, with resolution of disagreement through discussion or if necessary through third party adjudication (Dr. Gordon Guyatt will function as the independent adjudicator). Once agreement is reached, or adjudication completed, the definitive information will be documented (this task will be

assigned to one of the reviewers at the time of distribution of the proportion of recommendations to review) and the information will be forwarded to the lead researcher for entry into the master MS EXCEL repository database. The original abstracted information will be maintained by the lead researcher to ensure documentation of disagreements.

13.) Extent of agreement/disagreement tabulations will be made and calibrated via a kappa statistic for all steps of this project that entail duplicate review.

14.) Prior to beginning PHASE II data abstraction we will conduct a calibration exercise as described in item 10 above with a focus now on the strong recommendations.

15.) PHASE II data abstraction will include issues of the outcomes chosen, whether resource use and feasibility were considered, and value and preference statements associated with individual recommendations. The taxonomy in Tables 1 a and 1 b, modified as described in point 8, will be applied to all strong recommendations based on low or very-low confidence in estimates that were made by WHO and emerging from the initial characterization. The aim is to focus on the recommendations that were made as strong based on low or very low confidence and assess if they are consistent or inconsistent with GRADE guidance in terms of fitting Tables 1 a of 1 b.

16.) The study team will review all recommendations that do not fit the taxonomy.

If there is no issue that arises to question this judgment, these recommendations will be categorized as inappropriate in Table 1 b. In other words, these will be regarded as having been made by WHO as inconsistent with GRADE guidance based on reviewer judgements against the existing taxonomies. It is these inappropriately made recommendations that would be the focus of further study. Note, Tables 1 a and 1 b will be updated in the near term in terms of titling and adjustments to content of 1 b.

Table 1 a: Reasons for strong recommendations with Low/Very Low confidence in estimates

Paradigmatic situations in which strong recommendations with low or very low quality evidence are appropriate given typical values and preferences		
	Condition	Example (s)
1	When low/very low quality evidence suggests possible mortality reduction in an acute (rather than chronic) life-threatening situation (i.e. high risk of death) and when adverse effects and costs are not prohibitive and an alternative life-prolonging intervention is not available	Fresh frozen plasma or vitamin K in a patient receiving warfarin with elevated INR and an intracranial bleed (low quality evidence)
2	When low/very low quality evidence suggests possible benefit and high quality evidence suggests important harm or a very high cost	Head to toe CT /MRI screening for cancer. Low quality evidence that it may benefit early detection but high quality evidence of possible harm and/or high cost (strong recommendation against this strategy)
3	When low/very low quality evidence suggests equivalence of two alternatives, but high quality evidence of less important harm for one of the competing alternatives	H. 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 in comparison to the alternatives of radiation therapy or gastrectomy but with high evidence suggests less harm/morbidity
4	When high quality evidence suggests equivalence of two alternatives and low/very low quality evidence suggests important harm in one alternative	Hypertension in women planning conception and in pregnancy: high quality evidence of equivalence of alternatives methyldopa, labetalol, and nifedipine ; low quality evidence of harm due to angiotensin converting enzyme inhibitors and angiotensin receptor blockers (strong recommendation against the harm alternative)
5	When low to high quality evidence suggests possible benefits in one important outcome (outcome A) and low/very low quality evidence suggests possibility of harm in critical outcome (outcome B) and the harm regarding outcome B is valued much more highly than any benefit vis a vis outcome A.	Testosterone in males with prostate cancer - low quality evidence suggests possible increase cancer spread (strong recommendation against this strategy) Dopaminergic agents in pregnancy in women with prolactinomas - low quality evidence of possible fetal harm (strong recommendation against this strategy)

Table 1 b: Inappropriate reasons for strong recommendations based on Low/Very Low confidence in estimates and misclassification of confidence

Paradigmatic situations in which inappropriate strong recommendations based on low/very confidence in estimates are made		
	Condition	Example (s)
1	Best practice recommendation	We recommend avoiding tests and treatments that provide little or no value to the patient
2	The tendency by policy makers to make strong recommendations when in fact they are weak driven by the desire to influence decision makers to ensure availability of funding for an intervention or to ensure the recommendation is taken seriously	Uncovering examples of this paradigm is challenging but we hold strong views that this is occurring although not formally documented
3	Novel intervention or new evidence is being brought to an existing long standing practice, that predates GRADE and Cochrane (already widely adopted)	Intermittent iron supplementation for women of reproductive age has long been recommended with only low quality evidence
4	Misclassification of confidence (strong recommendation based on high or moderate confidence yet incorrectly classified as strong based on low or very low confidence by panelists)	We recommend that clinicians prescribe and support intensive lifestyle (dietary, physical activity, and behavioral) modification to the entire family and to the patient, in an age-appropriate manner, and as the prerequisite for all overweight and obesity treatments for children and adolescents). A meta-analysis of randomized pediatric trials, commissioned by this Task Force, of combined lifestyle interventions (diet and exercise) for treating obesity showed a modest but significant effect on obesity (equivalent to a decrease in BMI of 1.5 kg/m2; $P < 0.00001$) when these interventions targeted family involvement. This was graded as low quality when in fact there is at least moderate evidence for benefits.

References

1. GRADE working group. Grading the quality of evidence and the strength of recommendations. url: http://www.gradeworkinggroup.org/intro.htm (Accessed September 10th 2012).

2. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. ***Journal of Clinical Epidemiology.*** **Volume 64, issue 4, 383-394, April 2011. Guyatt et al. (2011).** url: http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract (Accessed July 20th 2012).

3. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008; 336(7650):924-6.**

4. WHO guidelines approved by the Guidelines Review Committee. url: http://www.who.int/publications/guidelines/en/index.html (Accessed July 16th 2012).

5. Guideline: Use of multiple micronutrient powders for home fortification of foods consumed by infants and children 6–23 months of age. WHO. http://whqlibdoc.who.int/publications/2011/9789241502047_eng.pdf (Accessed on September 24th 2012).

CHAPTER 3:

Phase I: World Health Organization recommendations are often strong based on low confidence in effect estimates (2007-2012)

[1]Paul E Alexander, [2]Susan Norris, [3]Lisa Bero, [4]Victor M. Montori, [5]Juan Pablo Brito, [6]Rebecca Stoltzfus, Benjamin [7]Djulbegovic, [1,8]Ignacio Neumann, [9]Gordon Guyatt

[1] Doctoral student, Health Research Methods (HRM)
Department of Clinical Epidemiology and Biostatistics
McMaster University, 1200 Main Street West,
Hamilton, Ontario, L8N 3Z5, Canada.

[2] Guidelines Review Committee Secretariat
World Health Association (WHO),
Geneva, Switzerland

[3]Professor
University of California, San Francisco
Suite 420, Box 0613
3333 California Street
San Francisco, CA 94118 USA

[4]Professor of Medicine
Director, Healthcare Delivery Research Program
Investigator, Knowledge and Evaluation Research Unit
Divisions of Endocrinology and Diabetes and Health Care and Policy Research

Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA

[5]Researcher and clinician
Mayo Clinic
Rochester, Minnesota, USA
Plummer 3-35, 200 First Street SW, Rochester, MN 55905, USA

[6]Professor
Provost's Fellow for Public Engagement
Director, Global Health Program
Director, Program in International Nutrition
Division of Nutritional Sciences
120 Savage Hall
Cornell University
Ithaca NY 14853, USA

[7]University of South Florida
H Lee Moffitt Cancer Center
Florida, USA

[8]Department of Internal Medicine
School of Medicine
Pontificia Universidad Catolica de Chile

[9] Distinguished Professor, McMaster University Health Sciences Centre
1200 Main Street West, Room 2C12
Hamilton, Ontario, L8N 3Z5, Canada.

Corresponding authors: Paul E. Alexander and Dr. Gordon Guyatt

E-mail of Dr. Guyatt: guyatt@mcmaster.ca

Citation:

Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014 Jun; 67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

The present chapter is the Phase I of the overall thesis project and represents my/our initial attempt to provide a descriptive epidemiology of all WHO guidelines that used GRADE methods and were published from 2007 to 2012, in order to document the distribution of recommendations' strength and confidence in effect estimates. In so doing, this allowed access to specifically those recommendations that were strong and based on low or very low confidence in effect estimates (and a confirmation or not of our initial impressions and evidence). Moreover, it allowed the opportunity to lay the groundwork for further study on whether such recommendations were developed consistent or inconsistent with GRADE guidance (addressed in Phase II). This Chapter three (Phase I) can thus be regarded as a core component of the overall thesis for it provided the basis for the overall thesis.

Abstract:

Objectives: Expert guideline panelists are sometimes reluctant to offer weak/conditional/contingent recommendations. GRADE guidance warns against strong recommendations when confidence in estimates of effect 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 WHO guidelines that used the GRADE approach and graded both strength and confidence (GRADEd).

Methods and setting: We reviewed all WHO guidelines (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 where 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.

Key words: GRADE, strength of recommendation, confidence in effect estimate, public health guidelines, clinical practice guidelines, high uncertainty, World Health Organization

Background:

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, PHG on populations. To produce credible recommendations, CPG or PHG must follow rigorous quality standards[1] in their development. Such standards[1] include use of an evidence-based approach with rating of the confidence in estimates of effect (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 adopted by WHO.[5,6] GRADE provides guidance in standardization of guideline development including rating confidence in estimates of effect 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 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 labelled 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 paper), other relevant factors include the burden of illness, accessibility, feasibility, the extent of current suboptimal practice, and the impact on health inequities.

GRADE guidance warns against strong recommendations in the face of low or very low confidence in estimates for critical outcomes. 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). Additionally, there is strong evidence emerging from an examination of The Endocrine Society (TES) guidelines, which is that

guideline panels often prefer to make strong rather than weak recommendations even in the face of low confidence estimates.[7] This is challenging because the low confidence in estimates of effect signals large uncertainty regarding effects of the recommended action. Given that strong recommendations are courses of actions recommended to all or almost all patients and circumstances, this 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, very low). We are particularly interested in identifying strong recommendations based on low or very low confidence (and whether it was for or against an action).

Methods:

Data source:

In December 2012 we retrieved all the available WHO PHG. This included WHO guidelines that are available on the main WHO website[5] (under 'WHO Guidelines: a

selection of evidence-based guidelines';
http://www.who.int/publications/guidelines/en/index.html) as well as those retrieved from an internal WHO database which includes all final guidelines approved by the WHO guideline review committee (GRC) that covered 2007 to 2012. For this study, a guideline was defined as a document produced by WHO and available publicly on their website or as part of the GRC retrieved dataset 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).

Data abstraction:

For each eligible guideline, two reviewers (PA and SR) (independently and in duplicate) abstracted the recommendations and noted the confidence in estimates/quality of evidence (high, moderate, low, very low), and strength of recommendation (strong, weak/conditional). 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 topic designated by WHO: maternal and reproductive health, child

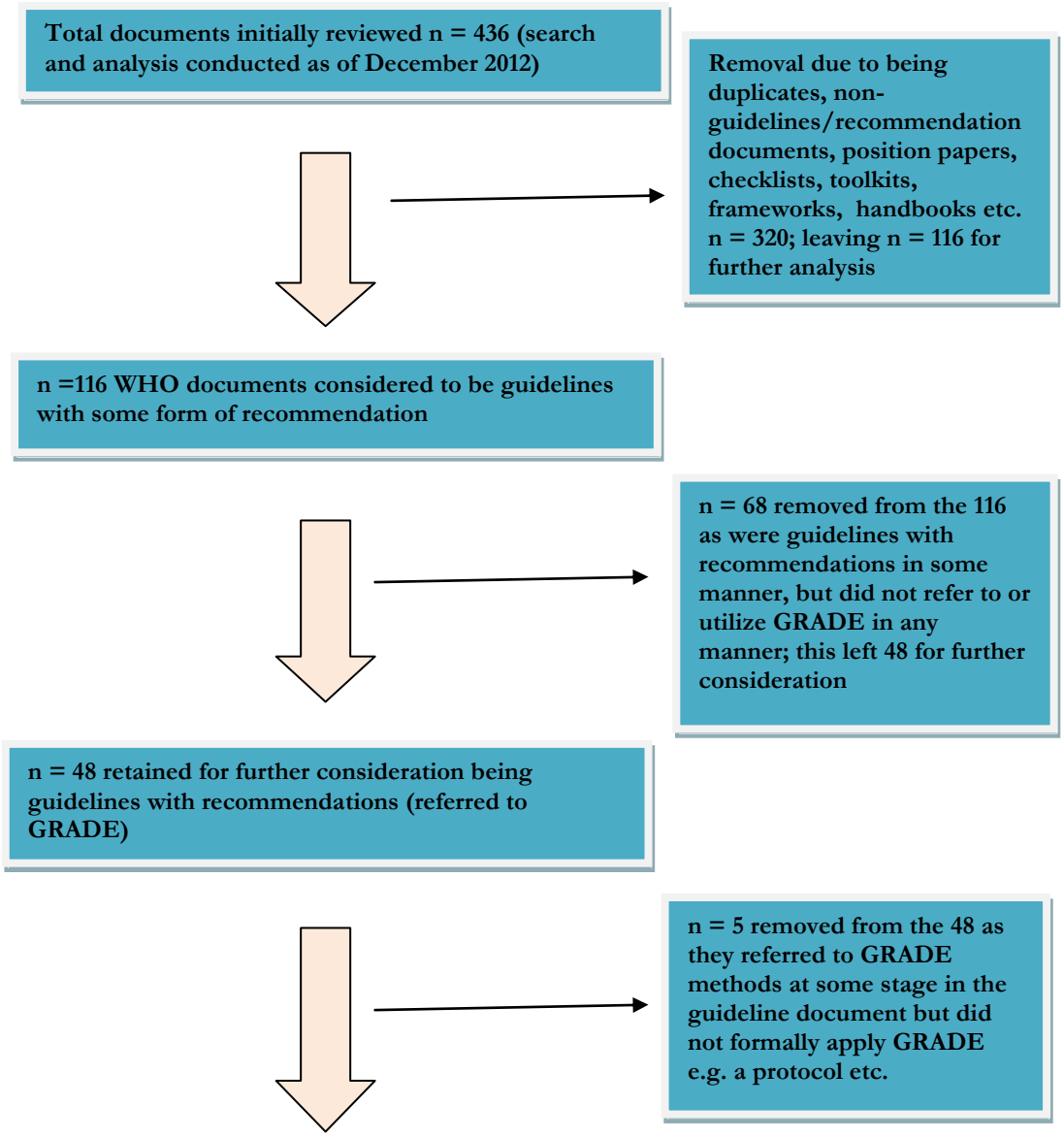
health, HIV/AIDS, tuberculosis related, pandemic influenza and nutrition guidelines etc. All abstracted guideline data was entered into a master MS EXCEL (2010) spreadsheet for management and cleaning (assessment of eligibility, removal of duplicates etc.) as well as during the duplicate data abstraction. Following abstraction, both reviewers consulted to address discrepancies in classifications of the strength and confidence and resolved this by consensus. A 3rd party adjudicator was not required given the clarity of the data. Kappa agreement score was computed. MS EXCEL was used for basic descriptive analysis of data classifications (frequencies, percentages).

Results:

We reviewed all 436 WHO documents (69 from the public WHO website and 367 from the internal GRC dataset) for eligibility of which 320 (73.3%) proved either duplicates or not guidelines, including position papers, checklists, toolkits, 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 5 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 (Figure 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%) based on very low confidence in effect estimates (Table 1). The kappa agreement score was 0.81 for abstraction phase. 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 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.

Figure 1: Flow Diagram of WHO Guidelines (documents) used for the WHO GRADE Guidelines Project (strong recommendations based on low/very low confidence in estimates)



n= 43 guidelines applied GRADE methods resulting in BOTH a strength of recommendation and confidence in estimate; these guidelines were used as the FINAL cohort for analysis (in description of classifications of recommendations)

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

Strong recommendations (n=289):	n (% column)	Weak recommendations (n=167):	n (% column)	Totals (%)
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%)
Totals (%)	289 (100%)	Totals (%)	167 (100%)	456 (100%)

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 over 50% of their recommendations as being strong; of these >50% were based on low or very low confidence in estimates (Table 2). These four guideline categories accounted for 88% of all strong recommendations based on low or very low confidence. Pandemic influenza and nutrition guideline recommendations were all strong and all 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).

A rapid update of available WHO guidelines was performed in late August 2013. In brief we found (via duplicate abstraction and full agreement): 94 recommendations emerging from 11 of 12 guidelines that were GRADEd (2013 release date but 2012 publication dates); 64 of 94 (68%) recommendations were strong; 52 of those 64 (81%) were strong with low or very low confidence (principally within the areas of chronic or non-communicable diseases, maternal and reproductive health guidelines, and mental health); when placed within the context of the existing 2007-2012 data, then the % WHO recommendations that were strong increases to 64% (from 63%) and the % strong that were based on low or very low confidence increases to 60% (from 55%). The number of

WHO guidelines using GRADE methods increases overall to 42% when examined from 2007 to 2013.

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

WHO guideline sub-category topic	Number Strong Low/Very Low/ total	% strong Low/Very Low	
WHO 6 of 9 guideline sub-categories examination period 2007-2012			
Maternal and reproductive health	28/54	50%	These 4 sub-categories account for 88% of all GRADEd strong recommendations
Child health	49/93	53%	
HIV/AIDS	32/66	49%	
TB	24/42	57%	
Nutrition	2/2	100%	
Mental health & substance abuse	6/8	75%	
WHO additional guidelines with no clear sub-category 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%	
Notes: a.) The 3 guideline topic areas i) chronic diseases, injuries and disabilities ii) patient safety and iii) environmental health, reported no GRADEd strong recommendations with low or very low confidence.			

Discussion:

The predominant finding was that WHO guideline panels often make strong recommendations (63%); over 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.

The large proportion of strong recommendations based on low confidence evidence may be inappropriately restricting the discretion of public health decision makers. When guideline panelists make strong recommendations they are suggesting that front line 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 estimates of effect Indeed, the GRADE working group has identified 5 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 of very low confidence estimates meet these conditions would be helpful.

Additional work on other determinants of strong recommendations that might be enlightening include the role of intellectual or financial conflicts of interest (balancing the need to utilize 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 confidence estimates, and the reasons why guideline panelists are making these recommendations (whether appropriate or inappropriate) is warranted.

Recent findings by Sinclair et al.[10] suggest that the WHO guideline development process has improved from the prior 2007 criticisms of WHO guidelines being based mainly on expert opinion, seldom using systematic evidence-based methods, or failing to follow the organization's prescribed guideline development process. Indications are that the WHO guideline development process and guideline quality as a whole has improved,[10] becoming more systematic and transparent. Our own cursory finding that 11 of 12 (91%) recently released WHO guidelines (2013 release date but 2012 publication dates) employed GRADE methods suggests a shift in the right direction in applying this systematic approach to grading evidence and recommendations. The challenge however,

and as demonstrated by the classification data we have presented for the examination period, is while GRADE may be increasingly applied within the guideline development process at WHO, there are concerns as to whether it is being appropriately applied especially within the context of high uncertainty.

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 in comparison 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)dall 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.

References

1. Institute of Medicine. Of the national academies. url: http://www.iom.edu/?ID=68004 (Accessed on April 25th 2013).

2. WHO Handbook for Guideline Development. 2012 url: http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf (Accessed on April 25th 2013).

3. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. ***Journal of Clinical Epidemiology.*** **Volume 64, issue 4, 383-394, April 2011. Guyatt et al. (2011).** url: http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract (Accessed March 10th 2013).

4. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.*** **2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

5. WHO. WHO guidelines approved by the Guidelines Review Committee. url: http://www.who.int/publications/guidelines/en/index.html (Accessed on April 24th 2013).

6.) WHO. Emergency response to antimalarial drug resistance. url: http://www.who.int/en/ (Accessed on April 24th 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

8. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M,Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G. GRADE guidelines 15: Going from evidence to recommendation-determinants of a recommendation's directionand strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

9. Guyatt G, Akl EA, Hirsh J, Kearon C, Crowther M, Gutterman D, Lewis SZ, Nathanson I, Jaeschke R, Schünemann H. The vexing problem of guidelines and

conflict of interest: a potential solution. ***Ann Intern Med.*** **2010;152(11):738-41. doi: 10.1059/0003-4819-152-11-201006010-00254. Epub 2010 May 17.**

10. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World healthorganization guideline development: an evaluation. ***PLoS One.*** **2013; 8(5):e63715. doi: 10.1371/journal.pone.0063715. Print 2013.**

Acknowledgement:

This manuscript has not been submitted for publication anywhere else but the JCE.

Conflicts 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. Dr. Susan Norris is a current employee of the WHO within the guideline development and review committee department; all authors are connected in some manner to guidelines in terms of their area of research, review for WHO or another organization, or development.

Financial conflicts of interest:

Paul Alexander (principal author): none
Gordon Guyatt: none; developer of the GRADE methods
Susan Norris: none; but employed by the WHO within the guideline development department
Ignacio Neumann: none
Victor Montori: none
Rebecca Stoltzfus: none; functioned as a guideline developer for WHO in the past
Benjamin Djulbegovic: none
Lisa Bero: none; functioned as a guideline developer for WHO in the past
Juan Pablo Brito: none

Supriya Rave (SR) (BSc in Biology University of Toronto, MSc in EBHC University of Oxford) has contributed to the double screening of the guidelines and recommendations as to strength and confidence classification.

Authors' contributions:

Mr. Alexander contributed principally to the conception of the chapter three/study, data collection, analysis, and drafting of the article. Dr. Guyatt provided guidance. All of the coauthors contributed to the drafting of the manuscript and where needed, collaboration on the methods.

CHAPTER 4: Phase II:

World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance

[1]Paul E Alexander, [2]Juan P. Brito, [1,3]Ignacio Neumann, [4]Michael R. Gionfriddo, [5]Lisa Bero, [6]Benjamin Djulbegovic, [7]Rebecca Stoltzfus, [8]Victor M. Montori, [9]Susan L. Norris, [10]Holger J Schünemann, [11]Gordon H Guyatt

[1] Health Research Methods (HRM)
Health Sciences Building (HSB)
Department of Clinical Epidemiology and Biostatistics
McMaster University, 1280 Main Street West,
Hamilton, Ontario, L8N 3Z5, Canada.

[2]Assistant Professor of Medicine

Investigator, Knowledge and Evaluation Research Unit
Divisions of Endocrinology, Diabetes, Metabolism and Nutrition
Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[3]Internal medicine specialist
Department of Internal Medicine
Pontificia Universidad Católica de Chile
Santiago, Chile

[4]Mayo Graduate School, Mayo Clinic
Knowledge and Evaluation Research Unit, Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[5]Professor (Chair of Medicines Use and Health Outcomes)
The University of Sydney
Charles Perkins Centre
Camperdown NSW 2006

[6]Distinguished Professor
University of South Florida
H Lee Moffitt Cancer Center
Florida, USA

[7]Professor
Provost's Fellow for Public Engagement
Director, Global Health Program
Director, Program in International Nutrition
Division of Nutritional Sciences
120 Savage Hall
Cornell University
Ithaca NY 14853, USA

[8]Professor of Medicine
Director, Healthcare Delivery Research Program
Investigator, Knowledge and Evaluation Research Unit

Divisions of Endocrinology and Diabetes and Health Care and Policy Research
Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[9] Guidelines Review Committee Secretariat
World Health Association (WHO)
Geneva, Switzerland

[10]Professor and Chair
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
Michael Gent Chair
in Healthcare Research
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON L8S 4K1, Canada

[11] Distinguished Professor
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
McMaster University Health Sciences Centre
1200 Main Street West, Room 2C12
Hamilton, Ontario, L8S 4K1, Canada.

Corresponding author: Paul E. Alexander, doctoral candidate.

E-mail: elias98_99@yahoo.com

This manuscript was accepted for publication in the JCE journal, January/February 2015.

The present chapter four (Phase II) builds on the information gained in chapter three (Phase I) by being able to focus on the strong recommendations based on low or very confidence in effect estimates via a taxonomic assessment of consistency with GRADE guidance. Based on a prior Endocrine Society Guidelines study, and based on our initial anecdotal impressions and the elevated number of strong recommendations made by WHO, we were concerned about whether they were being suitably made in accordance with GRADE guidance. This chapter four study thus allowed us the chance to test this out and to consider further study that would deepen our understanding of why such recommendations are being made by WHO in the first place.

Abstract

Background

In 2003 the World Health Organization (WHO) adopted the GRADE system for development of public health guidelines. Previously we found that many strong recommendations issued by WHO are based on evidence for which there is only low or very low confidence in the estimates of effect (discordant recommendations). GRADE guidance indicates that such discordant recommendations are rarely appropriate but suggests five paradigmatic situations in which discordant recommendations may be warranted.

Objective

To provide insight into the many discordant recommendations in WHO guidelines.

Data sources

We examined all guidelines that used the GRADE method and were approved by the WHO Guideline Review Committee between 2007 and 2012.

