BMJ Global Health

Institutionalising an evidence-informed approach to guideline development: progress and challenges at the World Health Organization

Unni Gopinathan,[1,2,3] Steven J Hoffman[3,4,5]

To cite: Gopinathan U, Hoffman SJ. Institutionalising an evidence-informed approach to guideline development: progress and challenges at the World Health Organization. *BMJ Glob Health* 2018;**3**:e000716. doi:10.1136/bmjgh-2018-000716

Handling editor Stephanie M Topp

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjgh-2018-000716).

Received 10 January 2018
Revised 28 June 2018
Accepted 29 June 2018

For numbered affiliations see end of article.

Correspondence to
Dr Unni Gopinathan;
unni.gnathan@gmail.com

ABSTRACT

This study explored experiences, perceptions and views among World Health Organization (WHO) staff about the changes, progress and challenges brought by the guideline development reforms initiated in 2007. Thirty-five semistructured interviews were conducted with senior WHO staff. Sixteen of the interviewees had in-depth experience with WHO's formal guideline development process. Thematic analysis was conducted to identify key themes in the qualitative data, and these were interpreted in the context of the existing literature on WHO's guideline development processes. First, the reforms were seen to have transformed and improved the quality of WHO's guidelines. Second, independent evaluation and feedback by the Guidelines Review Committee (GRC) was described to have strengthened the legitimacy of WHO's recommendations. Third, WHO guideline development processes are not yet designed to systematically make use of all types of research evidence needed to inform decisions about health systems and public health interventions. For example, several interviewees expressed dissatisfaction with the insufficient attention paid to qualitative evidence and evidence from programme experience, and how the Grading of Recommendations Assessment, Development and Evaluation (GRADE) process evaluates the quality of evidence from non-randomised study designs, while others believed that GRADE was just not properly understood or applied. Fourth, some staff advocated for a more centralised quality assurance process covering all outputs from WHO's departments and scientific advisory committees, especially to eliminate strategic efforts aimed at bypassing the GRC's requirements. Overall, the 'culture change' senior WHO staff called for over 10 years ago appears to have gradually spread throughout the organisation. However, at least two major challenges remain: (1) ensuring that all issued advice benefits from independent evaluation, monitoring and feedback for quality and (2) designing guideline development processes to better acquire, assess, adapt and apply the full range of evidence that can inform recommendations on health systems and public health interventions.

Key questions

What is already known?

- Previous studies of WHO's guidelines have primarily focused on the outputs of the guideline development processes, while the experience of WHO staff with these processes have received relatively less attention.

What are the new findings?

- WHO's Guidelines Review Committee—which oversees formal guideline development at WHO—is a key institutional mechanism for securing independent evaluation, monitoring and feedback.
- WHO's guideline development processes (especially about complex interventions) are not yet designed to systematically make use of all types of relevant research evidence.
- WHO's technical departments and scientific advisory committees still issue guidance without subjecting these to WHO's formal guideline development processes.

What do the new findings imply?

- WHO should promote systematic sharing of experiences and learning among its departments, systematically apply the full range of research evidence and consider centralised quality assurance processes for all products with normative content.
- Other technical health agencies and institutions could learn from WHO's mechanisms for ensuring independent evaluation, monitoring and feedback for process and quality in guideline development.

INTRODUCTION

In 2007, the World Health Organization (WHO) embarked on far-reaching reforms of its guideline development process. The major driving factor was a study published by Oxman, Lavis and Fretheim in *The Lancet* that identified at least four flaws in the agency's guideline development processes.[1] First, systematic and transparent methods for retrieving, appraising, synthesising and interpreting evidence were rarely used. Second, WHO's guideline development processes rarely involved methodologists or representatives of populations affected by the

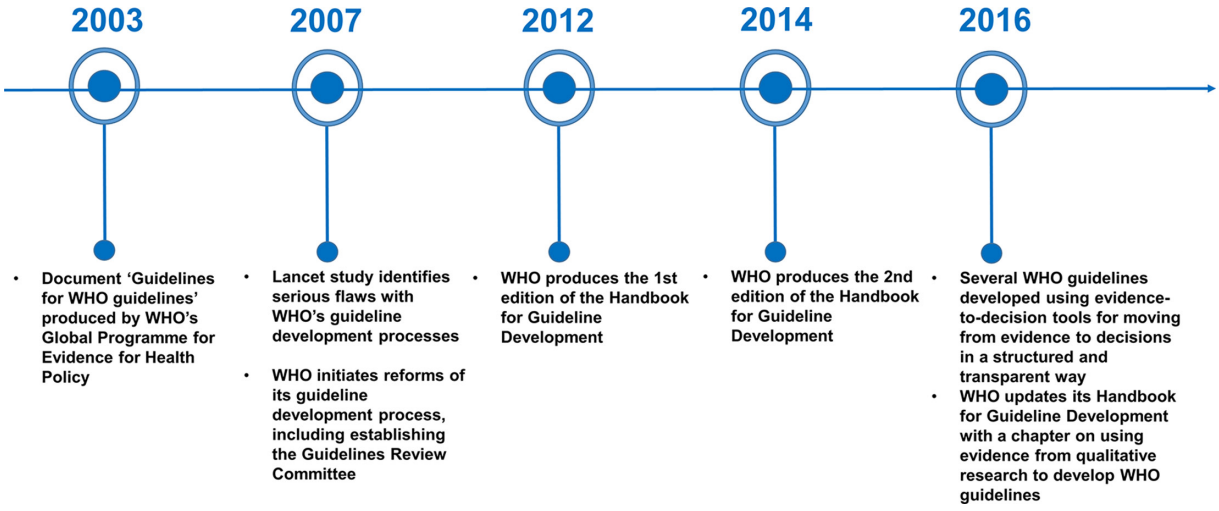

Figure 1 Key events and changes during the evolution of WHO's guideline development process.

recommendations. Third, limited efforts were made to adapt global recommendations to local needs, conditions, resources, costs and values. Fourth, dissemination and implementation strategies, and rigorous evaluations of these, were largely absent.

WHO quickly acknowledged the criticism and promised an immediate response.[2] Several key changes to strengthen WHO's guideline development process followed (figure 1). A new guideline development process was set up to involve different groups—internal and external to the agency—with specified roles and responsibilities (box 1). To ensure a transparent and evidence-informed decision-making process for every formal guideline issued by the agency, WHO established a Guidelines Review Committee (GRC). The GRC was set up in 2007 as a scientific oversight committee tasked with independently reviewing all guideline proposals from WHO's departments prior to initiating the guideline development process and to approve the final guidelines once completed. An essential part of ensuring transparency and rigour was prioritising the use of systematic reviews to inform guideline development and the use of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) tool.[3] The GRADE approach represents a systematic and transparent framework for appraising the quality of the evidence informing recommendations and making judgements about the strength of the recommendations. Two editions of a *Handbook for Guideline Development* have also been produced.[4 5]

It is now over 10 years since WHO implemented its guideline development reforms. Previous studies of WHO's guidelines have focused on the outputs of guideline development processes, including assessing the quality of the evidence underlying recommendations,[6–10] identifying explanations for issuing discordant recommendations[11 12] and identifying challenges with implementing the GRADE approach.[9–11 13–15] Only one study, published 5 years ago, focused on examining the views and perceptions of WHO staff about the reforms.[9] That study identified mixed views about whether WHO needed a single quality assurance mechanism, and uncertainties and lack of capacity among staff in applying the GRADE approach, concluding that quality assurance standards set by the GRC were yet to be fully embedded within the agency.[9] This study builds on previous studies and was conducted as part of a larger research project examining the design and effectiveness of WHO's scientific advisory committees. Specifically, the research objectives of this study were to investigate WHO staff's experience with, perceptions of and views about the changes brought by WHO's guideline development reforms, the progress and impact of these reforms to date, and key challenges that need to be addressed.

