The GRADE approach and Bradford Hill's criteria for causation

Holger Schünemann,[1] Suzanne Hill,[2] Gordon Guyatt,[1] Elie A Akl,[3] Faruque Ahmed[4]

[1]Departments of Clinical Epidemiology & Biostatistics and of Medicine, McMaster University Health Sciences Centre, Hamilton, Ontario, Canada
[2]World Health Organization, Geneva, Switzerland
[3]State University of New York, Buffalo, New York, USA
[4]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Correspondence to
Holger J Schünemann, Department of Clinical Epidemiology & Biostatistics, McMaster University Health Sciences Centre, Room 2C10B, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada; schuneh@mcmaster.ca

Accepted 19 July 2010
Published Online First 14 October 2010

ABSTRACT

This article describes how the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to grading the quality of evidence and strength of recommendations considers the Bradford Hill criteria for causation and how GRADE may relate to questions in public health. A primary concern in public health is that evidence from non-randomised studies may provide a more adequate or best available measure of a public health strategy's impact, but that such evidence might be graded as lower quality in the GRADE framework. GRADE, however, presents a framework that describes both criteria for assessing the quality of research evidence and the strength of recommendations that includes considerations arising from the Bradford Hill criteria. GRADE places emphasis on recommendations and in assessing quality of evidence; GRADE notes that randomisation is only one of many relevant factors. This article describes how causation may relate to developing recommendations and how the Bradford Hill criteria are considered in GRADE, using examples from the public health literature with a focus on immunisation.

> "Scientists believe in proof without certainty; most people believe in certainty without proof." (Ashely Montagu; from http://meds.queensu.ca/medicine/obgyn/links/criteria_for_causation.htm)

It has been proposed that the Grading of Recommendations Assessment, Development and Evaluation (GRADE) for public health questions should consider the Bradford Hill criteria for causation and that GRADE requires adaptation.[1] In this article, we describe the relation of the Bradford Hill criteria to the GRADE approach to grading the quality of evidence and strength of recommendations. The primary concern seems that evidence from non-randomised studies may provide a more adequate or best available measure of a public health strategy's impact, but that such evidence might be graded as lower quality in the GRADE framework. We would like to reiterate that GRADE presents a framework that describes both criteria for assessing the quality of research evidence and the strength of recommendations. In assessing quality of evidence, GRADE notes that randomisation is only one of many relevant factors. Furthermore, GRADE is not specific to the narrow field of therapeutic interventions. Indeed, it likely is the most broadly applied framework for evaluation of evidence and developing recommendations.

We would like to clarify several issues that have been raised in a recent editorial published in the Journal of Epidemiology and Community Health.[1] First, concern has been expressed that herd immunity as a result of immunisation and indirect effects on the co-circulation of other pathogens are typically ascertained through the use of observational epidemiological methods. Although we do not disagree with this assessment, we would like to point out that, innovative randomised controlled trials (RCTs) using cluster-randomisation can be conducted to provide such information.[2] Second, concern is expressed that a quasi-RCT that found a 94% protective effect of a live, monovalent vaccine against measles was classified as 'moderate level of scientific evidence'. However, GRADE's strength of association criteria can be applied to quasi-RCTS and observational studies with no major threats to validity to upgrade the quality of evidence (see below). Such a judgement would be possible in this situation. Third, it is implied that GRADE ratings do not give credit to the 'gradient of effects with scale of population level impact compatible with degree of coverage'. However, we would like to clarify that GRADE's dose-response criterion is not limited to clinical dose only, and that it can be applied to such gradients at the population level to upgrade the quality of the evidence.

Finally, it has been speculated that anti-vaccination lobby groups may abuse the GRADE ratings.[1] Although, abuse of any system is possible, in the case of GRADE it is equally likely that increased transparency provided by the GRADE framework can strengthen, rather than undermine, the trust in vaccines and other clinical interventions.[3 4]

How does GRADE see the relative role of observational studies and RCTs in judging quality of evidence? The GRADE framework applies to all study designs, but randomisation as a methodological approach to protect against bias and confounding is considered very important. Nevertheless, in certain situations observational studies may provide more relevant information than RCTs (eg, situations of long term follow-up and when RCTs only provide very indirect data).