Methods

Teams of reviewers independently abstracted data from eligible guidelines and classified recommendations either into one of the five paradigms for appropriately-formulated discordant recommendations or into three additional categories in which discordant recommendations were inconsistent with GRADE guidance: 1) the evidence warranted moderate or high confidence (a misclassification of evidence) rather than low or very low confidence; 2) good practice statements; or 3) uncertainty in the estimates of effect would best lead to a conditional (weak) recommendation.

Results

The 33 eligible guidelines included 160 discordant recommendations, of which 98 (61.3%) addressed drug interventions and 132 (82.5%) provided an explicit rationale for the strong recommendation. Of 160 discordant recommendations, 25 (15.6%) were judged consistent with one of the five paradigms for appropriate recommendations; 33 (21%)

were based on evidence warranting moderate or high confidence in the estimates of effect; 29 (18%) were good practice statements; and 73 (46%) warranted a conditional, rather than a strong recommendation.

Conclusions

WHO discordant recommendations are often inconsistent with GRADE guidance, possibly threatening the integrity of the process. Further training in GRADE methods for WHO guideline development group members may be necessary, along with further research on what motivates the formulation of such recommendations.

Word count: abstract: 306; main text: 2,687

Introduction

The World Health Organization (WHO) produces evidence-informed public health guidelines for clinicians, policy makers, and programme managers. [1] The goal is a transparent, systematic, and evidence-based process for decisions involved in developing guidelines, incorporating an in-depth analysis of the desirable and undesirable outcomes of public healthcare options.[2,3]

Criticism of their guideline development process in 2007[4] prompted WHO to formulate the Guidelines Review Committee (GRC) to oversee guideline development.[5] In that same year, WHO heightened the use of GRADE (Grading of Recommendations

Assessment, Development and Evaluation) system of guideline development, GRADE being adopted by WHO in 2003.[6-8] GRADE is an explicit, comprehensive, transparent, and pragmatic approach to guideline development that has been adopted by over 80 organizations worldwide (www.gradeworkinggroup.org) and provides detailed guidance on how to rate the confidence in estimates of effect (quality of evidence[6,7]) and how to develop recommendations based on evidence.[9,10] GRADE does not seek to eliminate subjective judgments – such judgments are an inevitable part of rating evidence and making /grading recommendations – but rather to make judgments transparent/explicit.

GRADE rates the confidence in effect estimates for benefits and harms as high, moderate, low or very low[7]; the overall confidence is based on the lowest confidence of the outcomes critical for decision making (if there is more than one critical outcome, the confidence in the overall estimates is based on the outcome with the lowest confidence). Recommendations are classified as strong (desirable consequences clearly do or do not outweigh undesirable consequences) or conditional (the balance of desirable and undesirable consequences is less certain).[11] Alternative designations for conditional are weak, discretionary, or contingent recommendations.

Determinants of the strength of recommendations include confidence in estimates of treatment effects, magnitude of the desirable and undesirable consequences of alternative courses of action, value and preference judgements required in trading off desirable and undesirable consequences, and resource use considerations.[10,11] In the context of public health decision-making, other factors including the burden of illness,

accessibility, feasibility, acceptability, barriers and facilitators for implementation, the extent of current suboptimal practice, and the impact on health inequities, also require consideration.

When guideline panels are not confident regarding the balance between desirable and undesirable consequences—which will be the case whenever evidence warrants low or very low confidence— GRADE discourages strong recommendations. There may, however, be circumstances in which strong recommendations are warranted despite low or very low confidence in estimates of effect (discordant recommendations). The GRADE working group has identified five paradigmatic situations in which this may be the case (Table 1).[10]

A recent study[12] of WHO guidelines reported that among 456 recommendations contained in 43 GRC-approved guidelines using GRADE, 63.4% (289/456) were strong and the remainder conditional. Of the 289 strong recommendations, 160 (55.5%) were discordant recommendations.

Table 1: Paradigmatic situations (and frequencies from the present study) in which panels may reasonably offer strong recommendations based on low or very low confidence in effect estimates (discordant recommendations).

Paradigmatic situation	Confidence in effect-estimates for health outcomes (Quality of evidence)		Balance of benefits and harms	Values and Preferences	Resource considerations	Recommen-dation	Example of discordant recommendations from WHO guidelines*	Frequency emerging from the present study (%)
	Benefits	Harms						
Life threatening situation	Low or very low confidence	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation. Adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits	Strong recommendation in favor	In the treatment of patients with MDR-tuberculosis, a fluoroquinolone should be used.[13]	7/160 (4.4)
Uncertain benefit, certain harm	Low or very low	High or Moderate	Possible but uncertain benefit. Substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	Possible high incremental cost (or resource use) may further mandate a recommendation against the intervention	Strong recommendation against (or in favor of the less harmful/less expensive alternative when two are compared)	*No example from the present study of WHO guidelines; we include an example from elsewhere[14]: We recommend against screening for androgen deficiency in the general population.	0 (0.0)
Potential equivalence, one option clearly less risky or costly	Low or very low	High or Moderate	Magnitude of benefit apparently similar - though uncertain - for alternatives. We are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits, may further support recommendation for less harmful alternative	Strong recommendation for less harmful/less expensive	For management of post partum haemorrhage, oxytocin should be preferred over ergometrine alone, a fixed-dose combination of ergometrine and oxytocin, carbetocin, and prostaglandins.[15]	8/160 (5.0)
High confidence in similar benefits, one option potentially more risky or costly	High or Moderate	Low or very low	Established that magnitude of benefit similar for alternative management strategies. Best (though uncertain) estimate is that one alternative has appreciably greater harm.	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) of one alternative	Strong recommendation against the intervention with possible greater harm or cost	*No example from the present study of WHO guidelines; we include an example from elsewhere[16]: In women requiring anticoagulation and planning conception or in pregnancy, the AT9 guidelines recommended against the use of certain anticoagulants. For example, high confidence estimates suggests similar effects of different	0 (0.0)

							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).	
Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) of potentially harmful intervention may further justify recommendation for less harmful.	Strong recommendation against the intervention (or in favor of the less harmful/less expensive alternative when two are compared)	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.[17]	10/160 (6.3) Total=25/ 160 (15.6)

Though these findings raise the possibility of excessive use of strong recommendations in the face of uncertainty in estimates of effects, the reasons for the strong recommendations remain unclear. Most recommendations might, for example, be compatible with one of the situations outlined in Table 1. Other possibilities are that the guideline panel misjudged the confidence in effect estimates (i.e. that evidence actually warranted moderate or high confidence); the recommendations actually represent good practice statements in which we have high confidence in estimates, but that confidence is based on indirect evidence that would be excessively time-consuming to document and therefore best not subjected to the GRADE process; [9,11] or that conditional

recommendations would have been preferable (see Table 2). Therefore, using a previously developed taxonomy,[10] we classified WHO recommendations according to these possibilities**.**

Methods

We included all guidelines approved by the WHO GRC between 2007 and 2012 that applied GRADE methods.[1,12] The final study cohort comprised 33 guidelines containing 160 discordant recommendations there in (a previous study[12] displays the Flow diagram for included guidelines).

We evaluated the extent to which each guideline documented the reasons for the discordant recommendations. We considered the rationale to be transparent when the guideline development group provided a rationale (in some instances a few lines; in others, a more extensive description).

Study co-authors performed data abstraction and the taxonomy exercise independently and in duplicate. Early in the data abstraction process we encountered challenges when the comparators to the interventions were not explicit or obvious. We therefore classified the comparator as: i.) explicit, identified clearly in the recommendation ii.) not explicit in the recommendation, but obvious or easy to infer from the guideline text and iii.) not identified in the recommendation and unclear in the guideline text. For category (iii), reviewers used their best judgement to determine the

likely comparator. When there was a disagreement, discussions to achieve consensus took place and, when necessary, a third party adjudicated the discussion. Table 3 provides examples of each of the three categories of comparators in terms of explicitness.

Three reviewers working in pairs independently classified each of the 160 recommendations as either consistent with one of the five previously identified optimal categories for discordant recommendations (Table 1), or in one of three categories in which we judged discordant recommendations to be inconsistent with GRADE guidance: 1.) misclassification of evidence; 2.) good practice statements; or 3.) more likely conditional recommendations (Table 2). Reviewers resolved disagreement through consensus discussion or if necessary third party adjudication. We calculated chance-corrected agreement[18] for whether recommendations were or were not consistent with GRADE guidance (i.e. consistent with Table 1 or Table 2) using the kappa statistic.

Results

The guidelines in the study cohort covered a broad spectrum of healthcare topics including maternal and reproductive health, child health, HIV/AIDS, and tuberculosis. These four WHO guideline topic areas were responsible for 83% of the discordant recommendations (133/160); the proportion of discordant recommendations in these four areas varied from 49% to 57%.

Interventions included drugs (61.3% of recommendations), screening programs (14.3%), medical devices (8.1%), and other (16.3%) e.g. growth, nutrition, and development monitoring (1.9%); breast-feeding (1.9%); vitamins and mineral supplementation (1.9%); healthcare policy (10%), and manual therapeutic interventions (0.6%).

Most guidelines (82.5%) provided some form of rationale for making the discordant recommendations, in some instances more extensive than others. For example, one guideline indicated: "in women with histologically confirmed cervical intraepithelial neoplasia (CIN), panelists presented the following explanation: "The expert panel recommends cryotherapy over no treatment": "This recommendation is strong, despite the presence of very-low-quality evidence. The expected benefit of cervical cancer prevention is very high but there is uncertainty related to the occurrence of adverse outcomes. There was very low-quality evidence for the occurrence of spontaneous abortions and infertility but the risk appeared similar to that in the general population. Although neither the risk of HIV acquisition in HIV-negative women nor the risk of HIV transmission by HIV-infected women who undergo cryotherapy is known, the current limited data do not suggest that there is an increase in the risk of HIV acquisition/transmission".

Reviewers judged 25 (15.6%) of the 160 discordant recommendations to be consistent with one of the five paradigmatic situations in which it is appropriate to offer

discordant recommendations. The most common (10/42%) was paradigm 5 (potential catastrophic harm) (Table 1).

Reviewers judged 33 of the discordant recommendations (20.6%) to represent a misclassification of confidence (evidence warranted moderate or high confidence); 29 recommendations (18.1%) as good practice statements; and 73 (45.6%) as warranting conditional recommendations (Table 2).

The comparator was explicit in 28 of the 160 (17.5%) recommendations; not explicit but easily inferred in 43 (26.9%), and not easily identified in 89 (55.6%) (Table 3).

Kappa estimate for the taxonomic judgment regarding whether the recommendation was consistent or inconsistent with GRADE guidance was 0.68. Third party adjudication to determine the appropriate classification was required in 11 (7%) recommendations.

Table 2: Discordant recommendations judged to be sub-optimally made in WHO guidelines

Condition	Example from a WHO guideline	Frequency from existing study, n (% of 160)
1. Misclassification of the evidence as low or very low when in fact it should have been medium or high	Recommendation: 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). Taxonomy judgement: In a randomized trial, couples who received CHTC versus health information reduced unprotected sex, providing moderate quality evidence supporting the recommendation.	33 (20.6)
2. Good practice statement (panel should not apply GRADE methods). A large body of difficult to summarize indirect evidence (an extremely large volume of evidence that may also extend over a long period of time) indicates that the desirable consequences of the intervention are much greater than the undesirable consequences (i.e. confidence is actually high, but summarizing the evidence systematically would be a poor use of a panel's resources)	Recommendation: Triage people with tuberculosis symptoms (strong recommendation, low quality of evidence). Taxonomy judgement: This recommendation suggests that persons with a sufficiently high probability of having tuberculosis should be promptly separated from other patients and promptly undergo the appropriate investigations. In this example of a good practice statement, there are no randomized trials or observational studies that compare triage to no triage, therefore no direct evidence. There is, however, evidence warranting high confidence that signs and symptoms can distinguish those with substantial probability of tuberculosis from those with low probability, and that isolation procedures can reduce the spread of tuberculosis. It may not be a good use of a panel's time to collect and summarize these bodies of evidence.	29 (18.1)
3. Recommendations inconsistent with GRADE guidance (guidance suggests conditional recommendations and not strong)	Recommendation: Uterine massage is recommended for the treatment of post partum haemorrhage (strong recommendation, very low quality evidence). Taxonomy judgement: 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.	73 (45.6)
Total		135 (84.4%)

Table 3: Judgements of the extent of explicitness of comparators in WHO recommendations and guidelines

Situation	Recommendation example	Frequency, n (% of 160)
Comparator explicit in the recommendation	The expert panel recommends cryotherapy over no treatment (strong recommendation, very low quality of evidence).	28 (17.5)
Comparator not explicit in the recommendation but explicit or clear in associated guideline text	High-dose vitamin A supplementation is recommended in infants and children 6–59 months of age in settings where vitamin A deficiency is a public health problem (strong recommendation, low quality evidence) The associated text made it clear that the recommendation was for high dose over low dose vitamin A.	43 (26.9)
Comparator not explicit in the recommendation and the associated text also failed to clarify	Offer and promote postpartum and post-abortion contraception to adolescents through multiple home visits and/or clinic visits to reduce the chances of second pregnancies among adolescents (strong recommendation, very low quality of evidence). Neither the recommendation nor the associated text made it clear whether the comparator was no offer or promotion of post-abortion contraception, or less intense or alternative programs.	89 (55.6)

Discussion

In a prior study, we found that a majority of strong WHO recommendations that used the GRADE approach across a broad range of topics were discordant.[12] The finding that the four topic areas maternal health, child health, HIV/AIDS, and tuberculosis had a very similar frequency of discordant recommendations, approximately 50%, suggests that the phenomenon of discordant recommendations is a systemic issue across guidelines developed by WHO.

Most guideline panels (over 80%) offered a rationale for the strong recommendation despite the high uncertainty about the intervention effects. Despite the

rationale provided, we found that only a minority of discordant recommendations were consistent with the GRADE taxonomy of situations in which discordant are appropriate (Table 1). Almost half of the discordant recommendations were judged as warranting only a conditional (weak) recommendation (Table 2).

An additional important finding is that the comparators to the interventions of interest were neither made explicit in the recommendations nor adequately clarified in the associated text in more than half the recommendations. When comparators are not clear and explicit, guideline users must make inferences that may not always be correct.

Strengths and limitations

We examined the entire sample of WHO guidelines approved by the GRC between 2007 and 2012 [1,12] that used the GRADE approach. The taxonomy that we used has been successfully implemented in a prior study of clinical guidelines.[19] Our reviewers all had extensive formal clinical epidemiology training and were very familiar with the GRADE approach to rating confidence in estimates and grading strength of recommendations. The entire investigative team conducted extensive discussions to deepen understanding of the classification approach, and the reviewers participated in additional calibration exercises. This preparation resulted in a satisfactory chance-corrected agreement of 0.68 [18] in the classification of recommendations as consistent or inconsistent with GRADE guidance.

One limitation of our study is that our decisions were based only on information in the published guideline document. Guideline panels may have considered additional factors. Our study is also limited in that it examined only the first six years of WHO experience with GRADE. It is possible that with further experience and education, WHO sponsored guideline panels will make fewer decisions inconsistent with GRADE guidance. Furthermore, our results may have limited applicability to other groups using GRADE to produce practice guidelines. As we note below, however, another formal study of GRADEd recommendations revealed similar findings to this investigation.[19]

A final limitation is that, although for the first two categories of recommendations inconsistent with GRADE guidance (Table 2) we know the problem was an excessively low rating of confidence, the reasons in the third category are not clear. Explanations in the text were at times helpful, but not definitive. Consider, for instance, the recommendation for cryotherapy in women with histologically confirmed cervical intraepithelial neoplasia (CIN) and summarized in the results section. The panel's explanation leaves open the possibility that they did not consider the potential adverse consequences of cryotherapy critical to the decision, or they were sufficiently confident that these adverse consequences would not occur to support a strong recommendation.

There are a number of possible explanations for discordant recommendations made by WHO that we classified as warranting only conditional recommendations. One is that panels have a fundamental disagreement with GRADE's rating of confidence. For instance, guideline panellists may believe that observational studies (in some instances the

only type of evidence available) warrant moderate or high confidence, and though they follow the GRADE system in making their explicit confidence ratings, their recommendations are based on their own internal judgment. Alternatively, panellists may feel that, if they make conditional recommendations, decision makers will ignore their work and their suggestions. Another possibility is that they may feel wedded to long-established practices, and feel uncomfortable issuing any but strong recommendations regarding such practices. A related explanation is that they may be convinced that they know what is best for patients, and that they should do all they can to ensure optimal care. Finally, financial or non-financial conflicts of interest may be driving the recommendations. There may be other explanations beyond those we have suggested here.

Relation to previous work

One prior study involved a formal structured exploration of discordant recommendations using the GRADE approach.[19] This study found that of 357 recommendations produced by the Endocrine Society, 206 (57.7%) were strong and of these, 121 (59%) were discordant recommendations. Of the discordant recommendations, 35 (29%) were consistent with GRADE guidance (Table 1), 43 (35.6%) were good practice statements, 5 (4%) a misjudgement of confidence (greater confidence in estimates was warranted), 5 (4%) were recommendations for research, and 33 (27%) would more appropriately have been graded as conditional recommendations.

These results are very similar to our present WHO findings, suggesting that the concerns that arise from our investigation are not limited to WHO guidelines.

Implications

These WHO results suggest that consumers of guidelines should view discordant recommendations with skepticism. Of the three categories inconsistent with GRADE guidance – misjudgement of the quality of the evidence, good practice statements, and warranting only a conditional recommendation – the third is of most concern. Good practice statements may be appropriate when indirect evidence that is difficult to collect and summarize actually warrants high confidence in intervention impact, and when the gradient between desirable and undesirable consequences of an intervention is large. Thus, like a misjudgement of evidence quality as low or very low when moderate or high confidence is appropriate, the problem with good practice statements is not the strong recommendation, but rather the confidence in estimate judgement. In these two situations, the strong recommendation, and thus the most important message to the clinician and the policy maker – just do it – is warranted.

This is not true for the final category, recommendations that would optimally have been graded as conditional. This is not a small concern: 46% of WHO discordant recommendations, and 16% of all their recommendations made from 2007 to 2012, fall in this category.

If our reviewers' judgments are accurate, the message from discordant WHO recommendations to the policy-making community may be problematic. WHO member

states and sub-national decision makers often - and particularly in low resource settings – need to set priorities with regard to alternative interventions. When guideline panellists make strong recommendations, they are suggesting that front line decision-makers need not consider the issue any further, and should simply implement the suggested course of action. Public health officials who view WHO guidelines as authoritative may feel that, in the face of such strong recommendations, they should put aside concern that the recommendation may not be optimal, or should have only a low priority, in their setting. If they do respond in this way, these strong recommendations may be inappropriately restricting the discretion of public health decision-makers.

On the other hand, there may be reasons that the panellists are correct in their judgment and it is the GRADE guidance that is problematic. First, there is debate in the methodological community regarding the confidence warranted by observational studies. If observational studies really do warrant higher confidence than attributed by GRADE, then strong recommendations based on such evidence may be appropriate. Second, panellists may feel that their conditional recommendations will be ignored, and they may be right. Perhaps it is better to make recommendations that are inappropriately strong than to make recommendations that will be ignored. Research on how policy-makers interpret strong and conditional recommendations may be useful to strengthen communications between WHO panels and end users of the recommendations.

WHO and other guideline developers using GRADE need to make clear policies regarding the extent to which they will adhere to GRADE guidance. If they do adhere, further training of their panels on GRADE principles and their application may be necessary. Our findings suggest that WHO sponsored guideline development groups may benefit from increased awareness of the need to clearly specify not only interventions but comparators; of the five situations in which strong recommendations may be warranted in the face of low or very low confidence estimates (Table 1); and of the inadvisability of strong recommendations when these criteria are not met. Given that almost 21% of discordant recommendations represented instances of under-rating of confidence (should have been moderate or high rather than low or very low), WHO panelists may also benefit from education regarding standards of rating evidence and particularly the application of indirect evidence. Previous findings from the Endocrine Society guidelines[19] suggest that other panels may also benefit from similar awareness and education.

Conclusion

There are major limitations in the extent to which WHO sponsored guideline development groups adhere to GRADE guidance in issuing discordant recommendations. Prior evidence suggests that this is also true of other guideline panels. These results suggest that organizations using GRADE should conduct a formal review of the relevant GRADE principles. Following such a review of GRADE principles, additional training

of guideline panellists, perhaps enhanced by audit and feedback and ongoing methodological support, may be required.

References

1. World Health Organization (WHO). WHO guidelines approved by the Guidelines Review Committee. Accessed from url: http://www.who.int/publications/guidelines/en/; May 27th 2012.

2. Graham R, Mancher M, Wolman DM, Greenfield S, Steingerg E. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Institute of Medicine. Clinical Practice Guidelines We Can Trust. 1st ed. Washington, DC: National Academies Press; 2011.

3. Institute of Medicine. Of the national academies. url: http://www.iom.edu/?ID=68004 (Accessed on April 25th 2013).

4. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet*** **2007;369: 1883-9.**

5. WHO Handbook for Guideline Development. 2012 url: http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf (Accessed on April 25th 2013).

6. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. ***J Clin Epidemiol*** **2011; 64(4):380-2.**

7. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.*** **2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

8. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008 Apr 26;336(7650):924-6. doi: 10.1136/bmj.39489.470347.AD.**

9. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. ***J Clin Epidemiol.*** **2011; 64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026. Epub 2010 Dec 31.**

10. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori VM, Brito JP, Norris S, Elbarbary M, Post P, Nasser M, Shukla V,Jaeschke R, Brozek J, Djulbegovic B, Guyatt G. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. ***J Clin Epidemiol.*** **2013; 66(7):726-35. doi: 10.1016/j.jclinepi.2013.02.003. Epub 2013 Apr 6.**

11. Brozek JL, Akl EA, Alonso-Coello P, Lang D, Jaeschke R, Williams JW, Phillips B, Lelgemann M, Lethaby A, Bousquet J, Guyatt GH, Schünemann HJ; GRADEWorking Group. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. ***Allergy.*** **2009;64(5):669-77. doi: 10.1111/j.1398-9995.2009.01973.x.**

12. Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014 Jun; 67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

13. WHO. Guideline: Guidelines for the programmatic management of drug-resistant tuberculosis. Geneva, World Health Organization (WHO), 2011 (url: http://whqlibdoc.who.int/publications/2011/9789241501583_eng.pdf?ua=1).

14. Bhasin S, Cunningham GR, Hayes FJ, et al. Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. ***J Clin Endocrinol Metab.*** **2010;95:2536–2559.**

15. WHO. Guideline: WHO guidelines for the management of postpartum haemorrhage and retained placenta. Geneva, World Health Organization (WHO), 2009 (url: http://whqlibdoc.who.int/publications/2009/9789241598514_eng.pdf).

16. Bates SM, Greer IA, Middeldorp S, Veenstra DL, Prabulos AM, Vandvik PO; American College of Chest Physicians. VTE, thrombophilia, antithrombotic therapy, and pregnancy: Antithrombotic Therapy and Prevention ofThrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. ***Chest.*** **2012;141(2 Suppl):e691S-736S. doi: 10.1378/chest.11-2300.**

17. WHO. Guideline: Rapid Advice. Treatment of tuberculosis in children. Geneva, World Health Organization (WHO), 2010 (url: http://whqlibdoc.who.int/publications/2010/9789241500449_eng.pdf).

18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. ***Biometrics.*** **1977 Mar;33(1):159-74.**

19. 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

Acknowledgements: Ms. Supriya Rave played a role in screening and data abstraction. She had no role in the interpretation of findings or the final conclusion and the positions advocated by the larger research team are not to be ascribed to her. Dr. Lehana Thabane (McMaster University, Vice-Chair) provided some editorial and technical assistance of which we are very grateful.

Authors' contributions:

Mr. Alexander contributed principally to the conception of the chapter/study, data collection, analysis, and drafting of the article. Dr. Guyatt provided guidance. All of the coauthors contributed to the drafting of the manuscript and where needed, collaboration on the methods.

Chapter 5: Protocol for Phase III:

Understanding why WHO guideline panels make strong recommendations in the face of low or very low confidence (study quality) in effect estimates; a qualitative descriptive study using inductive content analysis

Written by Paul Alexander with guidance from Dr. Gordon Guyatt

This Chapter five protocol for Phase III follows the Phase I descriptive study (Chapter three) and Phase II taxonomy study (Chapter four) and sought to lay out a qualitative study design that would allow a deeper understanding of why WHO guideline panelists make strong low confidence in effect estimate recommendations. This was based on one-on-one interviews that went beyond the explicit information contained in the examined guideline documents that were part of the Phase II. The rationales for such recommendations were not clear in guidelines and thus we decided to go to source so that we can understand the reasons, anticipating that WHO panels may have been entirely reasonable in their judgements as well as were not developing the recommendations consistent with GRADE guidance.