METHODS

Study design

This study used a qualitative study design involving primary data collection consisting of semistructured interviews with senior WHO staff. The analytical approach involved identification and interpretation of themes in the qualitative data describing key experiences, perceptions and views about WHO's guideline development reforms. This study did not operate with explicitly defined theoretical frameworks to guide the scope of the study, the specification of the research questions, or support interpretation; instead, the major themes identified in the qualitative data were interpreted in the context of the existing literature on WHO's guideline development process and guideline development more generally.

To identify major themes in the qualitative data, we followed the five-cycle process described by Yin (described in greater detail in the Data analysis section).[16] Yin does not explicitly describe this process as 'thematic analysis'; yet this analytical process involves identification of themes capturing 'something important about the data in relation to the research question',[17] and representing an 'idea or concept that captures and summarises the core point of a coherent and meaningful pattern in the data'.[18] Accordingly, the methodological approach described

Box 1 Key features of WHO's guideline development process*

What is a WHO recommendation?

WHO describes a recommendation to tell 'the intended end-user of the guideline what he or she can or should do in specific situations to achieve the best health outcomes possible, individually or collectively' and to offer 'a choice among different interventions or measures having an anticipated positive impact on health and implications for the use of resources'.

Furthermore, WHO's handbook emphasises that 'guidelines must have a clearly defined target audience (end-user) which is identified early in the guideline development process' and that 'the recommendations need to be tailored to that audience'. The target audience of guidelines are typically either policy-makers, managers in the health sector or health professionals. WHO recommendations help end-users 'make informed decisions on whether to undertake specific interventions, clinical tests or public health measures, and on where and when to do so, and to help the user to select and prioritise across a range of potential interventions'.

Who initiates a WHO guideline development process?

Technical departments in WHO initiate and coordinate the guideline development process. WHO generally develops guidelines in response to needs expressed by its Member States, WHO country offices, external experts or other stakeholders for guidance on a clinical or public health problem or policy area.

What type of committees are established and how are these constructed?

Four different groups are established when developing WHO guidelines: (1) a steering group, (2) a guideline development group (GDG), (3) an external review group and (4) a systematic review team.

1. The *steering group* is led by the responsible technical officer from the technical department overseeing the guideline development process. Among its responsibilities include drafting the scope of the guideline, identifying the systematic review team and guideline methodologist, and submitting the planning proposal to the Guidelines Review Committee (GRC). The steering group is usually exclusively composed of members from all WHO departments and regional offices whose work deal directly with the topic of the guideline. In cases where the guideline is developed jointly with another UN agency, individuals from that agency will also be members of the steering group.
2. The *GDG* is composed of external experts, and its central task is to develop the evidence-informed recommendations. Its composition should reflect the technical skills, geographic representation and diverse perspectives (including end-users and representatives of people and communities affected by the recommendations) needed to produce the guideline. Its membership does not include employees of WHO or other UN agencies. The GDG is established early in the process once the steering group has defined the general scope and target audience of the guideline and drafted key questions. The GDG is responsible for finalising the scope and key questions of the guideline. Members do not receive any financial compensation other than for direct expenses associated with their work on the guideline.
3. The *external review group* is composed of technical experts, end-users, programme managers, advocacy groups and individuals from communities affected by the recommendations. Similar to the GDG, the external review group should be balanced in terms of geography and gender and should involve diverse perspectives. If important perspectives and stakeholders are missing from the GDG, these should be represented in the external review group. Members of the external review group can be asked to participate in different stages of the guideline development process, such as reviewing the guideline's scope and key questions, reviewing the final guideline for errors or missing data and commenting on implications for implementation.
4. The *systematic review team* are usually external groups commissioned by WHO to undertake systematic reviews of the evidence base underpinning the key questions identified by the steering committee and the GDG. Ideally, the systematic review team is identified early and involved in formulating the key questions and establishing a reasonable scope for the guidelines.

What role does the methodologist play during the guideline development process?

WHO Handbook for Guideline Development recommends that at least one methodologist—defined as an expert in guideline development processes and methods—should be involved in the development of WHO guidelines. The methodologist should be an expert in systematic reviews, GRADE (Grading of Recommendations Assessment, Development and Evaluation) and translating evidence into recommendations. The process should recruit them early to enable their participation in planning, scoping and developing the key questions. During the GDG meetings, the methodologist supports the GDG to ensure that recommendations are informed by the evidence in a transparent and explicit manner.

What is the composition and role of the GRC?

The GRC was established by WHO's Director General in 2007 to ensure that WHO guidelines are of high quality, are developed using a transparent and explicit process, and to the extent possible, are evidence-based. The GRC is composed of approximately 30 individuals. Five of these are external, while the remaining are WHO staff from headquarters and regional offices (Susan L Norris, personal communication, 2018). WHO's Member States do not have representatives in the GRC, and the GRC members serve in their individual capacity (Susan L Norris, personal communication, 2018). The GRC meets monthly to review submitted documents. All WHO publications containing recommendations must be approved by the GRC according to WHO policies and procedures. The GRC reviews every WHO guideline at the initial planning stage and again after the recommendations have been developed and the guideline document has been finalised and edited. GRC approval is part of WHO's internal clearance processes for the publication of guidelines. The GRC is supported by the GRC Secretariat, which provides WHO staff with technical advice on guideline development, sets benchmarks and evaluates guideline development processes.

*All content was gathered from WHO Handbook for Guideline Development (Second edition).[14]

below resembles Braun & Clark's six phases of thematic analysis: (1) familiarisation with the data, (2) coding, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes and (6) writing up.[17] Moreover, the way our findings have been identified, interpreted and presented under 'Results' is broadly consistent with the

Box 2 Two general recommendations and three specific actions for WHO

Two general recommendations

- Guideline development processes in technical health agencies and institutions should learn from WHO's vast experience with implementing independent evaluation, monitoring and feedback for process and quality to ensure the legitimacy of recommendations.
- Guideline development processes at WHO should be designed to better acquire, assess, adapt and apply the full range of research evidence that can inform recommendations about health systems and public health.

Three specific actions for WHO

- WHO should foster the systematic sharing of experiences and learning among its departments that are or are planning to engage with guideline development processes so as to promote continuous professional development of its staff.
- WHO should share its experience externally (such as with the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group and its subgroups) as part of an effort to further optimise the guideline development processes to meet the needs of health systems and public health interventions (eg, complex interventions).
- WHO should consider whether outputs from scientific advisory committees that currently operate outside of the formal guideline development rules should be subject to a centralised quality assurance process.

qualitative descriptive approach described by Sandelowski.[19 20] Here, the identified themes present a comprehensive and rich summary of experiences responding to the research objectives. However, Sandelowski proposes the use of content analysis as the main analytical approach in a qualitative descriptive study, a term used interchangeably with thematic analysis, but which many consider to be two distinct ways of approaching qualitative data analysis.[21]

The COnsolidated criteria for REporting Qualitative research (COREQ) was used to report on the characteristics of the research team, study design and data analysis (online supplementary file S1).[22] COREQ was originally developed to promote transparent and comprehensive reporting of qualitative health research involving the use of interviews and/or focus groups to explore preferences and needs of clinicians, healthcare providers, policy-makers and patients. However, its relevance goes beyond research on interactions within healthcare systems since COREQ's items clarify various choices important for understanding the study design, the collection and the interpretation of qualitative data more generally.