Does a judgement of moderate or low quality evidence preclude such evidence driving a recommendation in favour, or against, an intervention? In the GRADE approach, it does not. It is possible that post-licensure observation studies on long-term or rare serious adverse effects of a vaccine may receive lower evidence grades, but the possibility of a potentially serious harm may be judged sufficient by a guideline panel to make a recommendation against an intervention. In fact, when moving from evidence to recommendations, GRADE does not focus on research evidence only and the framework does not preclude action based on lower evidence

levels. GRADE also acknowledges the wide range of possible judgements a guideline panel may make, while application of the GRADE approach enhances transparency concerning the evidence that was considered and transparency in how judgements regarding the quality of evidence were made.

It has been requested that the Bradford Hill criteria for assessing causality be considered in the GRADE framework.[1] We agree that Bradford Hill's criteria remain, half a century after their description, relevant factors that influence our confidence in a causal relation. Some of these criteria influence our confidence that estimates of effects are correct or that the association between an exposure and an outcome are trustworthy before making a judgement about causality. We note, however, that there are steps between establishing or surmising causality and moving to interventions that act on the perceived causal relation. Establishing a causal relation between an exposure and an outcome does not always allow a confident inference that all methods of modification or removal of the exposure lead to changes in outcomes. This is in particular true for complex interventions that go beyond simple drug interventions. The judgements involved include the confidence that removal of the exposure can be achieved and are included in the judgements about directness in GRADE.

Nevertheless, GRADE has adopted most of Bradford Hill's criteria, some implicitly, others explicitly. However, we realise that the way GRADE incorporates the criteria for causation may not be evident to everyone. We will therefore describe how the Bradford Hill criteria are considered in the GRADE approach and, when they are not, provide a rationale for not considering the particular criterion.

GRADE defines the quality of evidence as the confidence in an estimate of effect (causal relation) from a body of evidence. We will, therefore, use the term upgrading when this confidence is increased and the term downgrading when this confidence is lowered (table 1).

(1) *Strength of the association*. Bradford Hill suggests that a strong association supports causality. This criterion is directly considered in GRADE through upgrading. In the GRADE system, strong associations between an intervention or exposure and an outcome can lead to upgrading the quality of evidence, ie, increases our confidence that the intervention causes a change in the incidence of that outcome. A second criterion, imprecision, which limits our confidence in an effect, even if strong associations are present, is indirectly related to this item in that it lowers our confidence in an association if the magnitude of the effect is uncertain enough to undermine our confidence.

(2) *Consistency*. Bradford Hill suggests that causation is more likely if the results from various research studies are consistent. This criterion is directly considered in GRADE. The GRADE approach suggests downgrading the quality of evidence when there is inconsistent evidence, ie, when studies of similar quality show unexplained heterogeneity in the estimates of effect.

(3) *Temporality or study design suitability*. Bradford Hill describes that there must be a temporal relation between the exposure and outcome. This criterion, usually better than observational studies, in particular if they are not well designed and conducted is indirectly considered in GRADE in at least three ways. First, evidence from randomised controlled trials—which by default establish this temporal relationship—start as higher quality than evidence from studies that do not establish this relationship in GRADE. Second, longitudinal observational studies that include concurrent control groups would likely provide higher quality evidence than cross-sectional studies. Third, GRADE requires the critical consideration of confounders and covariates that may be responsible for a spurious relation when evaluating observational study designs.[3]

(4) *Biological gradient*. As described by Bradford Hill, a biological gradient between an exposure and the magnitude of an effect increases the confidence in causality. GRADE's criterion of upgrading the quality of evidence for a dose—response relationship is a direct application of this principle.

(5) *Specificity*. According to Bradford Hill, causation is more likely if there is a specific outcome related to a specific exposure in that altering the cause alters the disease outcome. In GRADE, this criterion is indirectly considered in the evaluation of whether both the exposure and the outcome were measured directly and by formulating the question and selecting the population, intervention, comparator and outcome in the first place. However, single exposures or interventions are almost invariably related to many outcomes and vice versa. This criterion is not an important criterion for an evaluation of the effects of interventions.

(6) *Biological plausibility*. Whether the association is plausible or not influences causality in the Bradford Hill approach. GRADE does not consider the issue of plausibility in the strict sense as it was included by Bradford Hill. This is, in part, related to the fact that every relation can be described as plausible given that researchers will always think of an explanation once an association is observed. However, GRADE partially considers plausibility in the evaluation of how direct the intervention is related to a surrogate outcome. For instance, we would frequently accept surrogates that have repeatedly responded to interventions in the same way as patient important outcomes. For example, we accept the use of CD4 levels and HIV viral load as acceptable surrogates for mortality and other patient important outcomes, and one of the reasons for this acceptance is the biological plausibility that CD4 levels and HIV viral load are determinants of disease and therapy success. In addition, GRADE considers biological plausibility as a criterion for the evaluation of the believability of an observed subgroup effect. Furthermore, by asking the question of interest and identifying evidence for or against it, the item of biological plausibility is considered indirectly.