Background

The World Health Organization (WHO) remains a leader in the production and dissemination of public health guidelines that play a role in shaping and informing global public health policy. Such guidelines are particularly important as a guide to public health for low and middle income nations. However, in 2007, researchers concluded that WHO guidelines and the ensuing recommendations were being made based principally on expert opinion and not on systematic reviews of the relevant evidence.[1,2] This finding prompted WHO to initiate the Guideline Secretariat (Guideline Review Committee (GRC)) in 2007. This was accompanied by the heightened use of the Grading of Recommendations Assessment, Development and Evaluation (GRADE)[3-8] approach to guideline development (adopted by WHO in 2003). GRADE provides guidance in standardization of guideline development including rating confidence in estimates of effect and moving from evidence to recommendations.

The GRADE approach categorizes the confidence in effect estimates (also known as quality of evidence) as high, moderate, low, and very low. Within this grading system, randomized controlled trials (RCTs) begin as high confidence and observational studies as low confidence. High risk of bias, imprecision, indirectness, inconsistency, and likelihood of publication bias, can lower confidence in effect estimates. Confidence can increase if effect estimates suggest large intervention effects or 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 labelled 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.[3-5] In the context of public health guideline (PHG) development, other relevant factors (while not exhaustive) include the burden of illness, accessibility, feasibility, acceptability, equity, social and political context, the extent of current suboptimal practice, and the impact on health inequities.

GRADE guidance warns against strong recommendations when there is low or very low confidence in estimates of effect for critical outcomes.[3-6] There is a strong impression that guideline panels often prefer to make strong rather than weak recommendations even in the face of low confidence estimates. This has proved true of the Endocrine Society Guidelines.[9]

Statement of the problem

An understanding of when and why strong recommendations are formulated at WHO when the evidence is low quality (either consistent or inconsistent with GRADE guidance), is an important methodological issue that has implications for WHO and a wide variety of guideline development groups internationally. If WHO guideline developers are frequently making strong recommendations based on low or very low

confidence in effect estimates, it raises concerns about whether GRADE is being optimally applied in the WHO guideline development process. If rigorous and valid procedures are not being followed, WHO guidelines may not be optimally evidence-based, recommendations may not provide public health practitioners and policy makers an appropriate degree of discretion in their decision-making, practices that ultimately prove to be ineffective or harmful may be widely implemented, and research may be inhibited.

This study thus adds the opportunity for the guideline development group (GDG) chair to explain the rationale for those recommendations not clearly consistent with GRADE guidance, providing additional insight into these recommendations. This study of WHO guidelines will therefore help both WHO guideline developers and others to understand the process of developing optimal recommendations, and when strong recommendations based on low quality evidence are optimal and when they may not be optimal.

We view the GDG chair as the most suitable person to provide the viewpoints of the other panel members and have therefore selected the GDC chair as our focus. Remarks made by chairs during the interview will be considered as an overall impression of the panel members and not strictly their personal perspective. It may be that panelists were well positioned to make such recommendations in the first place and the study will help uncover what were the precipitating factors for making the recommendations. In some cases for example, the interviews may reveal that, for reasons not initially evident to

researchers, the recommendations are consistent with GRADE guidance. Furthermore, the responses may reveal instances when GRADE guidance is suboptimal and requires modification.

This qualitative interview study is the last Phase (III) of a larger study comprising three phases: Phase I (guideline and recommendation descriptive epidemiology) which has been completed and an article describing the results is in press; Phase II (reasons for strong recommendations based on low or very low confidence and taxonomic typology) nearing completion; and Phase III, a qualitative study which we describe fully below. In other words, this study provides the opportunity to better understand this important step in guideline development: the descriptive study (Phases I and II) might have led to an incorrect label either because the researchers did not have all relevant information, or because current GRADE guidance is suboptimal.

To best characterize recommendations, we have also adopted the following terms: "clearly consistent with GRADE guidance" and "not clearly consistent with GRADE guidance". Recommendations "not clearly consistent with GRADE guidance" will be further categorized as:

i. apparently best practice recommendations;
ii. apparently moderate or high confidence in estimates;
iii. apparently better presented as weak recommendations

What was found in Phases I and II that drives the need for this study? In the fall of 2012, we initiated study reviewing all WHO guidelines from 2007 to 2012 (those already publicly available as well as guidelines making their way through the approval process, and shared with the research team by the WHO GRC Secretariat).

In brief, results for the 2007 to 2012 examination period showed that of 116 GRC-approved WHO guidelines, 43 (37%) utilized the GRADE method, reporting both

a strength of recommendation and a confidence in effect estimates. These 43 guidelines included 456 recommendations of which 289 were strong (64%). Of these 289 strong recommendations, 160 (55%) were based on low or very low confidence. Four WHO guideline topic areas (of the nine presently housed on the public WHO website): maternal and reproductive health, child health, HIV and AIDS, and tuberculosis, accounted for near 90% of all strong recommendations based on low or very low confidence in effect.

In examining the GRADEd guidelines we have found: i) recommendations that appeared poorly structured ii) apparently best practice recommendations being presented as strong recommendations based on low or very low confidence iii) apparently limited understanding of indirect evidence and iv) strong recommendations in the face of low or very low confidence estimates not clearly consistent with GRADE guidance. These inferences regarding classification of recommendations based on the guidelines may or may not be accurate. Information from WHO panel chairs will help to determine the accuracy of the classification/inferences. Further, even if the determination is accurate, understanding the reasoning of the panel will help identify educational needs of panelists regarding optimal use of GRADE methodology. Finally, even if the determination is accurate, the reasoning of the panel may highlight limitations in GRADE guidance.

Thus, to build upon the findings emerging from Phases I and II and to further understand the process of making strong recommendations based on low or very low confidence in effect estimates, we propose to now interview guideline development panel chairs. When possible, which we anticipate is the norm, we intend to interview one

content and one methods panel chair per eligible guideline, asking about possible reasons for making the large number of strong recommendations based on low confidence estimates. In cases whereby the guideline does not have two panel chairs, or does not indicate having content or methods chairs or only indicates having one chair, we will interview the WHO personnel identified as the lead member. The number of interviews will be adjusted accordingly.

Objectives

Through the interviews with panel chairs who were involved in the development of guidelines in the four topic areas (HIV/AIDS, TB, maternal and reproductive health, and child health) which were characterized by a large proportion of strong recommendations based on low or very low confidence, we hope to gain an understanding of the reasons for the strong recommendations.

We anticipate the following possible outcomes of the interviews with the GDG chair/lead:

a. Our judgments as researchers on the reasons for the strong recommendations may be inaccurate and the recommendations are inconsistent with GRADE guidance; and/or
b. The reflections of the GDG chair may lead us to question whether GRADE guidance is optimal, and provides a possible stimulus for modifying GRADE; and/or
c. The GDG chair may acknowledge in retrospect that:
    i. GRADE guidance is reasonable and
    ii. The recommendations were inconsistent with GRADE guidance.

In these instances we will identify the reasons for the discrepancy between GRADE guidance and practice. This understanding will help us identify opportunities for improving GRADE guidance as well as areas for further education of GDG members on GRADE.

Outcome measures

We will undertake a qualitative descriptive analysis and summarization of the interview responses from guideline panel chairs regarding their perception of the process whereby their panels made strong recommendations in the face of low or very low confidence estimates of effect. Responses will be transcribed and codes, themes, and commonalities will be abstracted via manual coding and the use of NVivo 10 qualitative analysis software.

Methods

Rationale for guideline selection (eligibility)

For the phase III qualitative descriptive study, the WHO guideline publication years 2011 and 2012 will be used for guideline inclusion. Choice of these most current publications increase the likelihood that the panel chairs will be accessible and that they will have access to prior notes and be able to recall the panel process.

We will focus on the four guideline topic areas (maternal and reproductive health, HIV and AIDS, child health, and TB) given that they account for the majority of WHO strong recommendations based on low or very low confidence. From the four topic areas, we will select 12 guidelines that include examples of strong recommendations based on low or very low confidence in effect estimates that are not clearly consistent with GRADE guidance.

Selecting guidelines recommendations that focus on strong recommendations that are not clearly consistent with GRADE guidance

Among the strong recommendations based on low or very low confidence estimates, we will select those for the Phase III study as follows:

i.) With the completion of Phase II we will have a catalogue of recommendations in which reviewers agreed that strong recommendations were inconsistent with GRADE guidance and were characterized as best practice, misclassification, or mistaken recommendation, and were in the four content areas of maternal health, child health, HIV/AIDS, and TB;

ii.) We will select three guidelines from each of the four topic categories (for a total of 12 guidelines). We will seek guidelines that include at least one best practice, one misclassification, and a mistaken strong recommendation. If there are more than three guidelines in each category that meet this criterion, we will randomly select three guidelines from among the total.

If there are fewer than three guidelines that meet this criterion, we will include any that do and select the guidelines that have two of the three categories of strong recommendations based on low confidence that are inconstant with GRADE guidance. If there are more guidelines with two of the three categories than are required, we will then select at random from among those potentially eligible.

In the event that a panel chair of a selected guideline cannot be contacted or declines participation, we will interview the guideline chair who has agreed (if there is one). We will also select, on the basis of the strategy described above, another eligible guideline that most closely matches the content area of the guideline in which only one

chair was available. If this happens in two guidelines, we will seek a replacement for each. We will, however, conduct interviews for a maximum of 12 guidelines and will be guided by the ability to secure contact information.

In addition to interviewing the two guideline chairs of each panel, we will also interview the WHO lead technical officer (RTO) who worked with the panel.

Interviews

We will interview the panel chairs from the selected guidelines (one methods chair and one content chair from each guideline) and the RTO (also known as the LTO or lead technical officer). The panel chairs/leads were chosen given that they may be best positioned to give their own views/role/decisions, of the overall panel process, and potentially that of the panelists. It would not be feasible to interview all panelists. RTOs may provide additional insights into the guideline process and the reasons for panellists' decisions, as well as factors outside of the panel that may have influenced the final text of the recommendations.

We will therefore conduct 36 interviews, each of approximately one hour and fifteen minutes duration (75 minutes), for each of the 12 selected guidelines. We believe that since we are not seeking in-depth interpretation of the responses, then approximately three to four minutes per question and one hour in total for the interview phase should be sufficient. This plan may be modified depending on the experience with the first interviews. Ten to fifteen minutes at the start of each interview is set aside for introductions and interview house-keeping matters.

We will ask each panel chair and the RTO questions to elucidate the process by which they arrived at the strong recommendations that we deem is a best practice, misclassification, or mistaken recommendation. The designated interviewees will receive the guideline and interview questions electronically via e-mail two weeks before the scheduled interview, with an option to also receive hard copy by mail. This will give the panel chairs and RTO an opportunity to review the guideline and interview questions. This will allow suitable preparation and reflection for the interview and optimal use of interview time.

Study manoeuvres

Qualitative Research approach

The qualitative descriptive methodological approach (QD)[10-12] will be employed to address our research question. The QD approach uses low-level inference, and allows for staying closer to the surface of the data in order to optimize understanding of the panels' decision-making processes during recommendation development. The QD approach allows us to fully describe the responses of the panel chairs and RTOs in their own words and does not require interpretation of the responses.[13] The QD approach is optimal given our aim is to acquire direct descriptions of the phenomenon as the participants lived and experienced it, as well as their reflections and observations at a distance from the experience.

Recruitment

How will the panel chairs and RTOs be contacted for participation?

To contact panel chairs and RTOs, our initial action will be to refer to the 12 selected guideline (s) which should list the names of the panelists, the institutions they represent, and the contact information within the document. We have found that this information is not readily available in most guidelines. The WHO Secretariat supports the Guideline Review Committee (GRC) and thus is involved at all stages of guideline development. We will consult the Secretariat to obtain the contact information (e.g. e-mails) of the relevant guideline chairs and RTOs. The contact will be through the RTOs (technical officers listed as part of the WHO guideline steering group) who can assist us in gathering the contact information of the interviewees.

Once the RTO is contacted and affirms the contact information, each nominated panel chair and RTO per guideline will receive an invitation letter via e-mail. The letter will explain the background and aim of the project and that the project is being conducted at the behest of the WHO GRC. The letter will specify that there will be no financial remuneration for participation. The letter will include an attached consent form (see Appendix 1). If they agree to participate, the chairs and RTO will register their consent and attach the consent document via PDF, in a return email to the lead researcher. Panel chairs and RTOs will also have the opportunity to ask questions regarding the project which will be answered in a return email (s). Once we have obtained consent we will

arrange interview times to occur in the subsequent two months. Panel chairs and RTOs per guideline who agreed to participate will be sent an e-mail reminder of the interview three weeks before the scheduled interview.

The interview with panel chairs

The initial options we considered for the meeting included face-to-face and Skype (or similar formats). We rejected face to face interviews for reasons of feasibility and Skype because of concerns regarding technical problems. We also considered Go-to-Meeting which provides a much more stable connection platform than Skype but incurs a financial cost to operate and is not as yet a mainstreamed approach. Ultimately, we chose to conduct the interviews by telephone. The telephone format allows a certain level of anonymity which may foster more forthcoming introspection and reflection in the respondents.

As prior mentioned, each interview with each panel chair and RTO will have a 5-10 minute house- keeping preparatory phase where the interview will be explained more fully and a 60 minute question and responding phase, and comprise approximately fifteen to twenty questions regarding their thoughts on the issues relevant to the guideline and recommendation development.

The interviewer will make sure that all questions and concerns are answered and addressed to the participants' satisfaction prior to starting the interview. The interview will be considered complete once the participant indicates that they have completed their responding and requires no further clarification surrounding the questions and/or content

of the interview.[14] At the end of the interview session, the participants will be asked whether they could be approached again for validation of the data as soon as essential descriptions and themes are synthesized.[15] This form of data rechecking or member checking[15] as it is known, is an integral part of qualitative research and is used to enhance study rigor and credibility. The process involves a check with key informants (participants) typically toward the end of the study to ensure that the final presentation of the interview data correctly translates and reflects the true experience and viewpoints of the participant (s).

The interview questions (listed below) will be finalized by the project team and the final set of questions will be delivered to the panel chairs in a standardized format and via two interviewers (the project lead and an additional person with experience in delivering interviews). This will function to reduce variability/biases.

Publication plan

Following qualitative and descriptive analysis of the interviews, the plan is to publish the findings within a peer-reviewed journal, thereby making the manuscript results publicly available. Input and interpretation from the full research team will be sought for this. Upon request and strictly for purposes of guideline development improvements, the WHO Guidelines Secretariat would be provided additional basic descriptive raw data in bulk format used to underpin the manuscript.

While the interviews will be based on existing recommendations and the results of the prior Phase II taxonomy exercise (whereby emerging examples of optimal and suboptimal strong recommendations based on low or very low confidence in effect estimates will be displayed in the published manuscript), the intent is to mitigate any blame or trace-back to interviewees. The lead researcher or any researcher as part of this study, will not identify the interviewees by name in any manner, and in any reports using information obtained from the interviews, and that confidentiality as a participant in this study will remain secure. However, given that the guidelines and recommendations already exist within the public domain, in some instances, the reader of the intended published manuscript, may be able to link interview comments to a specific person. This is a real possibility given that the recommendations will be referred to and listed (and these recommendations will have background referenced guideline (s)), and thus possibly traced to the originating guideline and by extension, the interviewee (chair and possibly co-chair).

This listing and linkage is felt to be central to the study and to the readers' understanding of the research. The cohort is only 12 guidelines and the researchers feel that without the context of the specific guideline and recommendation, comments from the interviews will not be meaningful.

Subsequent use of records and data will be subject to standard data use policies which function to protect the confidentiality and anonymity of individuals and

institutions. WHO leadership (e.g. Directors/Managers/Supervisors of interviewees), any staff reporting to interviewees, and any lateral WHO staff, will neither be present at the interview nor have access to raw notes or transcripts. This precaution will prevent individual comments from having any possible negative repercussions. No one (the exception being the WHO Guidelines Secretariat) outside of the research team (i.e. lead researcher, assistant, and all team members) will have access to the interviewee's raw and individual responses and no one as part of the research team, will share responses with anyone without the interviewee's written permission.

Interview guide and questionnaire development

The interview guide and proposed questions will seek to ask open-ended questions in a semi-structured format. This format encourages participants to respond in their own words. Initial questions will be followed by more probing questions in order to obtain as complete understanding as possible.

WHO GRADE guidelines project Phase III-Interview Guide

Begin the interview introduction and set up:

Lead researcher: "Good day, and thank you for agreeing to participate in this WHO GRADE guidelines project interview. My name is Paul Alexander and I am a PhD student working at McMaster University, Hamilton, Ontario, Canada. I am part of the team of researchers conducting this qualitative study in partnership with WHO. We appreciate that you have taken the time to participate and prepare for the interview.

As part of this study, your responses are being audio-recorded so that we may transcribe them following the interview to allow for detailed qualitative analysis. The study will be fully anonymized and the recording will be destroyed after it is transcribed. Is it ok if we audio record you're the interview?

As you know, the purpose of this interview is to understand your experience, and those of your fellow panelists, in guideline and recommendation development. We are exploring specifically when strong recommendations are made on the basis of low or very low study quality. We are interested in the thought processes you and your panel went through in making such recommendations.

I am going to ask you some general questions about the guideline and recommendation development process in which you were involved. We will then focus on instances in which your panel made strong recommendations on the basis of low or very low study quality. We recognize that there may be factors that WHO guideline developers take into account, such as human rights, the political situation on the ground, accessibility, feasibility, and so on that may not be explicitly written in the guideline report. So, we are trying to fully understand how the guideline developers go from evidence to recommendation.

Do you have any questions for me/us before we proceed and do we have your permission to audio-record your interview responses?"

Allow the panel chairs time to reflect and respond and then proceed.

Begin interview questions here (only if the interviewee responds "yes")

"Let us begin:

1.) Briefly describe your public health/academic/research/clinical background as well as the general areas of focus within your regular work life.

2.) Can you tell me about your understanding in general, of the GRADE methods and how they are utilized within guideline development e.g. in terms of strength of recommendations and study quality? What do you think are its strengths and challenges of GRADE?

3.) As the lead or chair, how did the panelists in your group view GRADE and the use of GRADE as part of guideline development? Is it your sense that the panel members found GRADE application challenging?

4.) What do you think your panel members understood by the term "quality of evidence"? Or to use another phrase, "how good the evidence was"?

5.) What do you think your panel members understood as the significance or meaning of a strong recommendation?

6.) What do you think your panel members understood as the significance or meaning of a weak/conditional recommendation?

7.) Did you get the impression that your panel was required to apply GRADE to all your recommendations? (If answer is no:) What was your understanding of the circumstances under which you could make recommendations and not apply GRADE?

Now I would like to understand your thinking and your impression or conception of what your panelists were thinking - their rationale- when they made certain strong recommendations based on low or very low confidence.

The interviewer will ask question 8 for each of the relevant recommendations.

8.) Dr. Sunday, I'd now like to focus on a specific recommendation you made, a strong recommendation made with low quality evidence. The recommendation was for infants who were able to breast feed should be put to the breast as soon as possible when clinically stable. I'll be asking some specific questions, but can we start off with your thoughts about your panel's rationale for this strong recommendation despite the low quality evidence?

I'd then give him a chance to talk and you may have wanted to follow up from what he said, but I'd expect at some point I'd ask:

"Dr. Sunday, what do you think the alternative here is to starting breast feeding as soon as possible? Is it using the bottle for a temporary period?

- and then after the answer –

What do you think that temporary period is as the comparison? One day? One week? Or....?;

And then subsequently:

We seem to have high quality evidence of benefit of breast feeding versus no breast feeding, a 42% reduction in mortality. What do you think of the argument as follows: if breast feeding versus no breast feeding is beneficial with high quality evidence, then we

have moderate confidence – that is, moderate quality evidence - that starting early is better than starting late. What do you think of that logic – makes sense, or not?

Why we are interested in your view here is because there are four criteria which is the basis for the process of going from evidence to recommendation (determining recommendation strength). Do you think the panelists used these four in their thinking and deciding? In other words were you and they aware of the four criteria to determine strength? e.g. quality of evidence, the balance between the desirable and undesirable out comes (benefits versus risks), the variability in values and preferences, and costs (resource use, feasibility, burden etc.) ".

9.) What criterion then did your panel consider in moving from evidence to recommendation (to determine strength)? If they used the four and fully described, go to Q 10.

10.) Were there any modifications to your recommendation (s) that were made after your panel's meetings? If so, what were the modifications? Can you tell us the reasons for the modification.

11.) Even though the quality of evidence was low, you probably thought that the issuing this recommendation would do more good than harm? Can you elaborate?

12.) It is possible that WHO panelists may make strong recommendations in some instances when the quality of evidence is low because if they make a weak/conditional

recommendation policy makers and funders are likely to ignore the recommendation. What are your thoughts on this?

13.) It is possible that WHO panelists may make strong recommendations in some instances when the quality of evidence is low because a practice has become very well established. What are your thoughts on this?

14.)"Are you aware of the available training videos that panelists can review ahead of time (before the guideline development begins) and were these reviewed by you? Do you think the panel members knew of these videos and reviewed them? Do you think there is a need for GRADE orientation prior to participation in the development of a guideline?"

Inform the panel chair that the interview is coming to a close.

"Our interview has come to an end. Do you have any other comments you would like to make at this time?"

Allow the panel chairs time to respond and then proceed.

"Are there any other questions I should be asking that I may have missed?"

Allow the panel chairs time to respond and then proceed.

"Can we approach you again for validation of the data as soon as essential descriptions and themes are synthesized?"

Allow the panel chairs time to respond and then conclude the interview.

"You have been very helpful and thank you very much for participating."

The interview is ended at this point.

Data collection phase

The interview responses will be audio-recorded via the use of the speaker phone feature and two tape recorders (one will serve as a back-up) on the lead researcher's side of the interview. There will also be a note taker who will document the interview responses. In effect, the interview responses will be collected via three avenues. This process will be fully disclosed to the interviewee. The lead researcher's assistant will be responsible (with the lead researcher) for setting up the interview environment, including the phone and recording equipment. The lead researcher will be located in a private office at McMaster University in Hamilton, Ontario, Canada for the conduct of each interview, as well as the transcription phases. This will ensure a sterile and standard environment. Each interviewee will likely be internationally situated for the interview call and all efforts will be made to schedule the interview at a time (factoring in time zone differences) convenient to the interviewee that will allow at least 75 minutes of undisturbed time.

The WHO guideline Secretariat will provide some funding to help defray the costs of the interview transcription phase and recording equipment.

Data management

Our only source of data/information will be the interviews and thus triangulation approaches[16] (multiple data sources) will not be used. Interviews will be professionally transcribed and analyzed (details on analysis discussed below) as data are being collected from the 24 panel chairs and 12 RTOs. A codebreaker will be prepared by the transcriber to de-identify each participant and affiliations/names that were mentioned during each interview. The hand written interview notes as well as the extra interview tape recordings

(from speaker phone) will be stored in a double locking system (locked file cabinet in a locked office) and only be accessed should the primary tape recording method fail. These additional sources of interview data will be destroyed immediately upon study completion when the analysis has been completed. The lead researcher will be responsible for the interview data security and storage and will handle the data (until fully destroyed at study termination) with the highest level of control and confidentiality.

To ensure that an iterative process is in place for data collection and analysis, data will be analyzed as interviews are conducted and transcribed.[17] This will allow for the refinement or addition of interview questions, as the process unfolds. The transcribed interviews (via 36 separate, password-protected documents) will be secured by the transcriber and provided to the lead researcher on a rolling basis. The project lead will listen to the recorded interviews and simultaneously read the transcribed information, in order to assess whether the transcribed information accurately reflects what was recorded. If the lead researcher deems that what is transcribed does not match what was recorded, then the lead researcher and the transcriber will hold a discussion to address the discrepancy. A third party adjudicator will be used if resolution is not reached.

Data analysis

Once agreement is reached that the transcribed information matches what was recorded, the interview information will be imported to the qualitative research software NVivo 10[18] in order to carry out the next phases of the qualitative descriptive study. This program is used to organize the codes (essentially labels that are assigned to segments of

text in order to provide meaning) for large amounts of data, establish an audit trail, and compute intercoder agreement between coders.

Inductive, conventional content analysis will be used to analyze the qualitative data. Content analysis is the method of choice because it allows for us to gain direct information from the panel chairs without imposing preconceived categories or theoretical perspectives. (Hsieh & Shannon, 2005 - Three Approaches to Qualitative Content Analysis).

Coding will be carried out by two researchers (the lead researcher and another member of the research team), each working independently on a subset of transcripts. The coding occurs as interview information accumulates. During qualitative analysis, the coder will read the transcripts repeatedly to become familiar and immersed in the data, highlighting important quotes and taking note of salient concepts. Next, the interview data will be analysed to gather codes, themes, and commonalities. While the aim is to derive no more than 30 codes and an overall six themes within the qualitative research approach, this may vary for this project given the nature of the research question. The coders will pay special attention to identifying definitions, events, perspectives and processes uncovered in the data. Terms or phrases used repeatedly by the panel chairs will be coded in their own words. Any notes taken (from the field) will also be used to help coders gain insight into the phenomena under study.