Document review

To gain a deeper understanding of WHO's guideline development process, inform the development of the study's research objectives, and help interpret the main findings in a broader context, a comprehensive literature review was conducted to identify and summarise research articles that have previously examined WHO's experience with its guideline development reforms. PubMed, SSRN and Web of Science were searched combining the keywords 'World Health Organization' and 'guideline development'. Three additional studies not identified through this search were also included.[6 8 11] A summary of all included studies' methodologies and key findings are available in online supplementary file S2. Studies that did not evaluate the guideline development process, but only described the development of specific guidelines, were excluded.[23]

Semistructured interviews

This research project and the proposed research questions were discussed with one very senior leader within WHO, who made an initial email introduction inviting directors of technical departments and coordinators of various technical programmes to participate in the study. These directors and coordinators either agreed to participate themselves, forwarded the request to other WHO staff responsible for convening scientific advisory committees within their respective departments or did not respond. A list of 68 potential WHO interviewees was generated through purposive and snowball sampling, and all of these were invited to participate in the study. Of these 68 people, 1 was no longer at WHO, 3 felt their work was not relevant for the study, 3 declined due to busy schedules, 6 forwarded the request to other staff they deemed more suitable for the study and 14 did not respond to the invitation. In total, 41 senior WHO staff were interviewed between March and June 2016. Six of these interviews were conducted by SJH (male, PhD, lawyer) at WHO headquarters in Geneva to pilot the scope and relevance of the study, gain clarity on the range of scientific advisory committees at WHO and revise the interview questions to maximise their clarity and probative value. These pilot interviews did not intend to explore the inner workings of WHO's guideline development processes; accordingly, these six interviews were not audio recorded and transcribed, and the data was not included in the formal analysis. The remaining 35 interviews were conducted by UG (male, PhD, physician). Both authors have previous experience working with processes informing WHO guidelines and strategies,[24–26] and conducting and publishing studies involving qualitative methods to explore research questions on the role of national and international institutions in global health.[27 28]

Every participant was sent an email prior to the interview introducing the interviewer and the project, which included a concept note about the project. Thirteen interviews were conducted in-person at WHO's headquarters in Geneva and 22 interviews were conducted by telephone. The interview questionnaire (online supplementary file S3) was designed to address key questions pertaining to the design and effectiveness of scientific advisory committees. The interviews were conducted using a semistructured format, permitting some iterations to the questions depending on the response of the interviewees, and flexibility to address interesting themes emerging from responses. Participants were interviewed

individually, except for three participants who were based in the same WHO department and interviewed together. Only the interviewees were present in the room during the interviews. No repeat interviews were carried out. With informed consent, the 35 interviews were audio recorded, transcribed and anonymised. The interviews lasted between 45 and 60 min. We categorised interviewees into five main groups:

1. Directors of technical departments who had experience overseeing development of WHO guidelines and other forms of scientific advice (n=7).
2. Coordinators and team leaders of technical units within departments with responsibilities for managing guideline development processes or scientific advisory committees, including WHO's expert committees (n=19).
3. Technical officers responsible for supporting guideline development processes or other types of scientific advisory work at WHO (n=4).
4. WHO staff with other roles, such as senior advisory roles or organisational management (n=3).
5. Staff with leadership positions in programmes or partnerships hosted by WHO and with responsibilities for overseeing scientific advisory work (n=2).

Of these 35 interviewees, 16 had in-depth experience with WHO's guideline development processes, while the remaining interviewees primarily dealt with other types of scientific advisory functions at WHO (such as WHO's expert committees or scientific and technical advisory groups). On average, the interviewees had 13 years of experience working at WHO (range 0.5–32 years). Interviewees had experience working on various technical issues, including antimicrobial resistance, child health, environmental health, essential medicines and pharmaceutical policy, health workforce, HIV/AIDS, humanitarian response, immunisation and vaccine safety, maternal and reproductive health, non-communicable diseases and nutrition, polio, tobacco control, tuberculosis and social determinants of health.

Data analysis

The process for identifying major themes in the qualitative data followed the general five-cycle process described by Yin,[16] which represents an iterative approach involving five phases: (1) compiling, (2) disassembling, (3) reassembling, (4) interpreting and (5) concluding. While each step addresses specific aspects of data analysis, we moved back and forth between phases 2 and 4 as part of continuously revisiting the accuracy of initial coding and interpretation of the data. *Compilation* consisted of transcribing and further deidentifying each audio recording, including removing specific mentions of names, titles and departments, which may indirectly identify the interviewee. *Disassembling* consisted of open coding where level 1 codes, including in vivo codes, were assigned to words, phrases and larger fragments of each interview transcript. During this phase, relationships between level 1 codes from different interviews were identified and assigned level 2 codes, thereby facilitating an incremental understanding of the major themes from across the interviews. The *reassembling* phase consisted of bringing level 1 and level 2 codes together to identify themes representing a central concept and/or message that captured recurring patterns observed across the interviews. The codes and patterns deemed most relevant for the study questions were continuously refined through an iterative process and by using the constant comparison method inspired from grounded theory.[16] We observed that the last five to six interviews only introduced a few new codes and no new major themes, which we used as the indication for reaching data saturation. During the *interpretive* phase, we used the reassembled data to write a narrative around the study questions, while continuously assessing whether revisiting the disassembling and reassembling phases was needed to recompile the data. Field notes and memos documenting observations and reflections during and after the interviews (eg, whether interviewees emphasised particular aspects or the investigator's initial comparisons of issues raised in the interview with those raised in previous interviews) were considered during all stages of the analysis.

The qualitative data was transcribed and coded by one investigator. After initial identification of level 1 and level 2 codes and their associated themes by UG, the findings were discussed in detail with the second investigator (SJH) and refined until agreement was reached about the fit of the codes with the identified central themes and messages. We provided interviewees with two opportunities to comment, make suggestions or raise any concerns with our data analysis and interpretation. First, we sent each interviewee the full transcript and a 1–2 page summary of their interview. These summaries were sent approximately 1 year after the interviews were conducted. Second, we sent all interviewees an early draft of this manuscript prior to submitting it to *BMJ Global Health*. Overall, 14 of 35 interviewees confirmed receiving the summary and/or the manuscript, while the remaining interviewees did not respond. One interviewee explicitly expressed disagreement with how the findings were presented, although did not object to the content. To further validate our findings, we sent the manuscript to two WHO staff with extensive experience with the agency's guideline development processes—but who were not interviewed for this study—for their feedback. They expressed that they recognised the main findings of the study.

Four major themes pertaining to WHO's guideline development process were identified from the analysis of the interviews. The qualitative codes related to these themes are presented in table 1. During the concluding phase, the interpretation of the main empirical findings formed the basis for two main recommendations proposed under discussion.