(7) *Coherence*. According to Bradford Hill, causation is more likely if what is observed is supported by and in agreement with the natural history of the disease. GRADE does not consider this criterion explicitly but assessing the validity of surrogate outcomes includes these considerations implicitly as well as formulating appropriate healthcare questions. Furthermore, greater emphasis is placed on direct (eg, long-term population

Table 1 Bradford Hill criteria of causality and their relation to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) criteria for upgrading and downgrading

Bradford Hill criteria	Consideration in GRADE
Strength	Strength of association and imprecision in effect estimate
Consistency	Consistency across studies, ie, across different situations (different researchers)
Temporality	Study design, specific study limitations; RCTs fulfil this criterion better than observational studies, properly designed and conducted observational studies
Biological gradient	Dose—response gradient
Specificity	Indirectness
Biological plausibility	Indirectness
Coherence	Indirectness
Experiment	Study design, randomisation, properly designed and conducted observational studies
Analogy	Existing association for critical outcomes will lead to not downgrading the quality, indirectness

Table 2 Interpretation of the Grading of Recommendations Assessment, Development and Evaluation (GRADE)

Interpretation of strong and conditional (weak) recommendations	Strong recommendation	Conditional (weak) recommendation*
For patients	Most individuals in this situation would want the recommended course of action and only a small proportion would not.	The majority of individuals in this situation would want the suggested course of action, but many would not.
For clinicians	Most individuals should receive the intervention. Formal decision aids are not likely to be needed to help individuals make decisions consistent with their values and preferences.	Recognise that different choices will be appropriate for individual patients and that clinicians must help each patient arrive at a management decision consistent with his or her values and preferences. Decision aids may be useful in helping individuals making decisions consistent with their values and preferences.
For policy makers and developers of quality measure	The recommendation can be adapted as policy in most situations. Adherence to this recommendation according to the guideline could be used as a quality criterion or performance indicator.	Policy making will require substantial debate and involvement of various stakeholders. An appropriately documented decision making process could be used as quality indicator.
Interpretation of the categories of the quality of evidence		
High: ⊕⊕⊕⊕	There is high confidence that the true effect lies close to that of the estimate of the effect.	
Moderate: ⊕⊕⊕◯	There is moderate confidence in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.	
Low: ⊕⊕◯◯	The panel's confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect.	
Very low: ⊕◯◯◯	There is little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.	

*Guideline panels applying GRADE use the terms 'conditional' and 'weak' synonymously.

important outcomes) rather than short-term outcomes during the formulating of questions and the evaluation of the evidence.
(8) *Experimental evidence*. Experimental evidence enhances the probability of causation. GRADE places emphasis on rigorous experimental designs and this criterion is directly considered. RCTs provide the ideal experimental study design to establish causation where randomisation is the leading experimental factor that increases confidence in associations. Flaws in the experimental design or implementation of an RCT lead to downgrading the quality of evidence. Better experimentally designed observational studies with independent control groups will be graded as higher quality than poorly designed observational studies.
(9) *Reasoning by analogy*. Bradford Hill suggests that existing similar associations would support causation. This criterion is indirectly considered in the GRADE approach. The overall quality of evidence may not be lowered for a single critical outcome if higher quality evidence exists for other critical outcomes and the association is similar in direction.[5] For example, if an intervention to reduce exposure (eg, air pollution) is associated with mortality and chronic respiratory disease and this is based on moderate (⊕⊕⊕◯) quality evidence, but only low (⊕⊕◯◯) quality exists for a third outcome, such as stroke, but all associations are indicating similar effects, then the overall quality would not be lowered because of the single outcome of low quality even if it is critical. Furthermore, GRADE considers indirect evidence when direct evidence is not available.