Regular meetings will be held between the two coders as they proceed with coding the subset of transcripts. The goals of each meeting will include i) a discussion of potential discrepancies in codes, ii) sharing of interpretations and understanding of the phenomena being studied so as to generate newer and richer codes, and iii) intercoder agreement computation to assess reliability. This process enhances the dependability of the findings. Together, the coders will develop a codebook that consists of definitions for each code with sample segments from the transcripts. The codebook will evolve until firm agreement is reached (e.g. if the theme is loyalty, then loyalty will mean the same for both coders). Reliability of the coded data will be assessed using intercoder agreement measures (Kappa) and the acceptable agreement threshold for this study will be 0.70.[19]

Once substantial agreement has been reached between the coders, the lead researcher will proceed with coding the subsequent transcripts independently. As soon as codes have been generated, they will be sorted into categories based on their relationship with each other. Memoing (recorded trail of ideas about codes and their relationships as they occur to the coder during analysis) will be used in order to sort the codes into meaningful categories. Subsequently, codes and categories will be compared and contrasted until important themes emerge from the data. From there, the lead researcher will prepare the results and interpret the findings in the final reporting, working at all steps in close liaison with the full research team. The lead researcher will ensure that a person skilled in qualitative research is a member of the Phase III team and in order to

provide guidance and support at all stages of the interview, data collection, and analysis phases.

Ethics and confidentiality

There will be minimal or no risk to our participants. We will take particular measures to safeguard against privacy breach. Ethical review board (ERC) approval from WHO will be sought. The consent form will be e-mailed to the panel chairs and RTOs via which they can accept to participate and we will inform the panelists of their right of refusal and withdrawal from the study/interview at any time.

The audio recordings will be destroyed immediately following transcription of the interviews (and independent verification). While the names of the panel chairs and RTOs may be publicly available within existing WHO documents, we will seek permission from each chair and RTO for the use of the interview information outside of the study aims, and will seek to maintain confidentiality and anonymity for the interviewees (anonymized documents whereby limited or no personal identifiers will be used via the use of a coding system). Only the lead researcher will maintain the traceability between the identification codes and the interviewee (under lock and key) and this will only be done in case there is a need to return to an interviewee for response clarifications. All study information will be managed under lock and key by the project lead and information managed and manipulated within computer software programs will be password protected. Upon study completion, the identification codes will be destroyed.

Limitations and strengths

One limitation surrounds not conducting the study via in-person interviews. As mentioned, logistics prevents this and thus our decision is to conduct via telephone. This telephone format allows a level of anonymity that can function as a benefit regarding complete responding. Another limitation is that we are not interviewing all panel chairs and RTOs and basing our study on 12 selected guidelines. However, we are focusing on the 12 guidelines that report the highest number of strong recommendations based on low or very low confidence. A third limitation may be that we would not be able to contact one or more panel chairs or RTOs. Should this occur, then we will re-examine the topic category in question and select another guideline that reported strong recommendations based on low or very low confidence in estimates. Research team discussions will seek to devise the best next step so as to gather the most valuable information.

The key strength of the proposed study is that it will provide unique insight into the processes that led to the strong recommendations based on apparent low confidence. Additional strengths come from the use of strategies to increase rigor, confidence, and credibility within this qualitative research design (e.g. use of field notes, audit trails (memoing), intercoder agreement, and member checking). Further, qualitative description anticipates no more than 20-30 participants to reach information redundancy and data saturation (stage where no new themes or categories are likely to emerge from the data). We therefore anticipate that our sample size of 36 is more than ample and will provide

sufficient information to allow for adequate description of the panel chairs' and panelists' thoughts (those involved in the recommendation (s)) and decision making processes.

Implications

We have documented in Phase I that strong recommendations based on low or very low confidence in effect estimates are very frequently made in WHO guidelines. Our research will provide insights into the reasons why panelists make strong recommendations when there is high uncertainty in confidence. Our findings may increase our understanding of how guidelines and recommendations are being interpreted by guideline panelists and thus areas for improvement.

WHO guidelines are integral to public health practice globally and they inform and shape local, regional, as well as global public health policy. It is imperative therefore that this study be conducted and results be disseminated to guideline developers, both at WHO as well as globally.

Proposed Project Team

Paul E Alexander, Susan Norris, Lisa Bero, Victor M. Montori, Juan Pablo Brito, Rebecca Stoltzfus, Benjamin Djulbegovic, Ignacio Neumann, Gordon Guyatt, Shelly-Anne Li

References

1. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World health organization guideline development: an evaluation. ***PLoS One.*** **2013; 8(5):e63715. doi: 10.1371/journal.pone.0063715. Print 2013.**

2. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet.*** **2007 ;369(9576):1883-9.**

3. GRADE working group. Grading the quality of evidence and the strength of recommendations.
url: http://www.gradeworkinggroup.org/intro.htm (Accessed March 10th 2013).

4. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. ***Journal of Clinical Epidemiology.*** **Volume 64, issue 4, 383-394, April 2011. Guyatt et al. (2011).** url: http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract (Accessed March 10th 2013).

5. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008; 336(7650):924-6.**

6. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.

7. WHO. WHO guidelines approved by the Guidelines Review Committee. url: http://www.who.int/publications/guidelines/en/index.html (Accessed on April 24th 2013).

8.) WHO. Emergency response to antimalarial drug resistance. url: http://www.who.int/en/ (Accessed on April 24th 2013).

9.) Brito JP, Domecq JP. Murad MH., Guyatt GH, Montori VM. The Endocrine Society Guidelines: when the confidence cart goes before the evidence horse. Submitted to ***Journal of Clinical Endocrinology and Metabolism*** 03/15/2013.

10.) Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health, 23*(4), 334-340.

11.) Sandelowski, M. (2001). Real qualitative researchers do not count: The use of numbers in qualitative research. *Research in Nursing & Health, 24*(3), 230-240.

12.) Walker JL. The use of saturation in qualitative research. ***Can J Cardiovasc Nurs.*** **2012 Spring; 22(2):37-46.**

13.) Thorne, S., Reimer Kirkham, S., & O'Flynn-Magee, K. (2004). The analytic challenge in interpretive description. *International Journal of Qualitative Methods, 3*(1). Article 1. Retrieved August 21st 2013 from http://www.ualberta.ca/~iiqm/backissues/3_1/pdf/thorneetal.pdf.

14.) Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: SAGE Publications, Inc.

15.) Lincoln, Y. S, & Guba, E. A. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

16.) Denzin, N. K. (1970). The Research Act in Sociology. Chicago: Aldine.

17.) Miles, M. B., & Huberman, A. M. (1994). An expanded sourcebook: Qualitative data analysis (2nd ed.). Thousand Oaks, CA: SAGE.

18.) QSR International. NVivo 10. url: http://www.qsrinternational.com/products_nvivo.aspx (Accessed on August 19th 2013).

19.) Landis JR, Koch GG. The measurement of observer agreement for categorical data. ***Biometrics.*** **1977; 33(1):159-74.**

Appendix 1: WHO GRADE Guidelines Project Phase III

Consent to participate form

You are being invited to participate in the WHO GRADE Guidelines Project PHASE III given your role as a panel chair/lead/lead technical officer in the development of specific WHO guidelines. The PHASE III is a continuation of PHASES I and II of this GRADE related project, whereby we are seeking to describe the distributions of WHO recommendations, and particularly to gain a better understanding of the factors and thoughts that go into panelists making strong recommendations. PHASE III is the interview portion of the project whereby panel chairs (content and/or methods) and RTOs will be asked to describe their thoughts and considerations in making strong recommendations based on high uncertainty, especially those that we think are more akin to being i.) best practices ii.) misclassifications (rated as low/very low but should be moderate or high confidence) and iii.) strong recommendations that are clearly inconsistent with GRADE guidance (should have been weak/conditional).

We are asking your consideration to participate and if you agree, please read and sign this consent form. Once you have signed in agreement, please e-mail a pdf version to the lead researcher (Paul E. Alexander, e-mail: elias98_99@yahoo.com; or pauleliasalexander@gmail.com). You are also invited to communicate with the lead researcher at any time leading up to the scheduled interview in order to help address any questions you may have. You may also communicate with the lead researcher post interview, should any questions arise. Please note that should you not wish to continue participation after you have consented, you are free to withdraw at any time and without prejudice to you. All interview information (written or recorded) will be destroyed immediately as analysis is completed and no records in any format, will be kept by anyone connected to the study. Names of all interviewees (whether listed openly in public guideline documents or not), will be held securely and all efforts to maintain confidentiality and anonymity will be undertaken, during and following the study. Once the signed consent form is received by e-mail, the lead researcher will sign and date the form and provide you with a copy by pdf return e-mail.

Consent for Participation in WHO GRADE Guidelines Project Phase III Interview

I volunteer to participate in the WHO GRADE Guidelines Project Phase III Interview conducted by Paul Alexander (doctoral student) from McMaster University, Hamilton, Ontario Canada. I understand that the PHASE III interview is designed to gather information about WHO strong recommendations based on high uncertainty and particularly those that the research team considers were strong recommendations that were not clearly consistent with GRADE guidance, as well as were consistent.

1. I will be one of approximately 36 people being interviewed for this research. My participation in this project is voluntary.

2. I understand that I will not be paid for my participation. I may withdraw and discontinue participation at any time without penalty. If I decline to participate further or withdraw from the study, no one other than the lead researcher and I, will know of this.

3. I understand that most interviewees will find the interview interesting, thought-provoking, and informative. If, however, at any time during the interview, I feel uncomfortable in any way, I have the right to decline to answer any question or to end the interview. I understand that all my prior responses if this occurs, will be destroyed immediately.

4. Participation involves being interviewed by Paul Alexander of McMaster University, Hamilton, Ontario, Canada. The interview will last approximately 60 minutes (at most 16 questions) with an additional 10-15 minute preparatory phase. During this prep phase, I can ask any clarification questions that would bear on my further participation. I understand that prior to the interview, my signed consent must be already provided to the lead researcher of this project.

5. Notes will be written during the interview and will be taken by an assistant to the lead researcher. An audio tape of the interview will be made. If I do not want to be taped, I will not be able to participate in the study. The purpose of taping the responses is to allow for a qualitative analysis of the responses.

6. Following qualitative and descriptive analysis of the interviews, the plan is to publish the findings within a peer-reviewed journal, thereby making the manuscript results publicly available. Input and interpretation from the full research team will be sought for this. Upon request and strictly for purposes of guideline development improvements, the WHO Guidelines Secretariat would be provided additional basic descriptive raw data in bulk format used to underpin the manuscript.

While the interviews will be based on existing recommendations and the results of the prior Phase II taxonomy exercise (whereby emerging examples of optimal and suboptimal strong recommendations based on low or very low confidence in effect estimates will be displayed in

the published manuscript), the intent is to mitigate any blame or trace-back to interviewees. The lead researcher or any researcher as part of this study, will not identify the interviewees by name in any manner, and in any reports using information obtained from the interviews, and that confidentiality as a participant in this study will remain secure. However, given that the guidelines and recommendations already exist within the public domain, in some instances, the reader of the intended published manuscript, may be able to link interview comments to a specific person. This is a real possibility given that the recommendations will be referred to and listed (and these recommendations will have background referenced guideline (s)), and thus possibly traced to the originating guideline and by extension, the interviewee (chair and possibly co-chair).

This listing and linkage is felt to be central to the study and to the readers' understanding of the research. The cohort is only 12 guidelines and the researchers feel that without the context of the specific guideline and recommendation, comments from the interviews will not be meaningful.

Subsequent use of records and data will be subject to standard data use policies which function to protect the confidentiality and anonymity of individuals and institutions. WHO leadership (e.g. Directors/Managers/Supervisors of interviewees), any staff reporting to interviewees, and any lateral WHO staff, will neither be present at the interview nor have access to raw notes or transcripts. This precaution will prevent individual comments from having any possible negative repercussions. No one (the exception being the WHO Guidelines Secretariat) outside of the research team (i.e. lead researcher, assistant, and all team members) will have access to the interviewee's raw and individual responses and no one as part of the research team, will share responses with anyone without the interviewee's written permission.

7. I understand that this research study has been reviewed and approved by the WHO Ethics Review Board (ERC) as well as the relevant ERB from McMaster University (e.g. studies involving human subjects).

8. I have read and understand the explanations/information provided to me regarding this project. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. If I have any additional questions, I will communicate them to the lead researcher via e-mail.

9. I understand that on signing, I will be given a signed copy of this consent form.

_____ _____

My Name, Signature, and Date

_____ _____

Name and Signature of the Investigator (and date)

For further information, please contact:

Mr. Paul Alexander, Principle Researcher or Dr. Gordon Guyatt,

elias98_99@yahoo.com, pauleliasalexander@gmail.com, guyatt@mcmaster.ca

Authors' contributions:

Mr. Alexander contributed principally to the conception of the chapter protocol study with guidance from Dr. Guyatt and the larger research team.

CHAPTER 6: Phase III:

WHO guideline panelist experience with GRADE methods when making strong recommendations based on low or very low confidence in effect estimates: A qualitative descriptive study

This qualitative interview study builds on the Phase I (Chapter three) and Phase II (Chapter four) studies and seeks to ask WHO guideline panelists involved in the uncovered recommendations, to explain (in their own words) the factors they take into consideration in arriving at such recommendations. We judged (via Phase II) that a significant portion of WHO's strong recommendations based on low or very low confidence in effect estimates are inconsistent with GRADE guidance. We thus needed to understand, via one-on-one interviews, why WHO guideline panels were making these recommendations. In our approach, we were trying to assess whether panels were correct in arriving at such recommendations (in other words our taxonomy judgements were incorrect), or were misapplying GRADE.

[1]Paul E Alexander, [2]Michael R. Gionfriddo, [1]Shelly-Anne Li, [3]Lisa Bero, [4]Rebecca J Stoltzfus, [1,5]Ignacio Neumann, [6]Juan P. Brito, [7]Benjamin Djulbegovic, [8]Victor M. Montori, [9]Susan L. Norris, [10]Holger J Schünemann, [11]Gordon H Guyatt

[1] Health Research Methods (HRM)
Department of Clinical Epidemiology and Biostatistics
McMaster University, 1280 Main Street West,
Hamilton, Ontario, L8N 3Z5, Canada.

[2]Mayo Graduate School, Mayo Clinic
Doctoral student
Knowledge and Evaluation Research Unit, Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[3]Professor (Chair of Medicines Use and Health Outcomes)
The University of Sydney
Charles Perkins Centre
Camperdown NSW 2006

[4]Professor
Division of Nutritional Sciences
120 Savage Hall
Cornell University
Ithaca NY 14853, USA

[5]Internal medicine specialist
Department of Internal Medicine
Pontificia Universidad Católica de Chile
Santiago, Chile

[6]Assistant Professor of Medicine
Investigator, Knowledge and Evaluation Research Unit
Divisions of Endocrinology, Diabetes, Metabolism and Nutrition
Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[7]Distinguished Professor
University of South Florida
H Lee Moffitt Cancer Center
Florida, USA

[8]Professor of Medicine
Knowledge and Evaluation Research Unit Divisions of Endocrinology and Diabetes and Health Care and Policy Research
Mayo Clinic
Rochester, Minnesota
Plummer 4-402, 200 First Street SW, Rochester
MN 55905, USA

[9] Guidelines Review Committee Secretariat
World Health Organization
Av. Appia 20
CH-1211 Geneva 27
Switzerland

[10]Professor and Chair
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
Michael Gent Chair
in Healthcare Research
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON L8S 4K1, Canada

[11] Distinguished Professor
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
McMaster University Health Sciences Centre
1200 Main Street West, Room 2C12
Hamilton, Ontario, L8S 4K1, Canada.

Corresponding author: Paul E. Alexander, doctoral candidate.

E-mail: elias98_99@yahoo.com

This chapter has been submitted for peer-review and publication in the JCE journal, January-March 2015 edition. Communication with the JCE journal editors have indicated that they will accept a submission for peer-review. WHO has cleared the manuscript.

Abstract

Background

Many strong recommendations issued by the World Health Organization (WHO) are based on evidence for which there is low or very low confidence in the estimates of effect (discordant recommendations). Many such recommendations are inconsistent with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidance that suggests that discordant recommendations are often inappropriate.

Objective

To gain insight into the process of making recommendations using GRADE and explore the process resulting in strong recommendations based on low or very low confidence (quality of evidence) issued by the World Health Organization (WHO).

Data sources

Panel members who were involved in the development of guidelines approved by the WHO Guideline Review Committee between 2007 and 2012 and included discordant recommendations.

Methods

Thirteen panel members participated in semi-structured interviews focusing on the use of GRADE and the reasoning behind, and factors contributing to, discordant recommendations. We recorded and transcribed interviews, and used inductive content analysis to derive codes, categories, and emergent themes.

Results

Four themes emerged: i) strengths of the GRADE approach, ii) challenges and barriers to GRADE application, iii) strategies to improve the use of GRADE, and iv) explanations for discordant recommendations. Reasons for discordant recommendations included skepticism about the value of making conditional recommendations; political considerations such as meeting the needs of ministries of health; a high certainty in benefits (sometimes warranted, sometimes not) despite having assessed the evidence as low confidence; a reluctance to make conditional recommendations for long-standing accepted practices; and concerns that conditional recommendations will be ignored.

Conclusions

GRADE provides a framework to support the development of trustworthy guidelines. WHO panelists are making discordant recommendations inconsistent with GRADE guidance for reasons that include a reluctance to make conditional recommendations and limitations in understanding of GRADE. Our findings suggest GRADE is being sub-optimally implemented. In order to reach its full potential, at WHO and elsewhere, is likely to require selecting panelists with a commitment to GRADE principles, and

additional training of panelists.

Background

The World Health Organization (WHO) produces and disseminates guidelines that play a role in informing global public health policy, particularly for low and middle income nations. In 2007, WHO created the Guideline Review Committee (GRC) in response to research indicating that guideline panels at WHO were sub-optimally using systematic reviews in their development process.[1,2] The creation of the GRC renewed a focus on the Grading of Recommendations Assessment, Development and Evaluation (GRADE)[1,2-4] approach to guideline development that WHO had begun to use in 2003. Presently, along with WHO and the Cochrane Collaboration, more than 80 other organizations are using the GRADE approach.

The GRADE approach facilitates judgements about the confidence in effect estimates (also known as study quality or certainty) of potential benefits and harms and offers guidance on grading the strength of recommendations. Confidence is rated as high, moderate, low, or very low. Strength is graded as strong or conditional (also labelled as weak or discretionary) and is determined by the magnitude of benefits and harms, values and preferences, confidence in effect estimates, and resource use considerations.[2-4] For public health guidelines, other factors including the burden of illness, acceptability, accessibility, feasibility, and equity may also play a role in determining the strength of recommendations.

Strong recommendations should generally be based on evidence in which we have high or moderate confidence. GRADE guidance therefore cautions against strong recommendations supported by low or very low confidence (labelled here as discordant recommendations).[3-6] The GRADE working group has, however, identified five circumstances[6] in which discordant recommendations may be appropriate to make (e.g., in a life threatening situation; potential catastrophic harm). [6]

A study of WHO guidelines that used GRADE and were published from 2007 to 2012[7] found that among 456 recommendations, 63.4% were strong and 36.6% were conditional/weak. Of the strong recommendations, 55.5% (n=160) were discordant. Of the 160 discordant recommendations, 25 (15.6%) were consistent with one of the five paradigmatic situations where discordant recommendations are deemed appropriate.[8] Of the remaining 135 recommendations, 33 (20.6% of the 160) were judged as a misclassification of confidence in effect estimates[8] (the quality of the body of evidence for the outcome is actually at least moderate), 29 (18%) as good practice statements,[8,9] in which an abundance of indirect but difficult to summarize evidence established that benefits are far greater than any possible harms, and 73 (45.6%) that would be appropriate as weak or conditional rather than strong recommendations.[8]

Thus, WHO guideline panels are frequently making discordant recommendations, with many deviating from GRADE guidance. These results are consistent both with anecdotal impressions and with empirical data from guideline published by the Endocrine Society.[10] Elucidating explanations for discordant recommendations may provide

insights relevant to not only WHO, but also other guideline panels. These insights may inform ongoing improvement of the GRADE process and efforts to improve and facilitate the uptake and implementation of GRADE. To acquire these insights we undertook a qualitative study with individuals involved with the development of guidelines at WHO to better understand the process and reasons underling discordant recommendations.

Methods

Qualitative research approach

We employed a qualitative descriptive methodological approach[11-15] that facilitates low-level inference and allows for staying close to the surface of the data. This approach generates a thorough description of the responses of interviewees. Previous experience has demonstrated that the qualitative descriptive approach can be useful in identifying critical information for refining existing guidelines and for program development.[16,17] The Ethics Review Boards at WHO, Geneva, Switzerland, and McMaster University, Hamilton, Ontario, Canada approved the study.

Sampling approach

We selected 12 guidelines (three from each topic area: maternal health, child health, tuberculosis, and HIV/AIDS) in which discordant recommendations were most frequent,[8] and sought assistance from the WHO GRC Secretariat to establish contact information on potential respondents.

Purposeful criterion sampling[18] guided the choice of one guideline content chair, one guideline methods chair, and the WHO lead technical officer (who worked with the panel) from each of the 12 selected guidelines, providing a cohort of 36 potential interviewees. We secured contact information for only 22 potentially eligible panel members. An e-mail invitation along with four reminders were sent to the 22 individuals, of whom 13 agreed to be interviewed.

Data collection

Once contact information was established with the help of the GRC, the research team then initiated e-mail contact and individuals who agreed to participate received the guideline and specific recommendations of interest via e-mail two weeks before the scheduled interview.

We collected data using individual, semi-structured audio-recorded interviews consisting of 14 questions. Interviews lasted approximately 60 minutes (Appendix A). We pilot tested the interview guide with an individual who has served as a WHO guideline panel member. The interview questions surrounded but were not limited to: i) understanding of guideline development and use of GRADE methods in guideline development, ii) understanding of how panel members viewed GRADE, iii) understanding of specific GRADE principles and iv) a focus on specific discordant recommendations in terms of factors panelists may have considered in arriving at such recommendations. For each question, interviewees provided their own understanding as well as that of the general panel specific to the discordant recommendation in question.

Analysis of interviews

Though interviewees understood that the examined guidelines were part of the public domain, a code-breaker was prepared by the independent transcriber to de-identify each participant and any affiliations/names that were mentioned during each interview.

Interviews were transcribed verbatim and analyzed as the interviews were conducted. We stored the hand written interview notes as well as the audio recordings in a locked file cabinet and password-protected laptop. Upon transcription and analysis, audio recordings were destroyed.

The qualitative descriptive approach does not specify a particular type of data analytic strategy. Because no previous studies explored the thinking and rationale of guideline panelists making discordant recommendations, we used inductive, conventional content analysis.[19] We did not explicitly impose preconceived theoretical perspectives on the data and findings were restricted to the information provided. The research team did not interpret any transcribed interview information, and all quotes are the verbatim words of the interviewees.

The coders (PEA, SAL) independently read each interview transcript in its entirety thrice. At the start of coding, each coder independently coded a subset of five transcripts to ensure that coders were unified in their approach. During the process, the coders highlighted important portions of raw text and quotes, took note of salient concepts, and finally chose a word or a short phrase to represent the meaning of a specific text segment. The coders met to finalize the preliminary list of codes. Codes were revised and new ones

added as concepts emerged from the remaining eight transcripts. Codes that had similar concepts were grouped together to form categories. Themes emerged from codes and categories.

Coders developed a codebook of definitions for each code with sample segments from the transcripts, beginning with a list of preliminary codes and revising the codebook as new concepts and codes emerged. Saturation was reached by the tenth interview; nevertheless, we interviewed all 13 consenting interviewees. We actively searched for deviant or negative cases/quotes and considered these in reporting the results. To maintain study rigor, we used field notes, audit trails, reflexivity, journal keeping, and assessed intercoder reliability[20] by coding a random subset (n=3) of transcripts in duplicate and independently and measured agreement using the kappa statistic. The kappa was 0.85, suggesting very high agreement.

Results

Of the 13 individuals who agreed to be interviewed 11 were content area expert panel chairs and 2 were WHO technical officers. There were six male and seven female interviewees with a wide variety of research and public health backgrounds. There were no guideline methods chairs as part of the 13.

Of the 16 discordant recommendations that we asked the interviewees to specifically focus on (10 interviewees reflected on one discordant recommendation, and three on two discordant recommendations each), seven recommendations were misclassifications[8] (our judgement was that panelists had rated confidence in effect

estimates down excessively), seven recommendations were inappropriately made strong recommendations that should have been weak,[8] and two recommendations were good practice statements.[8,9]

The interviews revealed four overarching themes (Table 1): i) Strengths of GRADE ii) Challenges/barriers to GRADE application iii) Strategies to support better application of GRADE and iv) Reasons for discordant recommendations.

i) *Strengths of GRADE*

Respondents perceived the use of GRADE as a positive step by the WHO leadership (routinely referred to by respondents as the "WHO GRC Guideline Secretariat") to improve evidence-informed guideline development. Panel chairs recognized GRADE as a highly structured, standardized, evidence-focused guideline development process. One respondent commented "I've felt that there wasn't a sort of robust systematic process with which we had been following so I found GRADE quite useful". Another stated "It is transparent, and it's permitting to use all of the evidence which is available in the literature. It makes the possibility of a fair discussion with the guideline development group on the recommendation that should be done".