Ethics

An exemption to ethical review was granted for this study by the University of Ottawa's Office for Research Ethics because the study was considered a programme evaluation in accordance with Canada's Tri-Council Policy Statement (TCPS2) on the Ethical Conduct for Research Involving Humans.[29]

Table 1 Major themes and corresponding qualitative codes

Themes	Level 2 codes	Level 1 codes
WHO's guideline development reforms represented a transformational shift in its approach to producing clinical and public health recommendations	Triggers of the reform	Dominance of expert opinion
	Impact of the reform	Handbook with comprehensive guidance
		Transformed WHO's guideline development process
		Institutionalisation of evidence-based principles
		More consistent use of systematic reviews
Independent evaluation and feedback by the GRC has strengthened the legitimacy of the decision-making processes underlying WHO's recommendations	Independent evaluation strengthens legitimacy	GRC process has helped recommendations stand up to criticism
WHO guideline development efforts are not yet designed to systematically make use of all relevant research evidence needed to inform decisions about complex interventions	Challenges with retrieving and appraising evidence to inform complex interventions	Nature of WHO guidelines becoming more complex
		Evidence from beyond RCTs needed to inform recommendations
		Challenges with formulating systematic review questions that capture broader range of evidence
	Challenges with and perceptions about GRADE	GRADE struggling with qualitative evidence
		Dissatisfaction with how GRADE evaluate non-randomised study designs
		Challenging with rigid application of GRADE
		GRC process perceived rigid/complicated
		Misperceptions about GRADE only being applicable to evidence from RCTs
		Need for more sophisticated understanding of GRADE
	Dialogues to address challenges	Increasing awareness within GRC about difficulties
		Constructive dialogue with GRC and methodologists crucial
		GRADE approach evolving to become more applicable to broader range of evidence
WHO's guideline development reforms do not currently apply to all outputs published from all of WHO's technical units and scientific advisory committees	Bypassing of formal guideline development process	Guidance being issued outside process overseen by GRC
		Tempting to circumvent GRC process
	All issued guidance could benefit from independent evaluation, monitoring and feedback	Disorganised approach to managing guidance produced outside GRC requirements
		Similar quality assurance needed for other guidance

GRADE, Grading of Recommendations Assessment, Development and Evaluation; GRC, Guidelines Review Commitee; RCT, randomised controlled trial.

RESULTS

WHO's guideline development reforms represented a transformational shift in its approach to producing clinical and public health recommendations

It was widely recognised among WHO interviewees that earlier criticisms of WHO's old guideline development process were appropriate and that reforms were needed due to the widespread reliance on expert opinion and ad hoc processes rather than structured processes informed by synthesised research evidence. For example, according to one WHO interviewee:

> … it was felt that this is just a bunch of expert opinions that is coming together. Who you happen to pick on that committee was going to define what the outcome was going to be. It was felt that WHO was handpicking and forcing themselves to a particular outcome that a small group of people at WHO wanted. (WHO Interviewee 9)

The consistent use of systematic reviews was described to be "the major change introduced with the guideline process", with the aim being to "stop WHO relying only on expert opinion" (WHO Interviewee 16). The reformed guideline development process was described to represent a transformational shift in how WHO developed clinical and public health recommendations, with one interviewee arguing that "how things are being done now is totally different from ten years ago" (WHO Interviewee 22). It was felt that the reforms overall had improved the quality of WHO's recommendations:

> Yeah, it is a long process that costs time and money because you have to commission the systematic reviews which take time. But then, at the end, you have a document and no one can come and tell you, 'OK, why this?' because you have the evidence, you can show the evidence supports the recommendations. I think that is a good thing for WHO. (WHO Interviewee 18)

Independent evaluation and feedback by the GRC has strengthened the legitimacy of the decision-making processes underlying WHO's recommendations

Several interviewees observed that guidelines from WHO frequently meet opposition, especially when recommendations challenge commercial interests and WHO subsequently face "strong push-back from the industry" (WHO Interviewee 13). Accordingly, many interviewees highlighted the important role of the GRC as a quality assurance mechanism for recommendations from WHO. Interviewees described the GRC to be an institutional mechanism for independent evaluation and feedback to ensure that appropriate procedures were followed. It was reported that the GRC had contributed to strengthening the legitimacy of decision-making processes underlying WHO's recommendations. For example, two interviewees expressed:

> It [the GRC] safeguards the procedural issues to say the members of the guideline development group were all vetted for conflict of interest, the meeting was done according to the procedures of the GRC, so we have that assurance. Even when we have serious challenges, letters of complaints, for example written to ethics committees, that helped us make our case on the process, but also on the technical recommendations. (WHO Interviewee 14)

> As long as you have something like a GRC mechanism that screens everything in the end, where people can go and say 'this is how we did it', these are the members, this is the process, and so it has been done well. (WHO Interviewee 22)

WHO guideline development efforts are not yet designed to systematically make use of all relevant research evidence needed to inform decisions about complex interventions

Although WHO interviewees generally expressed support for the guideline development reforms, many interviewees argued that the guideline development process in its current form is not designed to consider the full range of evidence that can inform health systems and public health recommendations. Two major challenges were highlighted by the interviewees.

First, some interviewees expressed that research questions for systematic reviews could be formulated very narrowly such that they did not capture the broader range of evidence needed to make nuanced recommendations about health systems and public health issues. One interviewee explained:

> I think that in some cases, there most probably is a tendency to pose too many peephole questions in areas where you're not going to have the evidence, so you're going through your systematic review and the outcome is very confusing. (WHO Interviewee 20)

Another interviewee described the challenge to be that "the definition of a systematic review is being set so narrowly that it's hard to bring in other kinds of relevant information to the table" (WHO Interviewee 9). Accordingly, interviewees emphasised the need for guidelines about health systems and public health interventions to consider evidence generated by non-randomised study designs and systematically documented programme experience, especially in order to produce guidance about how to implement and scale-up interventions. For example, two WHO interviewees explained:

> … because some of the richness of how you scale it up and make it work in the field is quite distinct from RCTs [randomised controlled trials], or when you review RCTs, you basically exclude a vast amount of knowledge which is powerful knowledge in the field. (WHO Interviewee 15)

> It is becoming challenging because on clinical interventions or something very medical, RCTs are fine and they produce the right sort of evidence. Increasingly, we are producing evidence around programmatic issues, service delivery programs, health systems functioning, social interventions, behavioural change, etc, and when you are looking at providing evidence on how programs should function, RCTs are generally not the best source of evidence. (WHO Interviewee 28)

Second, several interviewees expressed dissatisfaction over the current use of the GRADE approach for evaluating the quality of the evidence informing guideline development, raising two main concerns. The first was about GRADE being designed to assess the evidence for the effectiveness of interventions, but struggling "with how to interpret qualitative research findings" (WHO Interviewee 14). The second was about how the GRADE process rated the quality of evidence generated from non-randomised study designs and that most often the quality of evidence from these studies was rated as 'low'. It was argued that in these cases, consistent results from multiple well-designed observational studies should lead to a more favourable rating of the quality of the evidence, especially when interventions and policies cannot be tested with randomised controlled trials (RCTs) due to ethical, legal or logistical reasons. For example, two interviewees expressed:

> I still have a problem with, you know, if you have ten studies by ten different investigators, finding the same results, [and] not giving that benefit versus if you have only one study. Think about it the way you want, but for me there is more evidence with ten different studies from ten different investigators with ten different methodologies, so we try to advocate for increasing quality because of the consistency. But currently consistency is used only to downgrade. (WHO Interviewee 12)

> It does mean that when you score the grade using the GRADE tool, because it's not a randomized control trial, it tends to say that the 'oh, the level of evidence is low', when this is going to be always the highest-level of evidence that you can have and it's been replicated and done in many other places. (WHO Interviewee 31)

In relation to these concerns, interviewees expressed "there is too much rigidity in the way it's currently handled" (WHO Interviewee 12) and that some methodologists had "pushed a very rigid approach on the organization" (WHO Interviewee 16). Similarly, another interviewee argued that more flexibility and a better balance have to be found in order to adequately make use of all available evidence:

> I think the WHO swung the pendulum too far to the opposite of the extreme, trying to make things so pure and so unbiased that it left out the depth of understanding of the issues, and left out the depth of understanding the literature around an issue … We don't want it to be like it was before, but I think we pushed it too far, we need to come to something that is a little bit more balanced between those two approaches. (WHO Interviewee 9)