We appreciate the opportunity to provide clarification regarding how the GRADE framework applies to public health.[1] The GRADE framework—like other evidence-based systems—is an evolving system and we welcome input and insights from users on the strengths and challenges of applying GRADE to vaccines and other preventive public health programmes. Additional use in the field may improve GRADE, in particular in the field of public health and policy interventions, and will advance the field of guidance development. We have previously discussed some of the advantages and disadvantages of applying one approach across different questions.[6] In regard to the Bradford Hill criteria, we believe that the GRADE approach appropriately includes most of the considerations that Bradford Hill suggested.

Finally, there are two other issues that are relevant to this discussion and require emphasis. We remind users of GRADE that the approach separates the quality of evidence from the strength of recommendations and that for the appropriate emphasis we place the recommendation before the quality rating. No recommendation should come without appropriate interpretation (see table 2 for interpretation of the strength of recommendation and the quality of evidence). Guideline developers can (and sometimes should) make strong recommendations on the basis of low or very low quality evidence, but GRADE demands that these situation should be transparently described.

Furthermore, we understand that *labelling* quality as *low* or *very low* may be a valid concern and that other descriptors such as the symbols we presented above (eg, ⊕⊕⊕◯ for moderate) may help overcome reluctance to accept the underlying evidence due to labelling issues. We therefore suggest alternatives such as symbols or letters. Details about the GRADE system are published elsewhere, but in this article we have provided a brief guide for those who are dealing with observational study designs.[7–12] The approach has been used by many groups in the public health and policy sector, including guideline panels at WHO.

What this study adds

The GRADE approach to assessing the quality of evidence and grading the strength of healthcare recommendations has been used widely. This article deals with queries regarding it's applicability to public health questions and how to move from studies of exposure-risk assessment to recommendations when using the GRADE framework with special consideration for the Bradford Hill criteria for causation.

Acknowledgements The authors thank Andrew D Oxman for helpful input.

Competing interests HJS is co-chair of the GRADE working group; he supports the implementation of the GRADE approach worldwide. From non-profit organisations he has accepted honoraria and consulting fees for activities in which his work with GRADE may be relevant. SH is a staff member of the WHO. The authors alone are responsible for the views expressed in this publication and they do not necessarily represent the decisions, policy or views of the WHO or other organisations. The conclusions in this article are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention. GG is co-chair of the GRADE working; he supports the implementation of the GRADE

approach worldwide. On behalf of McMaster University, he has accepted honoraria and consulting fees for activities in which his work with GRADE is relevant.

Contributors HJS drafted the first version of the article. FA made substantial contributions to the first draft. All other authors made important additional contributions.

Provenance and peer review Not commissioned; not externally peer reviewed.

REFERENCES

1. **Durrheim DN,** Reingold A. Modifying the GRADE framework could benefit public health. *J Epidemiol Community Health* 2010;**64**:387.
2. **Sur D,** Ochiai RL, Bhattacharya SK, *et al*. A cluster-randomized effectiveness trial of Vi typhoid vaccine in India. *N Engl J Med* 2009;**361**:335—44.
3. **Hebert PC,** Levin AV, Robertson G. Bioethics for clinicians: 23. Disclosure of medical error. *CMAJ* 2001;**164**:509—13.
4. **Lopez L,** Weissman JS, Schneider EC, *et al*. Disclosure of hospital adverse events and its association with patients' ratings of the quality of care. *Arch Intern Med* 2009;**169**:1888—94.
5. **Brozek J,** Oxman A, Schünemann HJ. GRADEpro. [Computer program]. Version 3.2 for Windows. http://mcmaster.flintbox.com/technology.asp?Page=3993 and http://www.cc-ims.net/revman/gradepro (accessed 28 Mar 2011).
6. **Schünemann H,** Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. *Health Res Policy Syst* 2006;**4**:21.
7. **Guyatt GH,** Oxman AD, Kunz R, *et al*. Going from evidence to recommendations. *BMJ* 2008;**336**:1049—51.
8. **Guyatt GH,** Oxman AD, Kunz R, *et al*. Incorporating considerations of resources use into grading recommendations. *Br Med J* 2008;**336**:1170—3.
9. **Guyatt GH,** Oxman AD, Kunz R, *et al*. What is "quality of evidence" and why is it important to clinicians? *Br Med J* 2008;**336**:995—8.
10. **Guyatt GH,** Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008;**336**:924—6.
11. **Schunemann HJ,** Hill SR, Kakad M, *et al*. Transparent development of the WHO rapid advice guidelines. *PLoS Med* 2007;**4**:e119.
12. **Schunemann HJ,** Oxman AD, Brozek J, *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Br Med J* 2008;**336**:1106—10.