One interviewee commented,

> I think the strengths [sic] is that it formalizes a way that the quality of evidence is rated. It has a very systematic way of going about even for actually to have a proper scope and to formalize the question to put it into the research type of wording, guide the reviews which are being done.

Respondents also appreciated GRADE's focus on quantification and transparency, including the formulation of structured questions and emphasis on patient values and preferences. One respondent explained,

> The beauty of the one that's GRADE [sic] is that it is objective, so you can, everybody can see that and can understand that…. so and it's very, it's not only objective but it is also transparent, so nobody, people may not agree with you but everybody knows why you did something.

Respondents acknowledged the growing acceptance and use of GRADE in WHO.

ii) Ch*allenges/barriers to GRADE application*

Respondents spoke to unclear or insufficient GRADE guidance as a challenge or weakness of the overall process. They noted that at times they did not fully understand the GRADE process. Respondents sought further guidance at the start of guideline development in applying the GRADE approach to observational studies, topics with limited evidence, and when indirect evidence was the main source of evidence. One respondent elaborated, "So the type of studies that would be great to support this was not available so therefore what we have in looking at [sic] has been very indirect evidence...It is all descriptive studies. None of them had a logical comparative". In a related reflection, an interviewee commented, "One of the frustrations is that in the end we often have systematic reviews that don't adequately review the key question being considered for the guideline". Another respondent reported, "I think that there is some discomfort or misunderstandings or lack of understanding on when sometimes evidence is upgraded or

downgraded and I think the biggest concern is that in our area really a lot of the data are cohort data or observational data".

Respondents expressed concern regarding the GRADE approach to prioritizing critical versus important outcomes and how to determine the overall certainty of the evidence, as well as regarding the application of GRADE to all questions. For example, one respondent stated,

> WHO has endorsed the GRADE process and therefore recommendations need to all be made within GRADE and have a GRADE as the foundation for the recommendation. But it is my feeling for example that there are some questions that for subsets of questions that are not necessarily suitable to be GRADEd according to the GRADE process.

PICO (patient, intervention, comparator, outcome) question development posed challenges. Respondents experienced PICO questions as inflexible. One interviewee reflected, "It's time consuming because of the PICO questions, which in fact more time is spent on PICO than to have all of the evidence reviewed…PICO takes time because it is already difficult to find the level of detail that you want to address into [sic] the guidelines".

Challenges in considering contextual factors

Respondents indicated that they often face challenges in taking contextual factors into consideration, including accessibility of the intervention, and the costs involved in different countries. One interviewee commented,

> So how do you use the evidence to make the recommendations because the recommendations really involved many other factors, like values and preferences and costs, resources, the impact in terms of public health all that is a bit difficult to understand for the panel?

Power dynamics within the guideline development group

Respondents noted that panel members may subordinate their views to their peers who are more vocal, dominating, and experienced. One respondent shared,

> I've seen on many of these meetings that people who are dominating the discussions are the ones that feel comfortable assessing research evidence and know the GRADE systems because the people who do not, I think it is just my perception, sometimes feel a little bewildered by the whole thing and may stay quiet until you get to the final discussion about it.

Another respondent pointed to the impact on the process when the panel leader is not as experienced or as strong as more dominant panel members, sharing,

> And I think the thing is sometimes I think you know if there is a non-expert or junior person leading the process may find it [sic] very difficult to stand up [sic] the doyen of x y and z and say look that doesn't seem [sic] a reasonable position. And those people often, I have seen very, very strong dominant people who drive processes, who don't necessarily respect the guideline process as it was meant to be designed.

Recommendation decisions made based on age and experience

Interviewees noted that clinical/practical experience sometimes takes precedence over evidence. Some postulated a generational issue as a factor in what panel members would accept as evidence. Younger panel members appreciated properly conducted trials whereas older members were more likely to base their decisions on clinical experience. One interviewee recalled,

> There is certainly imbalance in the way that panel members have capacity to use GRADE methodology and the GRADE logic and coming up with recommendations and the practitioners will base their thinking on, not tested by any research, but my assumption is that the practitioners will base much more of their thinking on their clinical practical public health experiences from the field rather than the literature.

Respondents also expressed a sense of alienation when the methodologist did not consider the background of the panel nor take the time to revisit the GRADE approach when panelists deviated from the proper use of GRADE.

Lack of understanding

Interviewees expressed concerns about some panelists' level of understanding. One noted "I do think that the capacity to interpret findings through the GRADE lens varies a lot depending on what panel members we are talking about". Another stated, "So much of the challenge lies in the way that people understand it, especially in people who are in those guideline committees for the first time. It takes quite a bit for them to get their head around GRADE".

iii) *Strategies to support improved application and implementation of GRADE methods/guidance*

Respondents expressed concern regarding the provision of GRADE training, WHO leadership, and organizational support. They expressed a continued need for enhanced and focused GRADE training at the start of guideline development, particularly for those with little or no prior exposure to GRADE.

They also expressed that GRADE training methods should utilize specific examples with interventions for panelists to consider and use technology for training that is suitably supported by different locations and settings. They agreed that short and concise interactive training videos worked well for training and that there should be delivery of orientation to GRADE before the guideline development meeting begins. For example, one respondent stated "I think people have limited time, they are usually very

busy people with lots of other commitments so you need to be and only use incredibly a well thought [sic] through short videos". Others raised the issue of the amount of time spent on training versus guideline development, sharing "Sometimes we spend even all afternoon just to explain what is GRADE, what are the implications, what are the PICO...I didn't find it an ideal situation because it takes a long time to do that, and when you take this time from the meeting it's always a challenge so that's an issue yes".

There was strong agreement that building capacity around guideline development to promote the use of GRADE and engaging panelists requires leadership by the WHO. Respondents suggested that a strong panel chair and a highly skilled methodologist were necessary to improve the application of GRADE.

Importantly, respondents also agreed that they had failed to make considerations that they used in arriving at discordant recommendations fully explicit in the guideline, suggesting that comprehensive and fully transparent explanations by panelists should follow each discordant recommendation in the published guideline document.

iv) *Explanations for discordant recommendations*

General scepticism about conditional recommendations

One powerful theme that emerged was a general scepticism concerning conditional recommendations. For instance, one interviewee stated that "unless you achieve a strong recommendation you are really saying a conditional recommendation isn't worth the paper it is written on". Some respondents suggested that explaining what conditional means would require extensive time and training. Thus, it would be more

straightforward for panels to label recommendations as strong. One interviewee dismissed conditional recommendations in the context of public health guidelines:

> I think the weak part, contextual or conditional [sic] also doesn't help because every recommendation is contextual and conditional on the availability of resources, availability of people who are trained and so on. So maybe in the context of the guideline, national guideline or in practice it may work, but for international guidelines there are so many different constituencies.

Another interviewee suggested,

> I think weak recommendations is one [sic] which poses problems in almost all of WHO guidelines that use the GRADE system because weak by itself doesn't really have any meaning because when you recommend something you either recommend it for its use or you don't.

Political environment

Respondents stressed that political environments may drive strong recommendations. For example, one respondent commented that "sometimes the political environment, and not the evidence itself may push panelists and staff to make strong recommendations". Another interviewee reflected "but I think sometimes the political environment, the evaluation of the boss of the probability of implementation, in fact the understanding of the difficulties of former colleagues in the Ministry of Health of implementation may push bosses to push panelists to do strong recommendations".

Respondents suggested additional factors related to the political environment including being wedded to long-established practices, and the need for funding and policy formulation, drive discordant recommendations. One respondent stated,

> I certainly think there is the risk of that, as much as we are trying to move towards a more explicit evidence based approach, you are dealing with people who are

> generally practitioners who've got long standing views and there is a resistance to change which is implicit in any guideline process.

Regarding funding and budget concerns as drivers of discordant recommendations, one interviewee pointed out,

> They will not take it [sic] seriously because the term 'weak' and I think the reality is that in front of these policymakers there are so many decisions that they need to make that cost them. They kind of need their budget and if you have a few strong policies or recommendations and a few weak they will just throw out the weak ones.

Respondents also noted resistance from their home communities, ministries of health, and the media. For example, one reported on the reluctance of media to write about weak recommendations, sharing,

> The problem is [a journalist] for instance almost removed the role from the guidelines, because the journalist told me that he was not going to expose himself making weak recommendations about a thing [sic] because people will say that you don't know what you are doing.

Issues related to specific discordant recommendations

In addition to the issues reviewed thus far, other potential explanations for the discordant recommendations emerged from the discussion of particular recommendations including mistaken confidence ratings; minimal harm; longstanding practices; and fear of being ignored.

Mistaken confidence ratings

In some instances, the respondents arrived at low or very low GRADE confidence ratings based on GRADE principles despite in the interview expressing high confidence in estimates of desirable and undesirable consequences.

For example, for the discordant recommendation "Both paracetamol and ibuprofen need to be made available for treatment in the first step", our reviewers judged this as a good practice statement[9] and the interviewee expressed that, on reflection, a good practice statement might be appropriate, suggesting that the low confidence rating in the recommendation was misguided.

Regarding misclassification of confidence, an interviewee observed "so from that viewpoint I am sort of sharing some of the reasoning behind the recommendation which is not fully described in the remarks but perhaps at least partly and we also bring forward what has been pretty well established and described as evidence of moderate quality". This suggested to us that the interviewee recognized that the rationale for the discordant recommendation was not explicit and that despite having moderate confidence in the evidence, the panel still rated the evidence as low quality.

A number of respondents' comments suggested that they struggled with evaluating or using indirect evidence within the GRADE approach. For the discordant recommendation "all HIV infected infants and children exposed to TB through household contacts, but with no evidence of active disease, should begin isoniazid preventive (IPT) therapy", our review team considered that randomized trial evidence from adults could be applied to children, and thus confidence should be rated as moderate. The relevant guideline respondent agreed with our assessment, and yet the panel judged the evidence low confidence because there were no directly applicable

randomized trials. The interviewee stated, "but they rated the quality as very low...oh, I think they were more than moderately confident. I think that's why they felt that it be unethical [sic] not to almost on what we currently know. And frankly harmful to not, so that's why people made a strong recommendation".

For the recommendation "if bleeding does not stop in spite of treatment using uterotonics and other available conservative interventions (e.g., uterine massage, balloon tamponade), the use of surgical interventions is recommended" associated with a formal low confidence rating, the interviewee reported that the panel had high certainty in effectiveness of surgery on the basis of indirect evidence substantiates.

One respondent's comment appeared to capture three possible explanations for discordant recommendations. First, that biologic rationale actually warrants moderate confidence (contrary to GRADE guidance); second, that there may be an ethical mandate to recommend a potentially effective treatment for which there is only low quality evidence; and third, that if higher quality will never be available, a strong recommendation is warranted,

> I think it is more often that it ends up being a strong recommendation with low quality evidence when there is a compelling logic or and or [sic] an ethical argument would be we are doing the right thing and also when there are questions that are quite unlikely to be answered in the near future through high quality trials.

Minimal Harm

Respondents suggested that panel members' predominant view was that if no or minimal harm exists, even if low confidence in effect estimates, then a strong recommendation is warranted. One respondent commented,

> So you may come with a very strong recommendation because the other elements, besides the quality of evidence have a leaning towards have a strong recommendation. Let's say that you have something that is feasible, that is cheaper, that has no harm and has some benefits and it is valued by the population that is going to be benefitting with intervention, then even if the quality of evidence is low, we may come with a strong recommendation because of the combination of those elements.

Another stated "and if it helps and little harm, then why not do it? That is the thinking behind it". A specific example of this phenomenon was the recommendation "uterine massage is recommended for the treatment of post-partum hemorrhage"; a respondent argued that even though the available evidence warranted low confidence, with no cost and little discomfort to women, a strong recommendation was appropriate.

Longstanding practices

Interviewees expressed a reluctance to deliver conditional recommendations for established practices. For example, one panel took the position that in such circumstances, a strong recommendation was warranted unless there was evidence of lack of benefit, stating,

> I mean I think the worry is making a strong recommendation when you don't really know, I mean sure you can say...is a very long well established practice and it would be a ridiculous in a way not to, it would be very difficult to un-tick that without any evidence [sic].

Another interviewee described possible reactions to a conditional recommendation of an established practice: "they will say okay we've been treating the disease for the last fifteen or twenty years and now you say conditional?"

Fear of being ignored

Respondents suggested that their conditional or weak recommendations will be ignored, or that strong recommendations were required to ensure access to the treatment - particularly in resource poor settings where withholding the intervention may be punitive and depriving. For example, one respondent stated "There is some fear of inaction or depriving people of access to a particular intervention if the guideline appears to be conditional in any form or manner". Another shared,

> They will not take it seriously because the term "weak" and I think the reality is that in front of these policy makers there are so many decisions that they need to make that cost them. That kind of need their budget and if you have a few strong policies or recommendations and you a few weak, [sic] they will just throw out the weak ones.

Still another responded "There is a fear that that would send confusing signals into the public health arena and that particularly practitioners but also policy makers might misinterpret that. So it's better to be perhaps stronger on recommendations while being explicit about the quality of evidence that lies behind them".

Discussion

Although WHO guideline panelists fundamentally accept GRADE, and see merits in its structured, standardized, evidence-focused approach, there remains considerable scepticism regarding GRADE's provision for conditional recommendations. The result of this scepticism is that panels frequently ignore GRADE's caution against discordant recommendations.

Our interviews provided insight into some of the reasons for this scepticism, one of which is a conviction that a treatment is beneficial despite the panel making explicit ratings of evidence quality as low or very low. A second reason is a concern that policy and funding decision makers will ignore conditional recommendations. Further, WHO panelists sometimes feel wedded to long-established practices, and feel uncomfortable issuing any but strong recommendations regarding such practices. Related to both fear of being ignored and of raising questions about existing practices are political considerations of governments and Ministries of Health.

All respondents identified panelists' limitations in understanding of the GRADE system. This was sometimes exacerbated by unclear roles of the panel chair, inexperienced chairs, or chairs who struggle with more vocal and dominating panelists. Natural solutions regarding additional education emerged from the discussions.

Strengths and Limitations

Our study fulfilled the items of the consolidated criteria checklist for reporting qualitative research (COREQ)[21] and of the Critical Appraisal Skills Programme (CASP) critical appraisal tool for qualitative research (Appendix B).[22] We carried out analysis of the transcripts in duplicate with a high level of agreement.

CASP's question six focuses on the relationship between the researchers and conduct of the underlying study and any potential sources of bias. To varying degrees, most of the research team members are invested in GRADE methods and the production of WHO guidelines. We were cognizant of the potential resulting bias and endeavoured

to avoid possible influence at each stage of the study. To further ensure the trustworthiness and credibility of the findings, we invited a researcher (SAL) trained in qualitative research methods and who was not involved in GRADE or WHO guidelines, to function as a second coder.

Strategies to enhance study rigor included sampling to saturation, member-checking, use of reflexivity, stepwise replication, audit trails, deviant case analysis, and journal keeping. Our qualitative research design allowed key recurrent issues and themes to emerge fully from the interviews; no theme was specified in advance. We were successful in obtaining the participation of 11 of 12 guideline content chairs who were deemed to have the greatest insight into the processes of WHO panel recommendations. We recruited only 2 of 12 technical officers for the interviews, and no methods chairs.

Relation to other work

Our findings deepen our understanding of discordant recommendations that we have documented in prior studies with the endocrine guidelines[10]and WHO guidelines.[8] A number of recent articles have described guideline panels' experience working with GRADE, all of which presented views of both the strengths and the limitations of GRADE consistent with those of our respondents.[23-26] A description of the experience using GRADE by WHO-sponsored panels developing guidelines on mental, neurological, and substance use disorders[23] made no reference to discordant recommendations, nor did a description of the experience of the Canadian Task Force on Preventive Health Care.[24] Ansari et al., in a general discussion of GRADE, noted one of the circumstances when

discordant recommendations are appropriate: when high quality evidence of equivalence exists but evidence of serious harms of one alternative management strategy is only low quality, and the latter body of evidence drives the recommendation. [25]

In 2013, a published review of WHO guidelines found that although there was room for improvement, guideline quality had improved markedly since the GRC was implemented.[26] As part of their study, researchers selected guidelines published online during 2010 as the initial post-GRC sampling frame, explaining that the majority of guidelines published in 2008 and 2009 were only partially implemented. Interviews with 20 staff members found that some departments were persistently bypassing the procedures. Further, staff expressed uncertainties in applying the GRADE approach, and WHO GRC members expressed concerns that GRADE principles were not fully institutionalized. The authors concluded that the quality assurance standards the GRC had established were not yet fully embedded within the organization in 2012. These results are consistent with our findings.

In an accompanying article, we describe the results of interviews with five GRADE methodologists each of whom had participated in at least two WHO guideline panels. The findings suggest that panelists' conflicts of interest are another important contributor to discordant recommendations.

Implications

Guideline panelists often felt very confident of the effects of interventions despite their ratings of low or very low quality. Since by quality GRADE means confidence in

estimates of effect, this situation represents an unequivocal misunderstanding of GRADE.

In such circumstances panelists may be correct, or misguided, in their convictions. We found many instances of both. Panelists often had an intuitive understanding of indirect evidence, but did not realize that such evidence was clearly acknowledged within GRADE. In some such situations, a good practice statement would have been preferable.[9] In other situations, the confidence rating should have been moderate or high.

Other situations represent a disagreement with GRADE guidance. Anecdotal evidence, their own experience, or biological reasoning sometimes drove convictions regarding benefit. Panelists' concerns that decision-makers may disregard conditional recommendations represent a serious issue that requires empirical investigation. The first such investigation has provided reassuring findings of a highly appropriate greater uptake of strong versus conditional recommendations with, however, a very substantial adoption of conditional recommendations.[27]

Conclusion

The interview findings provide explanations for the considerable extent to which WHO sponsored guideline development groups fail to adhere to GRADE guidance in issuing discordant recommendations. One possible response to the problems that panels have with GRADE is to modify fundamental aspects of the GRADE approach. For

instance, one could formulate a different view of the relative merits of randomized trials and observational studies, or have only a single strength of recommendations.

If organizations such as WHO choose, however, to continue to use GRADE, our results suggest the need for enhanced GRADE training, raising awareness of existing training materials, use of evidence-to-decision frameworks - and possibly monitoring and feedback- as part of the guideline process at WHO and other organizations using GRADE. Ensuring panelists are truly committed to the GRADE process, and are free of important conflicts of interest, may also be necessary. Finally, education of guideline users regarding what panels mean by conditional recommendations, and how they can be applied, may be of use.

Table 1: Codes, categories, and themes emerging from panelist interviews

Theme	Category	Codes
Strengths of GRADE	GRADE as an assessment tool	• Rigorous structure of GRADE • Transparent, objective process for assessing existing evidence • Quantification
	Recommendations made for varying levels of quality of evidence	• Weak/Conditional/Contextual recommendations can be used for non-evidence-based considerations • Clear, unequivocal guidance for strong recommendations
Challenges/ barriers to GRADE application	Varying professional backgrounds among panelists in guideline development groups	• Observed resistance about GRADE from senior panelists • Lack of understanding in GRADE; mixed levels of training in GRADE among panelists across groups • Power dynamics within guideline development group
	Insufficient guidance in the application of GRADE	• Difficulty in applying GRADE for target outcomes with limited or no evidence, composite data or lack of controlled studies • Concern for rating up of observational, non-randomized studies • Uncertainty in GRADE's approach in prioritizing critical vs. important outcomes • Need for clearer rationale for strong recommendations based on low-confidence estimates
	PICO/research question(s)	• Time-consuming • Non-feasible • Inflexible (PICO questions cannot be altered post-hoc)
	Contributors of context	• Importance of context (accessibility of intervention, costs involved, etc.)when making recommendations
	Political environment	• Political environment may drive strong recommendations (may result in less adherence to GRADE) • Weak recommendations may create pushback from society, ministries, media
Strategies to support better application of GRADE	GRADE training	• Need for concrete examples during GRADE training • Panelist preparation prior to GRADE training • Use of suitable technology e.g. web-based interactive, to provide GRADE training in different countries
	Leadership and organizational support	• Presence of strong panel chair to lead and moderate the group, especially members who dominate • Leadership by WHO Secretariat required for building capacity around guideline development and to engage panelists

	• Presence of GRADE methodologist in each panel and for the entire duration of guideline development • Effective communication of using GRADE; provision of clear instructions to end users for understanding recommendation decisions
Explanations/ reasons for discordant recommendations	• Conditional or weak requires too much time/training for users to understand; easier to make discordant • Political environment driving discordants • Established practices driving discordants • Funding need or policy driving discordants • Desire to do minimal harm • Fear of recommendation being ignored if classified as weak or conditional • Resistance/push-back from media, ministries of health • Dominance by some panelists over inexperienced panel leader

Abbreviations:

GRADE: Grading of Recommendations Assessment, Development and Evaluation approach

PICO: Patients/Population, Interventions, Comparator, Outcome

WHO: World Health Organization

References

1.) Schünemann HJ, Fretheim A, and Oxman AD. WHO Advisory Committee on Health Research. Improving the use of research evidence in guideline development: 1. Guidelines for Guidelines. ***Health Res Policy Syst.*** **2006; 4:13.**

2.) Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet.*** **2007 ;369(9576):1883-9.**

3.) GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. ***Journal of Clinical Epidemiology.*** **Volume 64, issue 4, 383-394, April 2011. Guyatt et al. (2011).** url: http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract (Accessed March 10th 2013).

4.) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008; 336(7650):924-6.**

5.) Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.*** **2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

6.) Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M,Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G.
GRADE guidelines 15: Going from evidence to recommendation-determinants of a recommendation's directionand strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

7.) Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014;67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

8.) Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, Stoltzfus R, Montori VM, Norris SL, Schünemann HJ, Guyatt GH. World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. ***J Clin Epidemiol.*** **2014; In Press.**

9.) Guyatt GH, Schünemann HJ, Djulbegovic B, Akl EA. Guideline panels should not GRADE good practice statements. ***J Clin Epidemiol.*** **2014; In Press.**

10.) 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

11.) Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health, 23*(4), 334-340.

12.) Sandelowski, M. (2001). Real qualitative researchers do not count: The use of numbers in qualitative research. ***Research in Nursing & Health,*** **24(3), 230-240.**

13.) Walker JL. The use of saturation in qualitative research. ***Can J Cardiovasc Nurs.*** **2012 Spring; 22(2):37-46.**

14.) Thorne, S., Reimer Kirkham, S., & O'Flynn-Magee, K. (2004). The analytic challenge in interpretive description. *International Journal of Qualitative Methods, 3* (1). Article 1. Retrieved August 21st 2013 from http://www.ualberta.ca/~iiqm/backissues/3_1/pdf/thorneetal.pdf.

15.) Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: SAGE Publications, Inc.

16.) Lincoln, Y. S, & Guba, E. A. (1985). *Naturalistic inquiry*. Newberry Park, Beverly Hills, CA: Sage.

17.) Sullivan-Bolyai S, Bova C, Harper D (2005). Developing and refining interventions in persons with health disparities: the use of qualitative description. ***Nurs Outlook,*** **53(3), 127-33.**

18.) Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods* (2nd ed.). Thousand Oaks, CA: Sage.

19.) Hsieh, H.F. & Shannon, S.E. (2005). Three approaches to qualitative content analysis. ***Qualitative Health Research,*** **15(9), 1277-1288.**

20.) Hruschka, DJ, Schwartz D, Cobb St. John D, Picone E, Jenkins R, Carey J. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. Field Methods, Vol. 16, No. 3, August 2004 307–331. Miles, M. B. and A. M. Huberman (1994). Qualitative Data Analysis. 2nd edition. Thousand Oaks, CA: Sage Publications.

21.) Tong A, Sainsbury P, Craig J.
Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. ***Int J Qual Health Care.*** **2007;19(6):349-57. Epub 2007 Sep 14.**

22.) CASP. International Network on critical appraisal. Appraisal tools. url: http://www.caspinternational.org/?o=1012 (Accessed on November 26th 2014).

23.) Barbui C, Dua T, van Ommeren M, Yasamy MT, Fleischmann A, Clark N, Thornicroft G, Hill S, Saxena S. Challenges in developing evidence-based recommendations using the GRADE approach: the case of mental, neurological, and substance use disorders. ***PLoS Med.*** **2010;7(8). pii: e1000322. doi: 10.1371/journal.pmed.1000322.**

24.) Pottie K, Connor Gorber S, Singh H, Joffres M, Lindsay P, Brauer P, Jaramillo A, Tonelli M. Estimating benefits and harms of screening across subgroups: the Canadian Task Force on Preventive Health Care integrates the GRADE approach and overcomes minor challenges. ***J Clin Epidemiol.*** **2012;65(12):1245-8. doi: 10.1016/j.jclinepi.2012.06.018. Epub 2012 Sep 18.**

25.) Ansari MT, Tsertsvadze A, Moher D
Grading quality of evidence and strength of recommendations: a perspective. ***PLoS Med.*** **2009;6(9):e1000151. doi: 10.1371/journal.pmed.1000151. Epub 2009 Sep 15.**

26.) Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World health organization guideline development: an evaluation. ***PLoS One.*** **2013; 8(5):e63715. doi: 10.1371/journal.pone.0063715. Print 2013.**

27.) Umar Nasser SM, Cooke G, Kranzer K, Norris SL, Olliaro P, Ford N. Strength of WHO recommendations influences uptake in national guidelines: a systematic assessment. ***J Clin Epidemiol.*** **2014; In Press.**

Acknowledgement: WHO provided funding for the taping and transcription of interviews. Ms. Elizabeth Clow of Barrie, Ontario, Canada, transcribed all interviews.