However, others argued that the "GRADE process does allow for consideration of all types of evidence," and that "it's most probably the interpretation of the GRADE process that most probably distracts from a reasonable decision-making process" (WHO Interviewee 20). It was argued that the guideline development process and the GRC suffered from an internal perception problem in that "people think GRADE means RCTs" and that "it's a lot about people being a lot more sophisticated in their understanding of it" (WHO Interviewee 16). In line with this, other interviewees expressed that a constructive dialogue with the GRC and methodologists, and guidance about the GRADE approach had proved helpful in improving their understanding about the guideline development process and adapting it to their specific context and needs:

> Then the new methods expert that we had was much more flexible and was willing to work with us to see how GRADE can be applied to the kind of the studies which are mainly providing the evidence from our area of work. (WHO Interviewee 7)

> I think that it can look very difficult and challenging, but when you discuss with the committee [GRC], they can give you ways to address it and adapt it to your problem. (WHO Interviewee 18)

Several interviewees described that conversations were occurring both within various departments—and between those managing guideline development efforts and the GRC—about this challenge, but that it remained to be seen how the agency and the GRC will ultimately deal with it:

> I think the GRC has become more open to those kinds of dialogues, but it's kind of outside the normal process, so you have to set it up as 'here's an exception, and why do we have to justify this exception'. And I think we are moving in the right direction, but it is already a problem that the whole structure is kind of set up, it has to be handled as a special case as opposed to setting it up to bring a broad body of evidence to the table. (WHO Interviewee 9)

WHO's guideline development reforms do not currently apply to all normative outputs published from all of WHO's technical departments and scientific advisory committees

Many interviewees reported being involved with WHO advisory bodies known as 'scientific and technical advisory groups'. These bodies were described as "more informal than a guidelines review group because it's not tasked with making necessarily policy recommendations on treatments and use of diagnostics" (WHO Interviewee 15). Another interviewee characterised these bodies as not primarily dealing with scientific questions and technical issues, but rather the strategic direction of various WHO programmes:

> It's an advisory body, and it's not just science and technical issues but it's also strategic issues for our programmatic response, so in that way it operates slightly differently from a group that would convene around one piece of scientific policy. But it broadly reflects on guidelines produced and the overall strategic direction for our program and fulfilling WHO's core functions. (WHO Interviewee 17)

However, several interviewees described these bodies as also producing guidelines and policy recommendations that should be subjected to the formal guideline development process. For example, one interviewee expressed that "from time to time, we most probably are putting out guidance that should really be graded for guideline recommendations" (WHO Interviewee 19). Interviewees also described cases of strategic attempts by WHO staff to circumvent the mandated guideline development process that is quality assured by the GRC. One interviewee raised that the resources and rules associated with the guideline development process could tempt staff to frame guidance documents in ways that enabled them to bypass the formal route, which risked issuing WHO recommendations without bringing all the available scientific evidence to the table:

> So you're trying to figure out how do I frame this so that it doesn't really sound like a WHO guideline, so that I don't have to deal with all that process … Except that if you don't have to go through the rigorous process, then you're kind of on your own, you can do whatever you want to do. There's no oversight, there's no guidance on what ought to be part of that, and I think you end up not making the best recommendations. (WHO Interviewee 9)

Another interviewee expressed concerns about the agency having a "completely disorganized approach to managing

anything outside of the GRC or the expert committee processes" and that there exist "examples of groups that are informal that do not follow any procedures" (WHO Interviewee 16). Further concerns were raised on whether these scientific advisory committees produced advice that was in keeping with WHO's policies and standards (WHO Interviewee 1). It was suggested that there is scope for these outputs to be captured by a more centralised quality assurance process that cover all WHO outputs:

> If GRC was being set up again, you would try to say everything goes through the committee, you would probably have a slightly different looking committee, and a slightly different handbook, but the bypass route should not be allowed. (WHO Interviewee 16)

DISCUSSION

This study aimed to understand how WHO has responded to the reforms made to its guideline development process over 10 years ago, including the major changes it made in 2007, the progress and impact of these changes to date, and key challenges that need to be addressed. Informed by the semistructured interviews with WHO staff and previous studies on WHO's guideline development process, we propose and discuss two recommendations that can inform efforts to improve the guideline development efforts of technical health agencies like WHO and others at the local, national and international levels.

Recommendation 1: Guideline development processes in technical health agencies and institutions should learn from WHO's vast experience with implementing independent evaluation, monitoring and feedback for process and quality to ensure the legitimacy of recommendations

One major factor contributing to the legitimacy of a health recommendation is the underlying evidence base, or in other words, the extent to which recommendations are consistent with the quality of available research evidence. Our study did not quantitatively address this question, but other studies suggest that there is still room for WHO to improve. For example, one study found that over 50% of strong WHO recommendations are based on assessments of evidence that place low or very low confidence in effect estimates (known as 'discordant recommendations'), with the majority of these being inconsistent with the GRADE approach.[6 12]

However, the strength of the evidence alone is insufficient to ensure the legitimacy of a recommendation. A second major contributing factor to legitimacy is the decision-making process through which the recommendations are developed.[30] Good process is particularly important in cases where recommendations might implicate commercial or ideological interests. A strategy that interested actors often use to challenge the legitimacy of recommendations, even in cases where the underlying evidence base is strong, is to question the process by which guidelines were generated, including the selection of experts or the participation of relevant stakeholders.

One way of strengthening the legitimacy of decision-making processes and protecting recommendations from undue criticism is to ensure that the process followed was subject to independent evaluation, monitoring and feedback for quality. WHO has over 10 years of accumulated experience with implementing its GRC mechanism, which can be seen as an internal quality assurance mechanism augmented by external expertise. Its assessments are made independently of WHO's senior management and its Member States. Accordingly, the GRC can be seen to serve two roles: (1) represent an institutional mechanism for independent evaluation that can strengthen the legitimacy of the decision-making process underlying the recommendations and (2) by involving staff internal to the agency support gradual institutionalisation of evidence-informed principles and processes. Learning from WHO's experience with implementing independent evaluation, monitoring and feedback for process and quality to ensure the legitimacy of recommendations can be relevant for other technical agencies and institutions responsible for guideline development on health issues.

Recommendation 2: Guideline development processes at WHO should be designed to better acquire, assess, adapt and apply the full range of research evidence that can inform recommendations about health systems and public health

The second recommendation calls for adapting WHO's guideline development process to better enable assessment of the evidence base needed to inform health systems and public health interventions, many of which, if not all, are 'complex interventions'.[31] Unlike individual-level interventions which can more easily be evaluated by randomising study participants to receive either treatment or control, many health systems and public health interventions cannot be randomised for ethical, legal or logistical reasons, even if governing decision-makers are supportive of doing so. Accordingly, evidence from non-randomised study designs—for example a well-designed observational study, quasi-experimental impact evaluation or systematically documented evidence from programme experience—may represent the highest quality of evidence one can expect for a public health or health systems intervention.[9 32] This challenge is not unique to guideline development processes at WHO but has been debated at various points over the past years.[14 33–40] It represents an important factor explaining why WHO interviewees raised concerns over the guideline development process not being flexible enough to incorporate and appropriately evaluate evidence from non-randomised study designs and qualitative studies. Similar views have previously been reported to be held by WHO staff[9] and guideline panel members,[11 41] and confirmed by methodologists with experience serving on guideline panels.[15]

Two design features of WHO's guideline development process seem to particularly need adaptation to fully incorporate the evidence base needed to inform

recommendations. The first design feature is the use of systematic reviews to critically appraise relevant research underpinning WHO recommendations. Systematic reviews are among the cornerstones of guideline development processes and are critical for reducing the risk of bias and reaching reliable evidence-informed conclusions. By way of background, systematic reviews are commonly conducted by formulating a clear and specific question most commonly using the PICO format (population, intervention, comparison, outcome) and by having explicit inclusion/exclusion criteria. This rigorous approach helps to identify and include studies that are comparable and able to respond to questions about effectiveness of interventions (what works).[36] This approach works sufficiently well for evaluating the effectiveness of a medical treatment. However, many health systems and public health interventions are 'complex' interventions, characterised by multiple interacting components, requiring the involvement of different organisational levels, having a number of different points of interactions between interventions and the settings in which the interventions are implemented and affecting different outcomes.[31 37 38] For complex interventions, framing systematic review questions too narrowly and relying solely on assessing evidence of effectiveness risks excluding the broader range of relevant evidence needed to inform recommendations. This includes evidence for factors important for implementing an intervention, for bringing an intervention to scale, for assessing the resources needed to implement interventions across different settings, for understanding the feasibility and acceptability of an intervention, for identifying the interactions among various components of complex interventions and for probing the systems in which the interventions were implemented.[42] These represent factors that are not easily identifiable if systematic review questions are narrow and solely include experimental intervention studies focused on safety and effectiveness. Currently, the *WHO Handbook for Guideline Development* do not offer comprehensive guidance for adapting the PICO format or considering alternative frameworks when dealing with systematic reviews for health systems and public health interventions.[14] In light of new tools and frameworks for conducting systematic reviews for complex interventions,[43] WHO might consider adapting its guidance.

The second design feature of WHO's guideline development process that has led to it being viewed as inflexible involves the approaches and tools used to evaluate the quality of the evidence base. At WHO, the GRADE approach has been the main tool for assessing the quality of the evidence underlying recommendations. This study's findings highlight three aspects associated with GRADE that seem to have reinforced the view among many WHO staff that guideline development processes are not designed to incorporate a broader evidence base. The first is an insufficient understanding among many people involved in guideline development processes about the purpose, utility and implementation of the GRADE approach. This was also highlighted by Sinclair *et al* in their evaluation of WHO's guideline development process,[9] suggesting that better understanding of GRADE among WHO staff involved with guideline development needs continued attention. To this end, guidance to promote a more sophisticated understanding of GRADE has been issued (including by the GRADE Working Group),[40] but greater awareness and wider implementation is needed at WHO. The second aspect is GRADE's initial rating of certainty in the evidence from non-randomised study designs as 'low'. This is a feature of GRADE that guideline developers beyond WHO have raised concerns over, since in fields where RCTs are sparse or not feasible, the quality of the evidence rarely will be rated as 'high' or 'moderate'[13]; fortunately, this is a criticism that the GRADE Working Group have noted and are seeking to address.[44] The third aspect is that GRADE was not designed to evaluate the quality of evidence from qualitative studies, which is increasingly recognised as crucial for informing decision-makers about the needs, values, perceptions and experiences of stakeholders important for an intervention, and the system-level factors affecting implementation.[41 45] The reliance, at least until very recently, on GRADE as the sole tool for assessing the body of evidence might have created a perception that guideline development processes are not intended to incorporate a broader evidence base, including qualitative evidence. These concerns over GRADE raised by WHO staff in our study align with challenges highlighted by others.[14 32 46 47]

On all of these fronts, there have been recent developments worth noting. For evaluating non-randomised study designs, the ROBINS-I tool (Risk Of Bias In Non-randomized Studies-of Interventions) has been developed, with further extensions planned.[48] Using ROBINS-I as part of GRADE assessments can enable better comparisons of evidence from non-randomised study designs and RCTs, as well as more detailed assessments of different types of non-randomised study designs (such as rating evidence from a well-designed interrupted time series studies higher than conventional non-randomised study designs), thereby addressing one of the major concerns raised by WHO interviewees.

For evaluating the quality of qualitative evidence, WHO has together with collaborators taken a leadership role. Its own guideline development process for recommendations about optimising health worker roles for maternal and newborn health expanded the evidence base beyond safety and effectiveness,[41] leading to the development of a new approach for assessing the confidence that can be placed in qualitative evidence—the Grading of Recommendations, Assessment, Development and Evaluation—Confidence in Evidence from Reviews of Qualitative Research (GRADE-CERQual) tool.[49] GRADE-CERQual has later been further

developed and implemented as part of WHO guideline development processes,[45] addressing yet another concern raised by WHO interviewees. Moreover, the Evidence-to-Decision Framework developed by the DECIDE project creates space for the assessment of an intervention's acceptability and feasibility and is increasingly being used in WHO's guideline development processes.[50–52] Finally, the challenges with synthesising and assessing the quality of evidence for complex interventions, and the need for guidance and tools are increasingly recognised, both within and beyond WHO.[43 53]

Specific actions for WHO

We identify at least three specific areas where action could be taken by WHO (box 2). First, there should be more frequent and systematic sharing of experiences among WHO departments and between the GRC and the various departments that develop guidelines. Such sharing and continuous professional development for WHO staff would help address the many issues raised by this study and others.[9 11 15] Second, the guideline development process should be further enhanced to meet the needs of health systems and public health interventions, which is consistent with recent calls in peer-reviewed journals from senior WHO staff.[54] It is therefore timely that efforts are underway to examine extensions to the GRADE approach,[55] as well as efforts led by the GRADE Working Group to integrate GRADE assessment with the use of tools such as ROBINS-1.[44] Moreover, WHO has internally recognised this and other challenges with its guideline development process[56] and initiated its own process for improving retrieval, synthesis and assessment of evidence on complex health interventions, which might inform future changes to the design of WHO's guideline development process.[53]

Finally, WHO should consider whether *all* products containing advice and guidance that emerge from the plethora of WHO's technical departments and scientific advisory committees—many of which currently operate outside of the GRC's mandate—could benefit from a centralised quality assurance process independent of WHO Member States, similar to what is currently performed by the GRC for WHO's formal guidelines. This may improve quality and legitimacy, but it will also require resources, time and planning. On this front, a recent development is that WHO has proposed in its draft 13th General Programme of Work to 'establish guiding principles and quality assurance procedures for the design, formulation and dissemination/follow-up of all normative products (all normative products, including strategies, road maps and global action plans will be based on agreed standards and reviewed independently, as is the case for technical guidelines), including maximizing the use and engagement of top international experts' [57]—a proposal informed by a 2017 review of WHO's normative functions.[58]

Strengths, limitations and reflections on study design and data analysis

We identify three main strengths and two main limitations of this study. The first strength is the large number of interviewees (n=16) who had experience with WHO's guideline development process, as well as additional interviewees (n=19) working with other structures that produce scientific advice (eg, expert committees, scientific and technical advisory groups), which enabled us to consider WHO's broader context when interpreting our findings. A second strength is that the majority of the interviewees were senior WHO staff who have been working for the agency since before the guideline development reforms were initiated and therefore were able to inform our study with their experiences before and after the reforms. The third strength is the diversity in technical areas represented by the interviewees, which enabled us to identify themes that were relevant to guideline development processes across WHO's technical areas. The invited WHO staff who for various reasons decided to not participate in this study did not, with respect to roles and technical areas, differ from those who were interviewed since we managed to recruit interviewees from various levels and across the many technical areas of the agency. Overall, our analysis was informed by a large amount of qualitative data consisting of diverse sets of relevant experiences accumulated over a long period of time.