Authors' contributions:

Mr. Alexander contributed principally to the conception of this qualitative design chapter/study, data collection, analysis, and drafting of the article. Dr. Guyatt provided guidance. All of the coauthors contributed to the drafting of the manuscript and where needed, collaboration or feedback on the methods.

Appendix A (Interview Guide applied to all interviewees)

1.) Briefly describe your public health/academic/research/clinical background as well as the general areas of focus within your regular work life.

2.) Can you tell me about your understanding in general, of the GRADE methods and how they are utilized within guideline development e.g. in terms of strength of recommendations and study quality? What do you think are its strengths and challenges of GRADE?

3.) As the lead or chair, how did the panelists in your group view GRADE and the use of GRADE as part of guideline development? Is it your sense that the panel members found GRADE application challenging?

4.) What do you think your panel members understood by the term "quality of evidence"? Or to use another phrase, "how good the evidence was"?

5.) What do you think your panel members understood as the significance or meaning of a strong recommendation?

6.) What do you think your panel members understood as the significance or meaning of a weak/conditional recommendation?

7.) Did you get the impression that your panel was required to apply GRADE to all your recommendations? (If answer is no) What was your understanding of the circumstances under which you could make recommendations and not apply GRADE?

Now I would like to understand your thinking and your impression or conception of what your panelists were thinking - their rationale- when they made certain strong recommendations based on low or very low confidence.

The interviewer will ask question 8 for each of the relevant recommendations.
8.) If it's ok with you, I would now like to move on to consideration of a specific recommendation. The one I would like to deal with states that:

Insert recommendation here and state it for the interviewee so that they can locate it in their guideline with the respective background; you can prompt the interviewee by saying

It is located in the guideline: xxxxxx
It is located on page xx.

Just to clarify, the panel made a strong recommendation based on low study quality, that xxxx. Are we correct Sir/Madame in our assessment of the comparator here?

(Presumably, respondent will say yes and may provide further explanation on the comparator).

Can you describe what you think was the panel's rationale for making this a strong recommendation?

Follow up questions could be:

How confident do you think the panel felt that xxxx?

(if respondent says at least moderately (or highly or quite) confident then ask why e.g. why are you moderately confident?

After the respondent explains why you could discuss with the interviewee this way (adjust to the specific recommendation and evidence):

By quality of evidence, GRADE means confidence in estimates of effect. If in fact you are at least moderately confident of the benefit of xxxxx, perhaps you might have labelled this as moderate quality evidence. When we considered the xxxxx studies with strong effects we thought that perhaps this was a case for a strong recommendation based on moderate quality. GRADE allows for observational studies to be upgraded in the event that there are large effect estimates and a demonstrated dose response etc. What do you think of this rationale?

(if respondent says not confident you could ask:)

Can you please explain to me why, if you or the panel were not confident in the benefit of providing xxxx, you and the panel would still make a strong recommendation. We are trying to understand how you and the panel arrived at your decision to make this recommendation and in no way are we claiming that your judgement was unreasonable or incorrect. It is just that there is no explicit explanation in the guideline or recommendation background or notes, and we would like to understand so that we can assess for remediation in that it may have been sub-optimally made and inconsistent with GRADE guidance or GRADE requires modifications. As mentioned, we understand also that you may have been entirely correct and we thus seek clarification.

Use probe questions here to help clarify with the interviewee what the decision making process was.

9.) Why we are interested in your views is because there are four criteria which is the basis for the process of going from evidence to recommendation (determining recommendation strength). Do you think the panelists used these four in their thinking and deciding? In other words were you and they aware of the four criteria to determine strength? e.g. quality of evidence, the balance between the desirable and undesirable out comes (benefits versus risks), the variability in values and preferences, and costs (resource use, feasibility, burden etc.).

What criterion then did your panel consider in moving from evidence to recommendation (to determine strength)?

If they used the four and fully described, go to Q 10.

10.) Were there any modifications to your recommendation (s) that were made after your panel's meetings? If so, what were the modifications? Can you tell us the reasons for the modification.

11.) Even though the quality of evidence was low, you probably thought that the issuing this recommendation would do more good than harm? Can you elaborate?

12.) It is possible that WHO panelists may make strong recommendations in some instances when the quality of evidence is low because if they make a weak/conditional recommendation policy makers and funders are likely to ignore the recommendation. What are your thoughts on this?

13.) It is possible that WHO panelists may make strong recommendations in some instances when the quality of evidence is low because a practice has become very well established. What are your thoughts on this?

14.) Are you aware of the available training videos that panelists can review ahead of time (before the guideline development begins) and were these reviewed by you? Do you think the panel members knew of these videos and reviewed them? Do you think there is a need for GRADE orientation prior to participation in the development of a guideline?

Appendix B: Critical Appraisal Skills Programme (CASP)[22] critical appraisal tool for qualitative research

Question	Assessment response
1. Was there a clear statement of the aims of the research?	Yes, pages 4 and 5 of this manuscript
2. Is a qualitative methodology appropriate?	Yes, we aimed to acquire an in-depth understanding of the panelists' reasons for making discordant recommendations and in so doing, describe the rationales for the discordant recommendations and perceptions of GRADE.
3. Was the research design appropriate address the aims of the research?	Yes, this is described on pages 5 and 6 of the manuscript
4. Was the recruitment strategy appropriate to the aims of the research?	Yes, we addressed how participants were selected and why they would be best positioned to answer the questions; we could not recruit the intended sample despite four reminders and recognize that this is in part due to the possible trace-back that WHO interviewees may have feared; we did however anonymize and de-identify all data, destroying audio-recordings.
5. Were the data collected in a way that addressed the research issue?	Yes, we can answer all aspects in the affirmative, including saturation
6. Has the relationship between researcher and participants been adequately considered?	Researchers in the current study were aware of the roles and the preconceived biases which may affect the interpretation of the findings. This is further described in page 14 of this manuscript.
7. Have ethical issues been taken into consideration?	Yes, we received ethical approval and in the introductory letter and consent form, interviewees were explained all aspects regarding anonymity etc.
8. Were the data analysis sufficiently rigorous?	Yes, we used inductive content analysis, seeking codes, categories, and emergent themes
9. Is there a clear statement of findings?	Yes, findings explicitly stated with reference quotes; we used member-checking and also, while this is the first such interview study on WHO guideline development, we framed the results relative to other similar research.
10. How valuable is the research?	It is very valuable to the improvement of WHO guideline development via the use of GRADE and especially the making of discordant recommendations; also very informative to guideline development elsewhere.

CHAPTER 7: Phase IV:

Experiences of senior GRADE methodologists as part of WHO guideline development panels: an inductive content analysis

[1]Paul E Alexander, [1]Shelly-Anne Li, [2]Michael R. Gionfriddo, [3]Lisa Bero, [4]Rebecca J Stoltzfus, [1,5]Ignacio Neumann, [6]Juan P. Brito, [7]Benjamin Djulbegovic, [8]Victor M. Montori, [9]Holger J Schünemann, [10]Gordon H Guyatt

[1] Health Research Methods (HRM)
Department of Clinical Epidemiology and Biostatistics
McMaster University, 1280 Main Street West,
Hamilton, Ontario, L8N 3Z5, Canada.

[2]Mayo Graduate School, Mayo Clinic
Doctoral student
Knowledge and Evaluation Research Unit, Mayo Clinic

Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[3]Professor (Chair of Medicines Use and Health Outcomes)
The University of Sydney
Charles Perkins Centre
Camperdown NSW 2006

[4]Professor
Division of Nutritional Sciences
120 Savage Hall
Cornell University
Ithaca NY 14853, USA

[5]Internal medicine specialist
Department of Internal Medicine
Pontificia Universidad Católica de Chile
Santiago, Chile

[6]Assistant Professor of Medicine
Investigator, Knowledge and Evaluation Research Unit
Divisions of Endocrinology, Diabetes, Metabolism and Nutrition
Mayo Clinic
Rochester, Minnesota
Plummer 3-35, 200 First Street SW, Rochester
MN 55905, USA

[7]Distinguished Professor
University of South Florida
H Lee Moffitt Cancer Center
Florida, USA

[8]Professor of Medicine
Knowledge and Evaluation Research Unit Divisions of Endocrinology and Diabetes and Health Care and Policy Research
Mayo Clinic
Rochester, Minnesota
Plummer 4-402, 200 First Street SW, Rochester
MN 55905, USA

[9]Professor and Chair
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
Michael Gent Chair
in Healthcare Research
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON L8S 4K1, Canada

[10] Distinguished Professor
Department of Clinical Epidemiology & Biostatistics
Professor of Clinical Epidemiology and Medicine
McMaster University Health Sciences Centre
1200 Main Street West, Room 2C12
Hamilton, Ontario, L8S 4K1, Canada.

Corresponding author: Paul E. Alexander, doctoral candidate.

E-mail: elias98_99@yahoo.com

This chapter/qualitative study has been submitted to the JCE for peer-review for publication in their January to March 2015 edition.

Following the Phase I (Chapter three), Phase II (Chapter four), and Phase III (Chapter six), we were unable to recruit methodologists trained in GRADE and who were part of the guideline panels emerging from Phases II and III. We were able to subsequently recruit senior GRADE methodologists who were willing to be interviewed as part of this study (but were not part of the Phase II and III guidelines). Methodologists emerged as key informants from Phase III and we thus embarked on the Phase IV so that our

understanding of the reasons for strong low confidence recommendations could be deepened. We decided to proceed given the methodologists' capacity to comment in general, about their experiences as part of WHO panels when strong recommendations based on low or very low confidence in effect estimates were being made.

Abstract

Background

The World Health Organization (WHO) classifies a substantial proportion of their recommendations as strong despite low or very low confidence (certainty) in estimates of effect. Such discordant recommendations are often inconsistent with GRADE guidance.

Objective

To gain the perspective of senior WHO methodology chairs regarding panels' use of GRADE, particularly regarding discordant recommendations.

Data sources

Senior active GRADE methodologists who had served on at least two WHO panels and were an author on at least one peer-reviewed published manuscript on GRADE methodology.

Methods

Five eligible methodologists participated in detailed semi-structured interviews. Respondents answered questions regarding how they were viewed by other panelists and WHO leadership, and how they handled situations when panelists made discordant recommendations they felt were inappropriate. They also provided information on how the process can be improved. Interviews were recorded and transcribed, and inductive content analysis was used to derive codes, categories, and emergent themes.

Results

Three themes emerged from the interviews of five methodologists: i) The perceived role of methodologists in the process, ii) Contributors to discordant recommendations and iii) Strategies for improvement. Salient findings included i) a perceived tension between methodologists and WHO panels as a result of panel members' resistance to adhering to GRADE guidance; ii) both financial and non-financial conflicts of interest among panel members as an explanation for discordant recommendations; and iii) the need for greater clarity of, and support for, the role of methodologists as co-chairs of panels.

Conclusions

These findings suggest that the role of the GRADE methodologist as a co-chair needs to be clarified by the WHO leadership. They further suggest the need for additional training for panelists, quality monitoring, and feedback to ensure optimal use of GRADE in guideline development at WHO.

Background

The World Health Organization (WHO) produces and disseminates public health guidelines, often focused on low and middle income nations. To improve their guideline development process, in 2003 the WHO initiated adoption of the Grading of Recommendations Assessment, Development and Evaluation (GRADE)[1,2-4] approach to guideline development, and in 2007 initiated the Guideline Review Committee (GRC). The advent of the GRC saw a heightened focus on the use of GRADE as part of WHO's guideline development.

The GRADE approach involves judgements about the confidence in effect estimates (study quality or certainty) as high, moderate, low, and very low and suggests two categories of recommendations: strong and conditional.[2-4]

GRADE guidance cautions against making strong recommendations when there is low or very low confidence in effect estimates (discordant recommendations).[3-6] The GRADE working group has, however, identified restricted circumstances[6] in which discordant recommendations may be appropriate.

In a recent study of WHO guidelines using GRADE published from 2007 to 2012,[7] we found that 63.4% of the recommendations were strong and of these 20.6% represented a misclassification of confidence[8] (should have been with at least moderate confidence), 29 (18%) good practice statements,[8,9] and 45.6% as more appropriately a conditional rather than a strong recommendation.[8]

To gain further insight into the reasons for discordant recommendations, we conducted an interview study[10] of WHO guideline panels involved with selected guidelines.[7,8] Based on interviews with 13 panel members, we found that reasons for discordant recommendations included skepticism about the value of conditional recommendations; political considerations such as meeting the needs of ministries of health; a high certainty in benefits (sometimes warranted, other times not) despite rating evidence as warranting low confidence; a reluctance to make conditional recommendations for long- accepted practices; and concerns that conditional recommendations will be ignored.[10]

In a presentation of the results of the interview study[10] at an international meeting, a number of GRADE methodologists who had participated in WHO panels made comments that suggested additional valuable perspectives. This, and our failure to include any methodology chairs in our initial sample of 13 respondents, motivated us to undertake an additional qualitative study to understand the experience of GRADE methodologists who have worked on WHO guidelines that made discordant recommendations.

Methods

A qualitative descriptive methodological approach was employed, given that it facilitates low-level inference and allows for staying close to the surface of the data.[11-15] This approach generates a thorough description of the experiences from the senior GRADE methodologists. The qualitative descriptive approach has previously been used

to improve guideline and program development.[16,17] The Ethics Review Boards at WHO, Geneva; Switzerland and McMaster University, Hamilton, Ontario, Canada approved the study.

Guided by purposeful sampling,[18] we identified and recruited senior active GRADE members who served as chairs on at least two WHO panels with at least one GRADE peer-reviewed methods publication. We constructed a schedule of eight open-ended questions for the interviews that focused on how methodologists saw their role in the WHO guideline panels on which they served, how they handled panels' inclination to make discordant recommendations, the drivers of discordant recommendations, and their suggestions about how to improve the guideline development process. Data were collected using individual, semi-structured audio-recorded interviews that lasted approximately 30 minutes.

Interviews were transcribed verbatim, de-identified using a code-breaker prepared by an independent transcriptionist, and analyzed as the interviews were conducted. We stored the hand-written interview notes as well as the audio recordings in a locked file cabinet and a password protected laptop. Upon transcription and analysis, the audio recordings were destroyed.

To analyze the data we used inductive, conventional content analysis[19,20] and did not impose preconceived theoretical perspectives on the data. The coders (PEA, SAL) independently read each interview transcript in its entirety thrice. During the coding process, the coders highlighted important portions of raw text and quotes, took note of

salient concepts, and finally chose a word or a short phrase to represent the meaning of a specific text segment. Codes were revised and new codes were added as new concepts emerged. Codes with similar concepts were grouped together to form categories. Themes emerged from codes and categories.

Coders developed a codebook of definitions for each code with sample segments from the transcripts, beginning with a list of preliminary codes and revising the codebook as new concepts and codes emerged. To increase study rigor and to ensure that coders were viewing the responses similarly, we used field notes, audit trails, and intercoder agreement. To further enhance credibility of the findings, we performed member checking to ensure the coders' interpretations of the interview content were accurate following each interview.

Results

We approached eight senior GRADE methodologists who all agreed to participate; of the five who met eligibility criteria (three had participated in only one panel), four were male and one female.

Themes and insights from GRADE methodologists

The findings revealed three overarching themes (see Table 1 for full list of codes, categories, and themes):

i) *The perceived role of methodologists in the process*

Methodologists felt initial push back from panelists regarding adherence to GRADE guidance but recognized that panelists also appreciated their contribution. For

example, one methodologist stated "there are mixed perceptions, ranging from scapegoats and initially a negative reaction". Another shared "I feel sometimes a bit on attack that they're attacking [sic] because you are forcing them to use the GRADE process". Often, however, this grew into appreciation as the process unfolded: "eventually I think they appreciate that you're there."

Methodologists reported that continuous efforts were necessary to ensure panelists' adherence to GRADE principles. One respondent shared,

> So it sometimes it works and sometimes I feel like I'm policing what they are writing, even what they are putting in presentations. So sometimes it doesn't really work and you are again you are turning into as you had said before an adversary against them.

Similarly another respondent reflected,

> They try to change the wording to make it seem like a strong recommendation, it seems like they try to find lots of other tricks to make it a strong recommendation in other publications or in other flow charts or other support materials for the guideline, they will actually change it to the wording of a strong recommendation and I am constantly having to double check whatever they produce to make sure that it stays as a weak recommendation.

ii) *Contributors to discordant recommendations*

Similar to the panel members,[10] methodologists felt that limited prior exposure to GRADE, limited understanding of GRADE, political agenda pressures, and contextual factors all played a role in driving discordant recommendations. They put greater emphasis, however, on conflict of interest. For instance, one methodologist elaborated, "We've seen in some cases there was a clear conflict of interest playing out and pushing a

strong recommendation, you know what should have been a conditional recommendation based on very low quality of evidence". Methodologists' perceived conflicts were largely related to funding: "there could also be conflict of interest of those WHO offices that have a very cozy relationship with industry".

Conflicts could also be related to political issues. As one respondent explained, "We know that there is [sic] a lot of political agendas and other agendas of guideline panel members and those can be the reasons why people want strong recommendations". Methodologists also observed that a desire to be invited back as a panelist can constitute a conflict:

> When they decide who to invite on the panel, they invite panelists who are typically quote unquote compliant and willing to go with what that person would like to see in terms of recommendation. Because that person knows that these panelists would like to come back. So it is kind of a bargain...person wishes in terms of what the recommendation should be and in turn they get a chance to come back again for the next guideline which for many as you apparently know is a privilege and people might be willing to go for that otherwise.

Time urgency also emerged as a concern for methodologists that was not vocalized by panel members:

> Well I think some of the problems that WHO has are homemade. They bring in, I don't know, five, six, seven, ten people from all over the world to Geneva, and they want a result within two days. And that creates an atmosphere where sometimes just getting it finished is more important than getting it finished in an objective and well-done way.

iii) *Strategies for improvement*

The methodologists raised issues associated with their roles, responsibilities, and expectations particularly when they functioned as guideline panel co-chairs. There were

times when the methodologists themselves were unaware of their role as co-chairs in addition to fulfilling the role as a health research methods expert. They suggested a need for explicit indications by WHO of their roles. For example:

> In the future, will be really, really helpful if WHO, if first of all the role of the methodologist co-chair will be clarified and more generally in a sense what WHO and panels expect from methodologists and then also to make sure that methodologists have appropriate support by WHO staff and also the methodology unit like the GRC within WHO when important discussions take place…because methodologists do feel lonely trying to fight for making good sense of the evidence and it is a bit tedious if you feel that it is like you doing that and nobody else really cares and in a sense that feeling was there from time to time.

Methodologists offered suggestions that did not emerge from other panelists' interviews. One was a need for specific tools that could aid methodologists:

> I think we may need tools to help the methodologists who are in there to also convey how recommendations are made or how they can, like you are asking me what did you do as a solution when you were faced with a panel that wanted to make a strong recommendation that didn't fit.

Methodologists also urged a more effective use of the remarks section:

> You know there is this remarks section underneath the recommendations, I actually think the remarks section needs to be better structured or formatted… If we said it's a weak recommendation and they still want to leave it strong, that if they see a good remark section that communicates well I think they feel less not reluctant, they feel okay in making a weak recommendation.

Discussion

This study highlights the challenges and tension faced by GRADE methodologists as part of WHO guideline panels. Similar to the findings from the WHO panel interview study,[10] methodologists observed that the main drivers for making discordant recommendations include a fear that decision-makers and funding bodies would overlook

conditional recommendations, commitment to long-established practices, and a perception that evidence warrants high confidence even when the formal rating is low confidence. Conflicts of interest, including both financial and non-financial conflicts, emerged as another prominent explanation for discordant recommendations (Table 1).

Also consistent with WHO panel interview findings,[10] all interviewed methodologists indicated that many panelists have a limited understanding of GRADE, underscoring the need for GRADE training at the start of the process. Along with highlighting the role of conflicts of interest, one of the core findings of this study is the respondents' identification of an often unclear co-chair role of the GRADE methodologist that they feel magnified the tensions (Table 1).

Strengths and Limitations

Our study followed the consolidated criteria checklist on reporting qualitative research (COREQ)[21] and considered the domains of quality recognized by the CASP critical appraisal tool for qualitative research.[22] Our qualitative inductive content analysis design allowed themes to emerge fully from the interviews. To maintain study rigor, we utilized field notes, audit trails, reflexivity, journal keeping, and assessed intercoder reliability. We also employed member-checking to ensure that the emerging codes and themes accurately reflected what the interviewees stated.

There are only a small number of very senior methodologists with in-depth experience and understanding of GRADE, and we were able to identify only five such

individuals who had been co-chairs of WHO guideline panels. The consistency of their comments suggests, however, that their experience was not idiosyncratic.

Relation to prior work

A number of more recent articles have described guideline panels' experience working with GRADE[23-26]; none of these had an important focus on discordant recommendations. There are two other studies that did provide a characterization of discordant recommendations worth mentioning. One is a study by Brozek et al.[27] that sought to develop explicit, unambiguous, and transparent clinical recommendations in a systematic manner for the treatment of allergic rhinitis (using GRADE) and the other was the Endocrine Society (TES) Guideline Study which sought to characterize strong recommendations of TES based on low or very low confidence evidence (using GRADE).[28] The latter study[28] found that 58% of the 357 recommendations made by TES between 2005 and 2011 were strong, and 59% of those strong were discordant. These findings mirror what we found for our WHO studies[7,8] though a noticeably higher proportion of the TES discordant recommendations were judged to be consistent with GRADE principles (29%). On the other hand, the allergic rhinitis study[27] (The Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines) appeared to contradict the WHO and TES experiences[7, 8, 28] whereby researchers graded 17% of recommendations as strong.[27,29] Researchers suggested that several reasons may account for the greater numbers of weak or conditional recommendations and lower number of strong

recommendations including the often lower-quality evidence for patient important outcomes that are critical to decision making, as well as the resulting lowered confidence that the benefits clearly outweigh the downsides.[29] In addition, the role of variability of values and preferences that decision makers would assign to the key patient outcomes and the potential uncertainty that the benefits and the required resources for an intervention may not outweigh less resource-intensive alternative actions.[29] Indications are that management of conflict of interest and the presence of a strong panel chair for the ARIA guideline development process[29] may have accounted for the lesser number of strong recommendations. Indeed this also emerged as key findings when methodologists were interviewed for this study.

Most relevant to this current Phase IV study however is a review of WHO guidelines that, although authors reported that guideline quality had improved since the GRC was implemented, also noted GRC members' concern that GRADE principles were not fully institutionalized and that some departments were persistently bypassing the procedures.[26] These latter observations are consistent with the results of our WHO panel interview study[10] and in particular with the current findings. Methodologists clearly perceived the extent of limited buy-in to GRADE.

Implications

These findings suggest a limited understanding of GRADE among many panelists, some disagreement with GRADE guidance, and a tension between methodologists and guideline panel members at WHO. If WHO, or other guideline

panels elsewhere, are committed to GRADE, they may institute further education of panelists, and include methodologists with clear and explicit roles on the panels (ideally, chairs or co-chairs).

Conclusion

Our findings provide further insight into the reticence of WHO panels to adhere to GRADE guidance and particularly when it comes to issuing discordant recommendations. WHO leadership and panel chairs must provide clarity and support to the role of the methodologists as co-chairs of panels. More rigorous standards for managing and limiting conflict of interest may also be advisable.