The first main limitation is that the study was initially conceived to examine the design features of WHO's scientific advisory committees in general, and not specifically to evaluate WHO's guideline development reforms. We may therefore have overlooked asking important questions that could have deepened insights about WHO's experience with its reformed guideline development process. For example, while all interviewees emphasised the importance of diverse representation in guideline development groups, we did not probe in detail their experience with involving populations affected by the recommendations during guideline development—which is another important factor affecting the legitimacy of recommendations. The second main limitation is social desirability bias—that WHO interviewees may have responded in a way that casts the agency in a favourable light, downplaying internal weaknesses and challenges. However, the majority of interviewees provided candid assessments of the agency's progress and challenges with guideline development and producing scientific advice. Moreover, some also spoke rather critically of the agency, such that we do not believe the results are unduly biased in this way.

Finally, choices made during data analysis are worth discussing in light of different views and traditions with respect to improving reliability in qualitative research.

In our study, both investigators discussed and reached agreement about the identified themes and the fit of the coding with these themes. However, we did not implement independent coding by two investigators; rather, one investigator undertook the initial coding and identification of preliminary themes, which were subsequently discussed and refined in dialogue with the second investigator. This strategy may be seen as a weakness by researchers who argue that multiple, independent coding and calculation of inter-rater reliability is a prerequisite for rigour and trustworthiness in qualitative research.[59–61] However, our approach is in line with strategies undertaken and advocated by many qualitative researchers, including Braun & Clarke who have developed an approach to thematic analysis which closely resemble the analytical strategy undertaken in this study.[18 62] They and others argue that there is no 'one' accurate way of coding and interpreting qualitative data and that it is unrealistic to expect different researchers to reach exactly the same insights from qualitative data, since they may differ in disciplinary backgrounds and theoretical starting points. What remains important is full transparency about choices made during data analysis so that others can evaluate how these choices may have affected analysis and interpretation. Moreover, it remains important to otherwise minimise the risk of misrepresenting the qualitative data. To address the former, we have described our approach to data collection and analysis in great detail in the Methods section. To address the latter, we implemented participant checking. While less than half of the interviewees responded to our queries, only one interviewee raised objections with the way the findings were presented. We assume, but cannot be completely certain, that if other interviewees had similar objections, they would have expressed these after receiving the interview summaries or the manuscript. Moreover, two WHO officials who were not interviewed for this study, but with extensive in-depth experience with WHO's guideline development process, reviewed the manuscript and reported to recognise the experiences and key challenges identified by our study. Overall, we believe that the reported findings and interpretation do not misrepresent the interview data, but accept that these findings could be interpreted differently by other researchers. We therefore invite continued debate on issues raised by this study.

CONCLUSION

Since WHO initiated reforms to its guideline development process, the agency has advanced towards a more transparent and rigorous evidence-informed process for crafting its recommendations. The 'culture change' senior WHO staff called for over 10 years ago appears gradually to have spread throughout the agency.[2] However, at least two major challenges remain for WHO: (1) ensuring that all issued guidance benefits from independent evaluation, monitoring and feedback for process and quality and (2) adapting its guideline development processes to better acquire, assess, adapt and apply the full range of evidence that can inform recommendations.

Author affiliations
[1]Department of Global Health, Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway
[2]Oslo Group on Global Health Policy, Department of Community Medicine and Global Health and Centre for Global Health, Institute of Health and Society, Faculty of Medicine, University of Oslo, Oslo, Norway
[3]Global Strategy Lab, Dahdaleh Institute for Global Health Research, Faculty of Health and Osgoode Hall Law School, York University, Toronto, Ontario, Canada
[4]Department of Health Research Methods, Evidence, and Impact, and McMaster Health Forum, McMaster University, Hamilton, Ontario, Canada
[5]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

Acknowledgements The authors thank Susan L Norris from WHO's GRC Secretariat for responding to questions about WHO's guideline development processes and updating us about ongoing developments.

Contributors UG and SJH had the original idea for the manuscript. Data collection was done by UG. Data analysis was done by UG and SJH. UG prepared the original draft of the manuscript, and SJH provided input and comments on successive drafts. Both authors read and approved the final draft.

Funding This study was completed as part of the research project 'Strengthening International Collaboration for Capitalizing on Cost-Effective and Life-Saving Commodities' (i4C) that is funded through the Research Council of Norway's Global Health & Vaccination Programme (GLOBVAC Project No 234608). SJH is additionally funded by the Canadian Institutes of Health Research and the Ontario Government's Ministry of Research, Innovation & Science.

Competing interests UG has previously supported WHO's guideline development process on 'Transforming and scaling up health professionals' education and training' as an intern in 2011 and led a multicountry case study for the Norwegian Knowledge Centre for the Health Services in 2011–2012, which informed the development of WHO's recommendations on 'Optimizing health worker roles for maternal and newborn health.' SJH has previously worked for WHO. Both authors have previously published articles examining WHO's role in global health governance.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement At the time of data collection, consent was not obtained from the interviewees to share the qualitative data beyond what is contained in this article. We are, therefore, unable to provide access to the raw data.

REFERENCES

1. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. *Lancet* 2007;369:1883–9.
2. Hill S, Pang T. Leading by example: a culture change at WHO. *Lancet* 2007;369:1842–4.
3. Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
4. World Health Organization. *WHO handbook for guideline development*. Geneva: World Health Organization, 2014.
5. World Health Organization. *WHO handbook for guideline development*. Geneva: World Health Organization, 2012.
6. Alexander PE, Bero L, Montori VM, *et al*. World Health Organization recommendations are often strong based on low confidence in effect estimates. *J Clin Epidemiol* 2014;67:629–34.

7. Burda BU, Chambers AR, Johnson JC. Appraisal of guidelines developed by the World Health Organization. *Public Health* 2014;128:444–74.
8. Hoffman SJ, Lavis JN, Bennett S. The use of research evidence in two International Organizations' Recommendations about health systems. *Healthc Policy* 2009;5:66–86.
9. Sinclair D, Isba R, Kredo T, *et al*. World Health Organization guideline development: an evaluation. *PLoS One* 2013;8:e63715.
10. Chang LW, Kennedy CE, Kennedy GE, *et al*. Developing WHO guidelines with pragmatic, structured, evidence-based processes: A case study. *Glob Public Health* 2010;5:395–412.
11. Alexander PE, Gionfriddo MR, Li SA, *et al*. A number of factors explain why WHO guideline developers make strong recommendations inconsistent with GRADE guidance. *J Clin Epidemiol* 2016;70:111–22.
12. Alexander PE, Brito JP, Neumann I, *et al*. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. *J Clin Epidemiol* 2016;72:98–106.
13. Akl EA, Kennedy C, Konda K, *et al*. Using GRADE methodology for the development of public health guidelines for the prevention and treatment of HIV and other STIs among men who have sex with men and transgender people. *BMC Public Health* 2012;12:386.
14. Rehfuess EA, Akl EA. Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health* 2013;13:9.
15. Alexander PE, Li SA, Gionfriddo MR, Stoltzfus RJ, *et al*. Senior GRADE methodologists encounter challenges as part of WHO guideline development panels: an inductive content analysis. *J Clin Epidemiol* 2016;70:123–8.
16. Yin RK. *Qualitative research from start to finish*. New York, NY: Guilford Press, 2011:348.
17. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;3:77–101.
18. Questions about thematic analysis [Internet]. https://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis/frequently-asked-questions-8.html#e41676c2ec9a2c4caae1664a24aa3a0a (cited 3 Mar 2018).
19. Sandelowski M. What's in a name? Qualitative description revisited. *Res Nurs Health* 2010;33:77–84.
20. Colorafi KJ, Evans B. Qualitative descriptive methods in health science research. *HERD* 2016;9:16–25.
21. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nurs Health Sci* 2013;15:398–405.
22. 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;19:349–57.
23. Carr Z, Clarke M, Akl EA, *et al*. Using the grade approach to support the development of recommendations for public health interventions in radiation emergencies. *Radiat Prot Dosimetry* 2016;171:144–55.
24. World Health Organization, World Health Organization, Reproductive Health and Research. WHO recommendations: optimizing health worker roles to improve access to key maternal and newborn health interventions through task shifting.[internet]. 2012 http://www.ncbi.nlm.nih.gov/books/NBK148518/ (cited 5 Mar 2018).
25. WHO. *Transforming and scaling up health professionals' education and training. World Health Organization Guidelines 2013.[Internet*. Geneva: World Health Organization, 2013. http://whoeducationguidelines.org/sites/default/files/uploads/WHO_EduGuidelines_20131202_web.pdf. (cited 2018 Mar 3).
26. Hoffman SJ, Røttingen J-A, Bennett S, *et al*. *A Review of Conceptual Barriers and Opportunities facing Health Systems Research to inform a Strategy from the World Health Organization. [internet]*. Geneva: Alliance for Health Policy and Systems Research, 2012. http://www.who.int/alliance-hpsr/alliancehpsr_backgroundpaperconceptualbarriersopportunities.pdf. (cited 3 Mar 2018).
27. Gopinathan U, Watts N, Hougendobler D, *et al*. Conceptual and institutional gaps: understanding how the WHO can become a more effective cross-sectoral collaborator. *Global Health* 2015;11:46.
28. Hoffman SJ. Strengthening global health diplomacy in Canada's foreign policy architecture: Literature review and key informant interviews. *Canadian Foreign Policy Journal* 2010;16:17–41.
29. Interagency Advisory Panel on Research Ethics. TCPS 2 (2014)—the latest edition of Tri-Council Policy Statement: Ethical conduct for research involving humans [Internet]. http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/ (cited 1 Mar 2016).
30. Dawson A, Verweij M, Health P. Public Health and Legitimacy: Or why there is still a place for substantive work in ethics. *Public Health Ethics* 2014;7:95–7.
31. Petticrew M. When are complex interventions 'complex'? When are simple interventions 'simple'? *Eur J Public Health* 2011;21:397–8.
32. European Centre for Disease Prevention and Control. *Evidence-based methodologies for public health - How to assess the best available evidence when time is limited and there is lack of sound evidence. [internet]*. Stockholm: ECDC, 2011. http://ecdc.europa.eu/en/publications/publications/1109_ter_evidence_based_methods_for_public_health.pdf. (cited 30 Nov 2016).
33. Durrheim DN, Reingold A. Modifying the GRADE framework could benefit public health. *J Epidemiol Community Health* 2010;64:387.
34. Rehfuess EA, Bruce N, Prüss-Üstün A. GRADE for the advancement of public health. *J Epidemiol Community Health* 2011;65:559.
35. Schünemann H, Hill S, Guyatt G, *et al*. The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health* 2011;65:392–5.
36. Petticrew M. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev* 2015;4:36.
37. Petticrew M, Anderson L, Elder R, *et al*. Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol* 2013;66:1209–14.
38. Petticrew M, Rehfuess E, Noyes J, *et al*. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol* 2013;66:1230–43.
39. Movsisyan A, Melendez-Torres GJ, Montgomery P. A harmonized guidance is needed on how to "properly" frame review questions to make the best use of all available evidence in the assessment of effectiveness of complex interventions. *J Clin Epidemiol* 2016;77:139–41.
40. Guyatt GH, Schünemann HJ, Djulbegovic B, *et al*. Guideline panels should not GRADE good practice statements. *J Clin Epidemiol* 2015;68:597–600.
41. Glenton C, Lewin S, Gülmezoglu AM. Expanding the evidence base for global recommendations on health systems: strengths and challenges of the OptimizeMNH guidance process. *Implement Sci* 2016;11:98.
42. Squires JE, Valentine JC, Grimshaw JM. Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol* 2013;66:1215–22.
43. Guise JM, Chang C, Butler M, *et al*. AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol* 2017;90:6–10.
44. Schünemann HJ, Cuello C, Akl EA, *et al*. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol* 2018.
45. Langlois EV, Tunçalp Ö, Norris SL, *et al*. Qualitative evidence to improve guidelines and health decision-making. *Bull World Health Organ* 2018;96–79–79A.
46. Movsisyan A, Melendez-Torres GJ, Montgomery P. Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol* 2016;70:191–9.
47. Movsisyan A, Melendez-Torres GJ, Montgomery P. Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol* 2016;78:22–33.
48. Sterne JA, Hernán MA, Reeves BC, *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
49. Lewin S, Glenton C, Munthe-Kaas H, *et al*. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med* 2015;12:e1001895.
50. Rosenbaum SE, Moberg J, Glenton C, *et al*. *Developing Evidence to Decision Frameworks and an Interactive Evidence to Decision Tool for Making and Using Decisions and Recommendations in Health Care*: Global Challenges:1700081.
51. WHO. *WHO recommendations on antenatal care for a positive pregnancy experience [Internet*: World Health Organization, 2016. http://apps.who.int/iris/bitstream/handle/10665/250796/9789241549912-eng.pdf?sequence=1. (cited 3 Mar 2018).
52. WHO. *Health worker roles in providing safe abortion care and post-abortion contraception. [Internet]*. Geneva: World Health Organization, 2015. http://apps.who.int/iris/bitstream/handle/10665/181041/9789241549264_eng.pdf;jsessionid=D82669E3011AB37478CD7C0319B02033?sequence=1. (cited 3 Mar 2018).
53. WHO. Retrieval, synthesis and assessment of evidence on complex health interventions. http://www.who.int/maternal_child_adolescent/guidelines/development/complex-health-interventions/en/ (cited 19 Apr 2018).

54. Norris SL, Bero L. GRADE Methods for guideline development: time to evolve? *Ann Intern Med* 2016;165:810.
55. University of Oxford. *GRADE Extension for Complex Social Interventions [Internet]*: Department of Social Policy and Intervention, 2016. https://www.spi.ox.ac.uk/research/details/grade-extension-for-complex-social-inter.html. (cited 19 Nov 2016).
56. Norris SL, Ford N. Improving the quality of WHO guidelines over the last decade: progress and challenges. *Lancet Glob Health* 2017;5:e855–e856.
57. WHO. *Draft thirteenth general programme of work 2019–2023. Promote health, keep the world safe, serve the vulnerable [Internet]*. Geneva: World Health Organization, 2018. http://apps.who.int/gb/ebwha/pdf_files/EB142/B142_3-en.pdf?ua=1. (cited 2 Feb 2018).
58. Nordic Consulting Group. Evaluation of WHO's Normative Function [Internet]. 2017 http://www.who.int/about/evaluation/who_normative_function_report_july2017.pdf (cited 1 Dec 2017).
59. Berends L, Johnston J. Using multiple coders to enhance qualitative analysis: The case of interviews with consumers of drug treatment. *Addict Res Theory* 2005;13:373–81.
60. Armstrong D, Gosling A, Weinman J, *et al*. The place of inter-rater reliability in qualitative research: an Empirical Study. *Sociology* 1997;31:597–606.
61. Mays N, Pope C. Rigour and qualitative research. *BMJ* 1995;311:109–12.
62. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ* 2001;322:1115–7.