Table 1: Codes, categories, and themes emerging from methodologist interviews

Themes	Category	Codes
Perceived role of methodologists during WHO guideline development	Mixed perceptions of self as methodologist in WHO guideline development	• Methodologist perceived initially as "scapegoat" for discordant recommendations • Lack of support for methodologists • Persistent efforts by panels to change weak recommendations to strong • Lack of clear roles for methodologists as co-chairs
Proposed contributors leading to discordant recommendations	COI	• Funding interest driving panel decisions • Panelists' own conflicts of interest driving strong discordant recommendations • Choosing panelists based on their ability to comply
	Panelists' limited understanding of GRADE and PICO	• Inefficient use of time for generating structured PICO question • General misunderstandings of GRADE methods • Use of strong recommendations to drive compliance • Limited experience producing evidence based guidelines
	Political environment	• Political statements being made by WHO • Lack of GRADE acceptance by WHO leadership • Cost savings, budget, policy as drivers of strong recommendations • Commitment to established practices
	Logistic issues	• Time urgency in producing recommendations
	Feasibility	• Each nation's financial status as major contributor to type of recommendations • Real-world application of recommendations
Suggested strategies for improvements/changing practice	Leadership and organizational support	• Additional GRADE training of panel may not be helpful • Improve WHO support for methodologists • Need more effective use of remarks (or rationale) section • Focus training on the definitions of weak/conditional recommendations

References

1.) Schünemann HJ, Fretheim A, and Oxman AD. WHO Advisory Committee on Health Research. Improving the use of research evidence in guideline development: 1. Guidelines for Guidelines. ***Health Res Policy Syst.*** **2006; 4:13.**

2.) Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet.*** **2007 ;369(9576):1883-9.**

3.) GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. ***Journal of Clinical Epidemiology.*** **Volume 64, issue 4, 383-394, April 2011. Guyatt et al. (2011).** url: http://www.jclinepi.com/article/S0895-4356(10)00330-6/abstract (Accessed March 10th 2013).

4.) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ.*** **2008; 336(7650):924-6.**

5.) Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.*** **2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

6.) Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M,Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G.
GRADE guidelines 15: Going from evidence to recommendation-determinants of a recommendation's directionand strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

7.) Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014;67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

8.) Alexander PE, Juan P. Brito, Ignacio Neumann, Michael R. Gionfriddo, Lisa Bero, Benjamin Djulbegovic, Rebecca Stoltzfus, Victor M. Montori, Susan L. Norris, Holger J Schünemann, Gordon H Guyatt. World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. ***J Clin Epidemiol.*** **2014; In Press.**

9.) Guyatt GH, Schünemann HJ, Djulbegovic B, Akl EA. Guideline panels should not GRADE good practice statements. ***J Clin Epidemiol.*** **2014; In Press.**

10.) Phase III study citation here; JCE indicated that they would accept submission.

11.) Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health, 23*(4), 334-340.

12.) Sandelowski, M. (2001). Real qualitative researchers do not count: The use of numbers in qualitative research. *Research in Nursing & Health, 24*(3), 230-240.

13.) Walker JL. The use of saturation in qualitative research. ***Can J Cardiovasc Nurs.*** **2012 Spring;22(2):37-46.**

14.) Thorne, S., Reimer Kirkham, S., & O'Flynn-Magee, K. (2004). The analytic challenge in interpretive description. *International Journal of Qualitative Methods, 3* (1). Article 1. Retrieved August 21st 2013 from http://www.ualberta.ca/~iiqm/backissues/3_1/pdf/thorneetal.pdf.

15.) Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: SAGE Publications, Inc.

16.) Lincoln, Y. S, & Guba, E. A. (1985). *Naturalistic inquiry*. Newberry Park, Beverly Hills, CA: Sage.

17.) Sullivan-Bolyai S, Bova C, Harper D (2005). Developing and refining interventions in persons with health disparities: the use of qualitative description. *Nurs Outlook,* 53(3),127-33.

18.) Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods* (2nd ed.). Thousand Oaks, CA: Sage.

19.) Hsieh, H.F. & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research,* 15(9), 1277-1288.

20.) Shi C, Tian J, Wang Q, Petkovic J, Ren D, Yang K, Yang Y. How equity is addressed in clinical practice guidelines: a content analysis. ***BMJ Open.*** **2014;4(12):e005660. doi: 10.1136/bmjopen-2014-005660.**

21.) Tong A, Sainsbury P, Craig J.
Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. ***Int J Qual Health Care.*** **2007;19(6):349-57. Epub 2007 Sep 14.**

22.) CASP. International Network on critical appraisal. Appraisal tools. url: http://www.caspinternational.org/?o=1012 (Accessed on December 7th 2014).

23.) Barbui C, Dua T, van Ommeren M, Yasamy MT, Fleischmann A, Clark N, Thornicroft G, Hill S, Saxena S. Challenges in developing evidence-based recommendations using the GRADE approach: the case of mental, neurological, and substance use disorders. ***PLoS Med.*** **2010;7(8). pii: e1000322. doi: 10.1371/journal.pmed.1000322.**

24.) Pottie K, Connor Gorber S, Singh H, Joffres M, Lindsay P, Brauer P, Jaramillo A, Tonelli M. Estimating benefits and harms of screening across subgroups: the Canadian Task Force on Preventive Health Care integrates the GRADE approach and overcomes minor challenges. ***J Clin Epidemiol.*** **2012;65(12):1245-8. doi: 10.1016/j.jclinepi.2012.06.018. Epub 2012 Sep 18.**

25.) Ansari MT, Tsertsvadze A, Moher D
Grading quality of evidence and strength of recommendations: a perspective. ***PLoS Med.*** **2009 ;6(9):e1000151. doi: 10.1371/journal.pmed.1000151. Epub 2009 Sep 15.**

26.) Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World health organization guideline development: an evaluation. ***PLoS One.*** **2013; 8(5):e63715. doi: 10.1371/journal.pone.0063715. Print 2013.**

27.) Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW, Casale TB, van Wijk RG, Ohta K, Zuberbier T, Schünemann HJ; Global Allergy and Asthma European Network; Grading of Recommendations Assessment, Development and Evaluation Working Group. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. ***J Allergy Clin Immunol.*** **2010;126(3):466-76. doi: 10.1016/j.jaci.2010.06.047.**

28.) 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

29.) Schünemann HJ. Guidelines 2.0: do no net harm- the future of practice guideline development in asthma and other diseases. ***Curr Allergy Asthma Rep.*** **2011;11(3):261-8. doi: 10.1007/s11882-011-0185-8.**

Acknowledgement: Ms. Elizabeth Clow of Barrie, Ontario, Canada, is a trained transcriptionist who transcribed all interviews.

Authors' contributions:

Mr. Alexander contributed principally to the conception of the chapter/qualitative study involving senior GRADE methodologists, the data collection, analysis, and drafting of the article. Dr. Guyatt provided guidance. All of the coauthors contributed to the drafting of the manuscript and where needed, collaboration or feedback on the methods.

CHAPTER 8:

Strong recommendations based on low quality evidence (discordant recommendations): guidance for WHO guideline developers

Paul E Alexander and Gordon Guyatt

This guidance chapter has been accepted for publication by the World Health Organization (WHO) Guideline Secretariat, Geneva, Switzerland. The guidance will be published in WHO's upcoming 2015 Guidelines' handbook.

This guidance document was requested by WHO for their panels as part of the guideline development process and specifically when confronted with making strong recommendations based on low or very low confidence in effect estimates. This guidance was developed based on the sum total of all research findings from Chapters three, four, six, and seven (and preceding work by the GRADE Working Group).

Background:

The World Health Organization (WHO) is a global public health leader and provides guidance to all nations, guidance that is particularly important to lower income nations. WHO produces public health guidelines geared toward healthcare practitioners, policy developers, and consumers.[1] WHO strives to develop guidelines that are supported by systematic reviews of the underlying evidence and of the highest methodological quality. The aim is to use a transparent, systematic, and evidence-based decision making process that allows for an in-depth analysis of the desirable and undesirable outcomes of healthcare options.[2,3]

In 2007, the guideline development process at WHO received a strong critique that documented guideline developers' failure to use a systematic evidence-based approach.[4] The critique prompted WHO to initiate the Guidelines Review Committee (GRC) Secretariat and commit to using standardized guideline procedures outlined in the WHO handbook for guideline development.[5,6] This is accompanied by GRC quality improvement efforts, and GRC involvement in the final approval of the guidelines.

Along with the GRC, to improve the guideline development process and to facilitate explicit, systematic, and transparent movement from evidence to guideline recommendations, in 2007 WHO heightened the use of the GRADE system that was initially adopted in 2003.[7-10] GRADE is now more widely used across WHO in guideline development and in a large number of organizations worldwide (see www.gradeworkinggroup.org for a list of over 80 organizations using GRADE).[11-15]

Following the GRADE approach,[16,17] best estimates of intervention effects come from systematic reviews of randomized controlled trials (RCTs) of the impact of alternative management approaches. Factors including risk of bias (consideration of randomization to treatment arms; allocation concealment of the generated randomization sequence; blinding of study participants, clinicians and outcome assessors; losses to follow-up; and selective outcome reporting), imprecision (95% confidence intervals); indirectness (similarity of patients, interventions, comparators, and outcomes studied versus those relevant to the recommendation); inconsistency (variability of effect estimates across studies); and publication bias determine certainty (confidence or quality) of the evidence. GRADE rates certainty as high, moderate, low and very low; except under unusual circumstances, observational studies provide only low certainty.[8,10,13]

GRADE provides guidance in moving from evidence to recommendations.[9] Panels using GRADE offer strong recommendations when they are confident that desirable consequences (benefits) do or do not outweigh undesirable consequences (risks/harms), and conditional (synonyms are weak, discretionary, or contingent) recommendations when the balance of desirable and undesirable consequences is less certain.

Using the GRADE approach, determinants of the strength of recommendations include certainty in estimates of effects, magnitude of the desirable and undesirable consequences of alternative courses of action, value and preference judgements required in trading off desirable and undesirable consequences, uncertainty regarding patients'

values and preferences, variability in these values and preferences, and resource use considerations.[9] In the context of public health guidelines, additional factors including the burden of illness, accessibility, feasibility, acceptability, social context, the extent of current suboptimal practice, and the impact on health inequities may require consideration.

GRADE suggests that the values and preferences that inform decisions should be those of the population (patients or the general public) to whom a recommendation will be applied. GRADE recommends a systematic review of research relevant to values and preferences. Beyond that, mechanisms for ensuring typical preferences of patients/populations informing guidelines are controversial and challenging. Guideline panels should make their underlying values and preferences as explicit as possible, ideally including quantitation.

Cause for Concern Regarding WHO Guidelines:

In 2013, a review of WHO guidelines found that guideline quality had improved markedly since the GRC was in place.[18] Interviews with 20 staff members found, however, that some departments were purposefully bypassing the procedures. Further, staff expressed uncertainties in applying the GRADE approach, and the GRC expressed concerns that GRADE principles were not fully institutionalized. The authors concluded that the quality assurance standards the GRC has set were not yet fully embedded within the organization.

Initial anecdotal evidence, as well as findings from another organization,[19] suggested a specific problem: WHO guideline panels have been making many strong recommendations based on low or very low evidence certainty (discordant recommendations). GRADE guidance warns against discordant recommendations: in the face of low certainty evidence, high confidence that an intervention does more good than harm may be misguided. Given that strong recommendations are "just do it" directives recommended to all or almost all guideline users and under all or almost all foreseeable circumstances, discordant recommendations may entrench practices that are sub-optimal. The result could be WHO guideline users adopting recommended actions detrimental to patients or populations.

Despite reservations regarding discordant recommendations, the GRADE Working Group has identified five paradigmatic situations in which such recommendations may be warranted (Table 1).[9] The question therefore arises whether WHO guideline panelists are or are not adhering to GRADE guidance when they offer discordant recommendations.

Descriptive summary WHO guidelines 2007-2012

An initial descriptive study of WHO`s recommendations conducted in 2012 confirmed the frequent use of discordant recommendations.[20] Of 456 GRADEd recommendations from 43 WHO guidelines created from 2007 to 2012, 289 (63.4%) were strong and 167 (36.6%) were conditional/weak. Of the 289 strong recommendations, 160 (55.4%) were discordant.

Detailed Examination of WHO Guidelines:

We followed up the descriptive study[20] with an in-depth examination of the 160 discordant recommendations.[21] First, we determined if the recommendations were consistent with one of the five Table 1 paradigms (discordant recommendations consistent with GRADE guidance). For recommendations judged to be inconsistent with GRADE guidance (i.e. not consistent with one of the five Table 1 situations[22-26]), we classified them using a taxonomy we had previously developed for an exploration of Endocrine Society guidelines[9,19] (Table 2). This taxonomy classifies recommendations as good-practice statements[27] (high certainty in estimates, but that certainty is based on indirect evidence best not subjected to the GRADE process), misclassifications of evidence (likely moderate or high certainty in effect estimates), or recommendations best graded as conditional rather than strong (Table 2).

Early in the data abstraction process we encountered challenges when comparators to the interventions were not explicit or obvious. We therefore classified the comparator as: i) explicit, identified clearly in the recommendation ii) not explicit in the recommendation, but obvious or easy to infer from the guideline text and iii) not identified in the recommendation and unclear in the guideline text. For category (iii), reviewers used their best judgement to determine the likely comparator. Less than 20% of the recommendations made the comparator clear in the recommendation, and in over 50% it was not clear even after review of the entire text associated with the

recommendation (Table 3). When there was a disagreement, discussions to achieve consensus took place and, when necessary, a third party adjudicated the discussion.

Reviewers, working in duplicate and independently, judged 25 (15.6%) of the 160 discordant recommendations to be consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations, most commonly (10/42%) corresponding to the fifth paradigm (Table 1).

Of the remaining 84.4% of the recommendations, reviewers judged 29 (18% of the 160 total) as best presented as good practice statements,[27] 33 (20.6%) as a misclassification of certainty (evidence warranted moderate or high certainty), and 73 (45.6%) as recommendations more appropriately graded as conditional.

Table 2 provides examples of each type of recommendation inconsistent with GRADE guidance (i.e. good practice, misjudgement of certainty, and recommendations more appropriately categorized as conditional); further examples follow.

Examples of other discordant recommendations we judged as good practice include: "Minimize time spent in health-care facilities"; "Laboratory monitoring for toxicity should be symptom directed"; and triage people with tuberculosis symptoms.

An example of an error in certainty rating highlights the issue of indirect evidence that most frequently underlies the mistaken rating of low or very low certainty. "All HIV infected infants and children exposed to TB through household contacts, but with no evidence of active disease, should begin isoniazid preventive therapy (IPT) therapy". Our

review team considered that the available randomized trial evidence from adults could be applied to children, and thus certainty should be rated as moderate.

Another example of a discordant recommendation that would best have been presented as a conditional recommendation was "Intermittent iron and folic acid supplementation is recommended as a public health intervention in menstruating women living in settings where anaemia is highly prevalent". Because evidence supporting iron and folic acid supplementation warranted only low confidence, our judgement was that a conditional recommendation would have been more appropriate.

Of the three categories inconsistent with GRADE guidance – good practice statements, misjudgement of the quality of the evidence, and warranting only a conditional recommendation – the third is of most concern. Good practice statements may be appropriate when indirect evidence that is difficult to collect and summarize actually warrants high certainty in intervention impact, and when the gradient between desirable and undesirable consequences of an intervention is large. Thus, like a misjudgement of evidence quality as low or very low when moderate or high certainty is appropriate, the problem with good practice statements is not the strong recommendation, but rather the certainty in estimate judgement. In these two situations, the strong recommendation, and thus the most important message to the clinician and the policy maker – just do it – is warranted.

This is not true for the final category, recommendations that would optimally have been graded as conditional. This is not a small concern: 46% of WHO discordant

recommendations, and 16% of all their recommendations made from 2007 to 2012, fall in this category.

In depth exploration (interview study findings):

Following the taxonomy exercise[21], we conducted 18 open-ended semi-structured one-on-one interviews with WHO panellists. Participants included 13 panelists in an initial rounds of interviews and, subsequently, five expert GRADE methodologists. We selected recommendations from among guidelines that revealed the highest proportions of discordant recommendations and were among the 135 we judged to be inconsistent with GRADE guidance (HIV/AIDS, TB, maternal health, and child health guidelines) (Table 2).

The initial 13 interviews with content expert panel chairs and technical officers revealed four overarching themes: i) strengths of GRADE, ii) challenges/barriers to GRADE application, iii) strategies to support improved use of GRADE, and iv) explanations for discordant recommendations. Explanations for discordant recommendations included a general scepticism regarding conditional recommendations; political considerations such as Ministries of Health; a high certainty in benefits (sometimes warranted, sometimes not) despite the official label of low certainty; a reluctance to make conditional recommendations for long-standing accepted practices; and concerns that conditional recommendations will be ignored.

The five methodologist interviewees provided additional key insights: i) a tension between methodologists and WHO panels as a result of hostility toward insistence on

adhering to GRADE guidance ii) both financial and non-financial conflict of interest as an explanation for discordant recommendations and iii) the need for greater clarity of the role of methodology co-chair and support of that role.

More specifically, for the 13 initial interviews, the themes revealed:

i) Panelists identified and appreciated the *strengths of GRADE* including its highly structured, open, analytical, standardized, evidence-focused approach and the attention to patient values and preferences in moving from evidence to recommendation.

ii) *Challenges/barriers to GRADE application* included insufficient GRADE guidance on making weak or conditional recommendations; difficulties dealing with situations in which RCT evidence was unavailable; and challenges developing PICO questions.

iii) Suggested *strategies to support improved use of GRADE methods/guidance* included the provision of GRADE training in general and particularly at the start of guideline development. Respondents suggested that such training use concrete examples in web-based interactive videos that would include instruction on developing PICO questions and matching them to the evidence. Respondents also identified the desirability of increased WHO Guideline Secretariat leadership and organizational support.

iv) Reasons *for discordant recommendations* included a reluctance to deliver conditional recommendations for established longstanding practices. Respondents identified exigencies related to funding/budgets and policy formulation that drove discordant recommendations. They were concerned that conditional recommendations will be ignored, or that strong recommendations were required to ensure access to the treatment.

They feared that, particularly in resource poorer settings, conditional recommendations would be considered punitive and depriving. Related to this concern, interviewees expressed a fear that conditional recommendations would send confusing signals to the public health arena and that practitioners and policy makers might misinterpret the recommendations.

Respondents frequently felt certain regarding benefits and harms, despite ratings of low or very low quality, a clear misunderstanding of the GRADE system. At times, they were correct in this assessment of certainty (that is, the rating of low or very low certainty was incorrect). This most often occurred because panellists intuitively understood the importance of indirect evidence, but were unaware of GRADE's explicit guidance in this area. At other times, their certainty was not supported by compelling evidence, direct or indirect, but rather by reliance on past experience or anecdotal evidence.

The five methodologists identified additional important issues:

i) Regarding the *role of methodologists in the process,* the methodologists felt under attack and a sense of pushback by panelists. Often, however, this grew into appreciation as the process unfolded. Methodologists also reported that they had to continuously police the recommendations against what they perceived as a constant panel goal of changing the conditional grading to strong.

ii) Regarding *contributors to discordant recommendations,* methodologists shared the experience of other interviewees that a limited prior exposure to GRADE, limited

understanding of GRADE, political agenda pressures, and contextual factors all played a role in driving discordant recommendations. They put greater emphasis, however, on conflict of interest related to funding, intellectual conflict, and political issues within the panels. By the latter, we mean a quid pro quo whereby panelists would acquiesce to recommendations with which they were not completely comfortable in return for ensuring repeated invitations to Geneva. Time urgency in terms of WHO requiring the guidelines to be completed in a short period of time, thus placing less emphasis on the objective systematic process, was a concern.

iii) *Strategies for improvement* included addressing the roles, responsibilities, and expectations of methodologists and especially as they function in a co-chair capacity (they were sometimes unaware that they were co-chairs rather than simply methods experts). Methodologists also agreed that additional panellist GRADE training was necessary, and suggested the need for tools that could aid methodologists when panels insist on discordant recommendations inconsistent with GRADE guidance. They suggested more effective use of the remarks section that accompanies each recommendation.

Table 1: Paradigmatic situations in which panels may reasonably offer (optimally made) strong recommendations on the basis of low or very low confidence in effect estimate.

Paradigmatic situation	Confidence in effect-estimates for health outcomes (Quality of evidence)		Balance of benefits and harms	Values and Preferences	Resource considerations	Recommen-dation	Example of discordant recommendations from WHO guidelines*	Frequency emerging from the Phase II study (%)
	Benefits	Harms						
Life threatening situation	Low or very low confidence	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation. Adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits	Strong recommendation in favor	In the treatment of patients with MDR-tuberculosis, a fluoroquinolone should be used.[22]	7/160 (4.4)
Uncertain benefit, certain harm	Low or very low	High or Moderate	Possible but uncertain benefit. Substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	Possible high incremental cost (or resource use) may further mandate a recommendation against the intervention	Strong recommendation against (or in favor of the less harmful/less expensive alternative when two are compared)	*No example from the present study of WHO guidelines; we include an example from elsewhere.[23] We recommend against screening for androgen deficiency in the general population.	0 (0.0)
Potential equivalence, one option clearly less risky or costly	Low or very low	High or Moderate	Magnitude of benefit apparently similar - though uncertain - for alternatives. We are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits, may further support recommendation for less harmful alternative	Strong recommendation for less harmful/less expensive	For management of post partum haemorrhage, oxytocin should be preferred over ergometrine alone, a fixed-dose combination of ergometrine and oxytocin, carbetocin, and prostaglandins.[24]	8/160 (5.0)
High confidence in similar benefits, one option potentially more risky or costly	High or Moderate	Low or very low	Established that magnitude of benefit similar for alternative management strategies. Best (though uncertain) estimate is that one alternative has appreciably greater harm.	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) of one alternative	Strong recommendation against the intervention with possible greater harm or cost	*No example from the present study of WHO guidelines; we include an example from elsewhere.[25] In women requiring anticoagulation and planning conception or in pregnancy, the AT9 guidelines recommended against the use of certain anticoagulants. For example, high confidence estimates suggests similar effects of different	0 (0.0)

							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).	
Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) of potentially harmful intervention may further justify recommendation for less harmful.	Strong recommendation against the intervention (or in favor of the less harmful/less expensive alternative when two are compared)	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.[26]	10/160 (6.3) Total=25/160 (15.6)

Table 2: Recommendations in which panels were judged to have made recommendations not consistent with GRADE guidance.

Situation	Condition	Examples from the Phase II WHO study	Frequency from Phase II, n (% of 160)
1	Good practice statement (panel should not apply GRADE methods). A large body of difficult to summarize indirect evidence indicates that the desirable consequences of the intervention are much greater than the undesirable consequences (i.e. confidence is actually high, but summarizing the evidence systematically would be a poor use of a panel's resources).	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 promptly undergo the appropriate investigations.	29 (18.1)
2	Misclassification of the evidence as low or very low when in fact it should have been medium or high	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.	33 (20.6)
3	Recommendations inconsistent with GRADE guidance (guidance suggests weak recommendations)	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.	73 (45.6)

Table 3: Judgements of the extent of explicitness of comparators in WHO recommendations and guidelines

Situation	**Recommendation example**	**Frequency, n (% of 160)**
Comparator explicit in the recommendation	The expert panel recommends cryotherapy over no treatment (strong recommendation, very low quality of evidence).	28 (17.5)
Comparator not explicit in the recommendation but explicit or clear in associated guideline text	High-dose vitamin A supplementation is recommended in infants and children 6–59 months of age in settings where vitamin A deficiency is a public health problem (strong recommendation, low quality evidence) The associated text made it clear that the recommendation was for high dose over low dose vitamin A.	43 (26.9)
Comparator not explicit in the recommendation and the associated text also failed to clarify	Offer and promote postpartum and post-abortion contraception to adolescents through multiple home visits and/or clinic visits to reduce the chances of second pregnancies among adolescents (strong recommendation, very low quality of evidence). Neither the recommendation nor the associated text made it clear whether the comparator was no offer or promotion of post-abortion contraception, or less intense or alternative programs.	89 (55.6)

Recommendations

General Comments:

Despite strong support for GRADE among many panellists and panel Chairs, we identified serious discomfort with use of GRADE among many panellists. This was clear from the responses of the Guideline chairs, but became even more evident from the comments of the GRADE methodologists who perceived a degree of hostility from the panellists and a necessity to act as "police" to ensure adherence to GRADE guidance. It is clear from the two major problems we identified – failure to specify comparators clearly, and lack of adherence to GRADE guidance regarding discordant recommendations – that any such policing for the guidelines on which we focused was to a considerable extent unsuccessful.

To us, this leaves a fundamental choice for WHO. The first possibility is that WHO make fundamental changes to their approach to guidelines, and move away from GRADE toward approaches with which panellists will be more comfortable. The latter is to recommit to GRADE and take steps to ensure that WHO recommendations adhere to GRADE guidance.

If WHO takes the second approach it would mean that modifications to GRADE guidance would be largely cosmetic. We identified one example of where changes might be desirable. This came not from the formal research project, but from the experience of one of the investigators in that project, Rebecca Stoltzfus. When, regarding a proposed

discordant public health recommendation, she challenged a WHO panel that she was chairing and showed them the GRADE guidance regarding implications of strong and conditional recommendations for policy-makers. This includes the following: *For policymakers, the implication of a conditional recommendation is that policy making will require substantial debate and involvement of many stakeholders.* The panel responded that one always needs the involvement of many stakeholders whatever the strength of the recommendation – and went ahead and gave a discordant recommendation. This suggests limitations in current GRADE characterization of the implications of strong and conditional recommendations for policy makers, and the possible desirability of a rewording.

In the remainder of this document we will assume that WHO will recommit to GRADE and will consider modifications in their process to ensure recommendations are consistent with GRADE guidance. We will separately identify goals and then thoughts as to how the goals might be achieved.

Goals

1) Ensuring explicit comparators

When making discordant recommendations (or any recommendation using the GRADE framework), panelists must ensure that the comparator in the recommendation is either clearly stated or obvious. When there is any doubt about the matter, panels should explicitly state the comparator in the recommendation.

2) Appropriate use of good practice statements

Currently, WHO guideline panels are issuing discordant recommendations that would be more appropriate as good practice statements. Leading members of the GRADE working group have recently suggested guidance for use of good practice statements.[27] This guidance, which includes detailed cautions regarding overuse of good practice statements, may be of use to WHO panels.

3) Appropriate rating of confidence

Guideline panellists often felt confident regarding the effects of interventions despite their ratings of low or very low quality. Since by quality GRADE means confidence in estimates of effect, this situation represents an unequivocal misunderstanding of GRADE. In many instances, however, panellists' intuitive assessment of the certainty of the evidence was accurate, and made on the basis of indirect evidence. As stated in the prior goal, in some such situations they would best have issued good practice statements.[27] In others, they would best have conducted a formal GRADE assessment and rated the confidence moderate or high using GRADE's explicit guidance on use of indirect evidence. Ensuring panellists understanding of the role of indirect evidence would represent an important forward step for WHO guidelines.

4) Diminished impact of competing influences

Interviews identified a number of factors that led to discordant recommendations inconsistent with GRADE guidance. These included an excessive value placed in personal experience; panellists' concerns regarding attitudes of Ministries of Health; panellists' reluctance to make conditional recommendations regarding long-established practices; panellists' concerns that conditional recommendations will be ignored; and financial and non-financial conflict of interest. Optimal recommendations will require minimizing these influences.

5) Optimal Definition of the Methodologist Role on the Committee

At least at this point, to ensure methodologic rigor and adherence to GRADE guidance, WHO guideline panels need, as a participant in the guidelines, an expert methodologist with skills in group process. The role of the methodologist should be well defined and, if the role is co-chair, the process of sharing the chair with the content expert should be well defined.

Possible Strategies for Achieving Goals

1) Education for Panellists

Many panellists have a limited understanding of GRADE and greater education/training on GRADE principles could remedy this problem. Education should focus on developing PICO questions, the meaning of strong and conditional recommendations (which may need modification), and the nature and value of indirect evidence and

observational study evidence. Educational materials should make extensive use of examples. Interactive videos would likely be popular for many, but possibly not all, panellists. The development of educational materials should involve panellists, and there should be repeated iterative feedback. Having the panel meet online before the face-to-face meeting with the specific goal of sorting out process and methodology issues may be of use.

2) Enhancing Methodologist Contributions

Methodologists' roles require clarification. Are they to act as consultants or advisors, or a co-chairs, and if the latter how are they to function in relation to the content expert co-chair? WHO staff need to collect feedback on the function of the methodologists, and methodologists must be aware of this feedback. This process should be placed in the context of the goals of improving the methodologists' performance in meeting the needs of the participants, as well as ensuring optimal recommendations consistent with GRADE guidance. Finally, it may be useful to have training materials, and training sessions, for methodologists. This would include the development of strategies to deal with situations when panels are inclined to make discordant recommendations inconsistent with GRADE guidance.

3) Choice of Content Chairs

Choice of panellists is important (see 4) but choice of panel Chairs is particularly important. Along with skills in managing a group and facilitating their interaction,

content Chairs should have as good an understanding of GRADE as possible, and a commitment to work with methodologists to ensure ratings of certainty and strength of recommendations are consistent with GRADE guidance.

4) Choice of Panellists

Our interviews suggested there are panellists who do not believe in the usefulness or appropriateness of conditional recommendations, and who are negative, and perhaps hostile, to the use of GRADE. Such individuals should not - if WHO is committed to the use of GRADE - sit on guideline panels. Agreement with the fundamental principles of GRADE, including the usefulness of conditional recommendations, should be a requirement for participation.

5) Conflict of Interest

Several methodologists, and at least one content Chair, highlighted concerns with conflict of interest. This suggests further rigor in addressing conflict of interest would be desirable. Attention should be focused on financial conflicts, intellectual conflicts, but also on the possible dynamic of agreement with particular positions and the likelihood of being invited back to subsequent panels.

6) Use of Established Paradigmatic Situations when Discordant Recommendations are Warranted

Considerable thought has gone into defining and refining characterization of situations when discordant recommendations are appropriate. GRADE has thus far defined five

such situations (Table 1).[9] It is likely that insisting that panellists, when they make discordant recommendations, define which of these five situations applies, would be useful. This is likely both to limit the number of such recommendations, and also to assist panels in the discussion of their recommendations, to make clear the rationale for discordant recommendations.

7) Matching Time Available to the Magnitude of the Task

It may be that the time available for developing recommendations is less than required for a thorough process and the development of trustworthy guidelines. If this is indeed the case, either decreasing the scope of a guideline associated with a particular meeting, or extending the time of the meeting, would be desirable.

8) Evidence to Recommendation/Decision Tables

GRADE has recently developed evidence to recommendation/decision tables (EtD).[28] The initial experience with these tables has been positive, particularly with organizations making public health recommendations. In our interviews, panellists repeatedly mention the wide variety of factors that may impact on WHO recommendations. The explicit acknowledgment of these factors through evidence to recommendations/decision tables, and then in the remarks section of the guideline, may be helpful for moving toward more trustworthy WHO guidelines.

9) Clearly identifying the target audience

Typical target audiences for WHO recommendations are clinicians and policy-makers, most often governmental policy-makers. WHO panels should make it clear to themselves and to others who is the target audience for each recommendation. For recommendations in which both groups are target audiences they may consider issuing separate recommendations for each group.

References:

1. World Health Organization (WHO). WHO guidelines approved by the Guidelines Review Committee. Accessed from url: http://www.who.int/publications/guidelines/en/; May 29th 2014.

2. Graham R, Mancher M, Wolman DM, Greenfield S, Steingerg E. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Institute of Medicine. Clinical Practice Guidelines We Can Trust. 1st ed. Washington, DC: National Academies Press; 2011.

3. Institute of Medicine (IOM). Standards for Developing Trustworthy Clinical Practice Guidelines. Standard 2. Management of Conflict of Interest. url: http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust/Standards.aspx (Accessed on January 3rd 2015).

4. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet* 2007;369: 1883-9.**

5. Hill S, Pang T. Leading by example: a culture change at WHO. Lancet; 2007: 369: 1842–1844.

6. WHO Handbook for Guideline Development. 2012 url: http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf (Accessed on April 25th 2013).

7. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. ***J Clin Epidemiol* 2011; 64(4):380-2.**

8. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. ***J Clin Epidemiol.* 2011; 64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. Epub 2011 Jan 5.**

9. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M,Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G.
GRADE guidelines 15: Going from evidence to recommendation-determinants of

a recommendation's direction and strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

10. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. ***BMJ*** **2008; 336: 924–926.**

11. WHO. Emergency response to antimalarial drug resistance. url: http://www.who.int/en/ (Accessed on April 24th 2013).

12. WHO. WHO guidelines approved by the Guidelines Review Committee. url: http://www.who.int/publications/guidelines/en/index.html (Accessed on May 20th 2014).

13. GRADE Working Group. url: http://www.gradeworkinggroup.org/ (Accessed on November 6th 2014).

14. WHO. Guidelines for the screening, care and treatment of persons with hepatitis C infection. url: http://apps.who.int/iris/bitstream/10665/111747/1/9789241548755_eng.pdf?ua=1&ua=1 (Accessed on November 5th 2014).

15. WHO. Guidelines for the identification and management of substance use and substance use disorders in pregnancy. ISBN 978 92 4 154873 1. url: http://apps.who.int/iris/bitstream/10665/107130/1/9789241548731_eng.pdf?ua=1 (Accessed on November 12th 2014).

16. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. ***J Clin Epidemiol.*** **2011 ;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026. Epub 2010 Dec 31.**

17. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y, Schünemann HJ. GRADE guidelines: 2. Framing the question and deciding on important outcomes. ***J Clin Epidemiol.*** **2011; 64(4):395-400. doi: 10.1016/j.jclinepi.2010.09.012. Epub 2010 Dec 30.**

18. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. Guideline development at the World Health Organization: an evaluation. ***PLoS One*** **2013; 8(5):e63715.**

19. 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

20. Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014;67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

21. Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, Stoltzfus R, Montori VM, Norris SL, Schünemann HJ, Guyatt GH. World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. ***J Clin Epidemiol.*** **2014; In Press.**

22. WHO. Guideline: Guidelines for the programmatic management of drug-resistant tuberculosis. Geneva, World Health Organization (WHO), 2011 (url: http://whqlibdoc.who.int/publications/2011/9789241501583_eng.pdf?ua=1).

23. Bhasin S, Cunningham GR, Hayes FJ, et al. Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. ***J Clin Endocrinol Metab.*** **2010;95:2536–2559.**

24. WHO. Guideline: WHO guidelines for the management of postpartum haemorrhage and retained placenta. Geneva, World Health Organization (WHO), 2009 (url: http://whqlibdoc.who.int/publications/2009/9789241598514_eng.pdf).

25. Bates SM, Greer IA, Middeldorp S, Veenstra DL, Prabulos AM, Vandvik PO; American College of Chest Physicians. VTE, thrombophilia, antithrombotic therapy, and pregnancy: Antithrombotic Therapy and Prevention ofThrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. ***Chest.*** **2012;141(2 Suppl):e691S-736S. doi: 10.1378/chest.11-2300.**

26. WHO. Guideline: Rapid Advice. Treatment of tuberculosis in children. Geneva, World Health Organization (WHO), 2010 (url: http://whqlibdoc.who.int/publications/2010/9789241500449_eng.pdf).

27. Guyatt GH, Schünemann HJ, Djulbegovic B, Akl EA. Guideline panels should not GRADE good practice statements. ***J Clin Epidemiol.*** **2014; In Press.**

28. Evidence to Decision (EtD) tables. Framework for clinical guidelines' development. url: http://www.slideshare.net/CharlieNeck/etd-20140327 (Accessed on December 26th 2014).

Authors' contributions:

Mr. Alexander contributed principally to the conception of this guidance chapter and drafting. Dr. Guyatt provided relevant guidance.

CHAPTER 9:

Discussion and summary conclusion

The understanding of when and why strong recommendations are formulated at WHO when the evidence is of low quality is an important methodological issue that has implications not just for WHO, but for a wide variety of guideline development groups internationally.

This thesis builds on two streams of prior research work: i) concerns about WHO guidelines[1-3] and ii) evidence of excessive use of strong l/vl or discordant recommendations by guideline panels.[4]

For example, an earlier study[4] that supported our Phase I WHO findings[1] involved a formal structured exploration of discordant recommendations using the GRADE approach. The study found that of 357 recommendations produced by the Endocrine Society,[4] 206 (58%) were strong and of these, 121 (59%) were strong l/vl or discordant recommendations. Of the strong l/vl, 35 (29%) were consistent with GRADE guidance; 36% were best practice statements inappropriate for grading; 4% were recommendations for additional research (also not subject to grading); 4% were recommendations in which moderate rather than low confidence was actually warranted; and 27% were strong recommendations that did not fit the existing paradigmatic situations, and thus were likely not consistent with GRADE guidance. These results are very similar to the WHO findings presented in this thesis,[1,5] suggesting that the concerns that arise from our thesis investigation are not limited to WHO guidelines.

This thesis set out to expand on the prior empirical work[4] in understanding when and why strong recommendations are formulated at WHO when the evidence is of low quality. In addressing the research concerns for this thesis, we conducted four studies:

Phase I: A descriptive epidemiological study[1] that focused on WHO GRC approved guidelines produced from 2007 to 2012 and which used GRADE methods.

Phase II: An expansion of the Phase I study via a taxonomic assessment[5] of the strong l/vl (and based on a previously developed taxonomy employed in the Endocrine Society study[4]) to establish consistency or inconsistency with GRADE guidance.

Phase III: A qualitative descriptive study whereby we dove deeper and sought interviews with guideline panel members involved in the strong l/vl that were inconsistent with GRADE guidance and based on the Phase I findings and Phase II taxonomic examination. The qualitative interviewing strove to gain an understanding of the factors WHO panelists considered (and process involved) when making strong l/vl.

Phase IV: An additional qualitative descriptive study whereby we interviewed senior GRADE methodologists to further our understanding of the guideline development process and the crafting of discordant recommendations from the perspective of senior GRADE methodologists. These methodologists were not recruited as part of the Phase III interview study but were deemed valuable to our understanding of reasons for strong l/vl made by WHO panelists.

Phase I findings (chapter three):

Based on the Phase I descriptive analysis, we confirmed the likelihood of excessive use of strong recommendations by WHO panels and specifically strong l/vl (chapter 3). We found that 63.4% of examined recommendations were strong and of these, 55.5% were strong l/vl.[1] Furthermore, while emerging across WHO guidelines in

general, strong l/vl are particularly frequent in certain content areas including maternal and reproductive health, child health, HIV/AIDS, and tuberculosis.[1] Finding that these four WHO content areas had a very similar frequency of strong l/vl or discordant recommendations, approximately 50%, suggests that the phenomenon is a systemic issue across guidelines developed by WHO.

Phase II findings (chapter four):

In expanding the Phase I analysis, we confirmed the excessive use of strong l/vl frequently inconsistent with GRADE guidance (chapter four). From the Phase II taxonomy study,[5] we found only 25/160 (15.6%) were judged consistent (optimally made) with one of the five paradigms for appropriate recommendations;[6] the remaining 135/160 (84.4%) being made inconsistent (sub-optimally made) with GRADE guidance: 33 (21%) were based on evidence warranting moderate or high confidence in the estimates of effect; 29 (18%) would have been best framed as good practice statements;[7] and 73 (46%) warranted a conditional, rather than a strong recommendation.[5]

Phase III findings (chapter six):

We sought to understand the reasoning behind the high proportion of strong l/vl that were inconsistent with GRADE guidance. We found explanations from the qualitative interview study[8] that included a general skepticism about the value of making

conditional recommendations, and political considerations, funding, policy, as well as being wedded to long-standing practices as drivers of strong l/vl.

Phase IV findings (chapter seven):

We found additional explanations for strong l/vl via interviews with senior GRADE methodologists[9] that substantiated the Phase III panel interview findings and added additional layers that included a perceived tension between methodologists and WHO panelists (as a result of panel members' resistance to adhering to GRADE guidance) and the presence of both financial and non-financial conflicts of interest among panel members. An unclear role for methodologists also emerged as a real concern.

WHO Guidance when making strong l/vl (chapter 8):

This thesis confirms that WHO makes a large number of their recommendations as strong l/vl[1] that are inconsistent with GRADE guidance[5] and make it clear that organizations producing guidelines and using GRADE (and indeed using any system that differentiates between "just do it" recommendations and those that are value and preference sensitive) must be alert to the danger of excessive use of discordant recommendations and take preventive action. WHO and other guideline developers using GRADE need to make clear policies regarding the extent to which they will adhere to GRADE guidance. If they do adhere, many panels may require further training on

GRADE principles and their application (e.g. focused training on developing PICO questions, the meaning of strong and conditional recommendations, and the nature and value of indirect evidence and observational study evidence). Our findings strongly suggest that WHO sponsored guideline development groups would benefit from increased GRADE training and awareness of the need to clearly specify not only interventions but comparators in the recommendations, of the five situations in which strong l/vl may be warranted,[6] and of the inadvisability of strong recommendations when these criteria are not met.

We therefore conclude the thesis with a guidance document. This guidance report has been commissioned on request by WHO (chapter eight) and provides suggestions of what guideline panels can do about the problem of making strong l/vl that may be inconsistent with GRADE guidance. We offer the guidance that is targeted for WHO, but the following are generalizable for any guideline panel and we separately identify goals and then thoughts as to how the goals might be achieved:

Goals

1) Ensuring explicit comparators

When making discordant recommendations (or any recommendation using the GRADE framework), guideline panelists must ensure that the comparator in the recommendation is either clearly stated or obvious. When there is any doubt about the matter, panels

should explicitly state the comparator in the recommendation and ensure that it is locatable within the background guideline document itself.

2) Appropriate use of good practice statements

Currently, WHO guideline development groups are issuing discordant recommendations that would be more appropriate as good practice statements.[7] Good practice statements typically represent situations in which a large body of indirect evidence, made up of linked evidence including several indirect comparisons, strongly supports the net benefit of the recommended action. The GRADE Working Group has guidance for the use of good practice statements,[7] including detailed cautions regarding overuse of such statements. The cautions, framed as questions that the guideline development groups and WHO staff should ask themselves before deciding to issue a good practice statement, are as follows:

> *Is the statement clear and actionable?*
> *Is the message really necessary?*
> *Is the net benefit large and unequivocal?*
> *Is the evidence difficult to collect and summarize?*
> *Are there specific public health issues that should be considered (e.g., equity)*
> *Is the rationale explicit?*
> *Should the quality of the evidence be formally assessed using GRADE?*

3) Appropriate rating of confidence

Guideline panelists may feel confident regarding the effects of interventions despite their ratings of low or very low quality. Since by quality GRADE means confidence in estimates of effect, this situation represents an unequivocal misunderstanding of GRADE. This may also be due to panelists' lack of understanding of the role of indirect

evidence. As stated in the prior goal, in some such situations these would be best issued as good practice statements.[7] In others, panels should conduct a formal GRADE assessment and rate the confidence as moderate or high using GRADE's explicit guidance on use of indirect evidence.

4) Diminished impact of competing influences

Our research has identified a number of factors that can lead to strong l/vl inconsistent with GRADE guidance. These may include an excessive value placed on personal experience; political considerations; panelists' reluctance to make conditional recommendations regarding long-established practices; funding and policy considerations; panelists' concerns that conditional recommendations will be ignored; and financial and non-financial conflicts of interest. Optimal recommendations will require recognizing that these influences may play a role and thus minimizing them.

5) Optimal Definition of the Methodologist Role on the Committee

To ensure methodologic rigor and adherence to GRADE guidance, guideline panels need, as a participant in the guideline development process, an expert methodologist with skills in GRADE as well as the group process/leadership experience. The role of the methodologist should be well defined, at both WHO and elsewhere.

Possible Strategies for Achieving Goals

1) Enhancing Methodologist Contributions

It may be useful to have training materials, and training sessions, for methodologists. This would include the development of strategies to deal with situations when panels are inclined to make strong l/vl inconsistent with GRADE guidance.

2) Choice of Content Chairs

Choice of panelists is important (see 4) but choice of guideline panel lead or Chair is particularly important. Along with skills in managing a group and facilitating their interaction, content Chairs should have as good an understanding of GRADE as possible, and a commitment to work with methodologists to ensure ratings of confidence/certainty and strength of recommendations are consistent with GRADE guidance.

3) Choice of Panelists

Our research suggests that there are panelists who do not believe in the usefulness or appropriateness of conditional recommendations, and who are negative to the use of GRADE. This is a concern and such individuals should not – if the guideline developer/sponsor is committed to the use of GRADE - sit on guideline panels. Agreement with the fundamental principles of GRADE, including the usefulness of conditional recommendations, should be a requirement for participation.

4) Conflict of Interest

Our research from the interview studies indicates concerns with conflicts of interest. Attention should be focused on both financial conflicts and intellectual conflicts.

5) Use of Established Paradigmatic Situations when Discordant Recommendations are Warranted

Considerable thought has gone into defining and refining characterization of situations when strong l/vl are appropriate. GRADE has thus far defined five such situations.[6] It is likely that insisting that panelists, when they make strong l/vl, define which of these five situations applies, would be useful. This is likely both to limit the number of such recommendations, and also to assist panels in the discussion of their recommendations, to make clear the rationale for discordant recommendations (increase transparency). This will be useful for a variety of stakeholders.

6) Evidence to Recommendation/Decision Tables

GRADE has recently developed evidence to recommendation/decision tables (EtD).[10] The initial experience with these tables has been positive, particularly with organizations making public health recommendations. Public health recommendations may involve a wide array of factors that include equity, accessibility, acceptability, feasibility, and resource use. The explicit acknowledgment of these factors through evidence to recommendations/decision tables, and then in the remarks section of the guideline will increase the transparency of and trust in the guideline.

7) Clearly identifying the target audience

The target audiences for WHO recommendations are typically health workers, programme managers and policy-makers. WHO-sponsored guideline development groups should make it clear to themselves and to the end-user of the guideline, who is the target audience for each recommendation. For recommendations in which both groups are

target audiences they may consider issuing separate recommendations for each group. A panel issuing different recommendations for different groups risk confusing both audiences; in general, it is likely best to avoid such situations and, when used, to provide extremely explicit rationale.

Conclusion/summary

This thesis study is the first to have focused on the issue of guideline developers, in this case WHO panelists, making strong recommendations when the underlying confidence or study quality is low or very low. Though such recommendations are to be made with caution, GRADE guidance does support the issuance of such recommendations and outlines instances when they are indeed appropriate.[6]

The Phase I-IV studies indicate that WHO guideline panels make more than one-half of their recommendations as strong, and a majority of them are strong l/vl. Our detailed analysis throughout the thesis indicates that there are major limitations in the extent to which WHO sponsored guideline development groups adhere to GRADE guidance in issuing strong l/vl or discordant recommendations. Prior evidence suggests that this is also true of other guideline panels[4] whereby panel members may prefer to issue strong "just do it" recommendations even when uncertainties in the evidence warrant a more cautious approach. This tendency may be equally prevalent in panels that do not use GRADE – the explicitness of the GRADE process may make the problem more

apparent. Pending further exploration, guideline panels should be alert to the potential for their recommendations to go beyond what the available evidence warrants.

The Phase III panelist interviews[8] and Phase IV GRADE methodologist interviews[9] reveal strong support for the structured, systematic, and transparent GRADE approach among many WHO panelists while at the same time identifying serious discomfort with use of GRADE. This could be due in large part to a lack of GRADE background or training and could thus be remedied with suitable GRADE training at the start of the guideline development process, panel preparation prior to arrival for guideline development (can be of use in sorting out process and methodology issues), and raising awareness of and the effective use of existing GRADE training tools.

If organizations such as WHO choose to persist with the use of GRADE, these results also suggest the need for a re-commitment to GRADE principles as part of the guideline process at WHO and other organizations using GRADE. Ensuring panelists are truly committed to the GRADE process, and are free of important conflicts of interest (as addressed in Standard 2 of the IOM's Standards for Developing Trustworthy Clinical Practice Guidelines[11]), are necessary (balancing the need to utilize expert input into guideline development while mitigating the impact of intellectual and financial conflicts). Finally, education of guideline users regarding what panels mean by conditional recommendations, and how they can be applied, may be of use. It is our hope that this

thesis work will be both informative and instructive in guiding WHO panelists (and guideline developers elsewhere) regarding the issues to consider when making strong l/vl.

References

1. Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, Neumann I, Rave S, Guyatt G. World Health Organization recommendations are often strong based on low confidence in effect estimates. ***J Clin Epidemiol.*** **2014;67(6):629-634. doi: 10.1016/j.jclinepi.2013.09.020. Epub 2014 Jan 3.**

2. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World health organization guideline development: an evaluation. ***PLoS One.*** **2013; 8(5):e63715. doi: 10.1371/journal.pone.0063715. Print 2013.**

3. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. ***Lancet.*** **2007; 369(9576):1883-9.**

4. 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(8):3246-52. doi: 10.1210/jc.2013-1814. Epub 2013 Jun 19.**

5. Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, Stoltzfus R, Montori VM, Norris SL, Schünemann HJ, Guyatt GH. World Health Organization strong recommendations based on low quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. ***J Clin Epidemiol.*** **2014; In Press.**

6. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori V, Brito Campana JP, Norris S, Elbarbary M, Post P, Nasser M,Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G.
GRADE guidelines 15: Going from evidence to recommendation-determinants of a recommendation's direction and strength. ***J Clin Epidemiol.*** **2013 pii: S0895-4356(13)00054-1. doi: 10.1016/j.jclinepi.2013.02.003. [Epub ahead of print].**

7. Guyatt GH, Schünemann HJ, Djulbegovic B, Akl EA. Guideline panels should not GRADE good practice statements. ***J Clin Epidemiol.*** **2014; In Press.**

8. Phase III panel interview study; JCE has agreed to accept submission.

9. Phase IV methodologist interview study; JCE has agreed to accept submission.

10. Evidence to Decision (EtD) tables. Framework for clinical guidelines' development. url: http://www.slideshare.net/CharlieNeck/etd-20140327 (Accessed on December 26th 2014).

11. Institute of Medicine (IOM). Standards for Developing Trustworthy Clinical Practice Guidelines. Standard 2. Management of Conflict of Interest. url: http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust/Standards.aspx (Accessed on January 3rd 2015).

Authors' contributions:

Mr. Alexander contributed principally to the conception and writing of the final chapter. Dr. Guyatt provided guidance and sincere gratitude goes to Dr. Guyatt for invaluable guidance throughout the thesis study. Similar gratitude is extended to the larger WHO GRADE Guidelines Project research team.