Allergy 2009: 64: 669–677

DOI: 10.1111/j.1398-9995.2009.01973.x

Review article

Grading quality of evidence and strength of recommendations in clinical practice guidelines

Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions

The GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) approach provides guidance to grading the quality of underlying evidence and the strength of recommendations in health care. The GRADE system's conceptual underpinnings allow for a detailed stepwise process that defines what role the quality of the available evidence plays in the development of health care recommendations. The merit of GRADE is not that it eliminates judgments or disagreements about evidence and recommendations, but rather that it makes them transparent. This first article in a three-part series describes the GRADE framework in relation to grading the quality of evidence about interventions based on examples from the field of allergy and asthma. In the GRADE system, the quality of evidence reflects the extent to which a guideline panel's confidence in an estimate of the effect is adequate to support a particular recommendation. The system classifies quality of evidence as high, moderate, low, or very low according to factors that include the study methodology, consistency and precision of the results, and directness of the evidence.

J. L. Brożek[1,2], E. A. Akl[3], P. Alonso-Coello[4,5], D. Lang[6], R. Jaeschke[7], J. W. Williams[8], B. Phillips[9], M. Lelgemann[10], A. Lethaby[11], J. Bousquet[12,13], G. H. Guyatt[7,14], H. J. Schünemann[3,14], for the GRADE Working Group

[1]Department of Epidemiology, Italian National Cancer Institute Regina Elena, Rome, Italy; [2]Department of Medicine, Jagiellonian University School of Medicine, Krakow, Poland; [3]Department of Medicine and Social and Preventive Medicine, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA; [4]Iberoamerican Cochrane Center, Servicio de Epidemiología Clínica y Salud Pública, Hospital de Sant Pau, Barcelona, Spain; [5]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Spain; [6]Allergy/Immunology Section, Respiratory Institute, Cleveland Clinic, Cleveland, OH, USA; [7]Department of Medicine, McMaster University, Hamilton, ON, Canada; [8]Department of Internal Medicine and Psychiatry, Duke University and Durham VA Medical Center, Durham, NC, USA; [9]Centre for Reviews and Dissemination, University of York, York, UK; [10]HTA-Zentrum, Universität Bremen, Bremen, Germany; [11]School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, New Zealand; [12]Service des Maladies Respiratoires, Hôpital Arnaud de Villeneuve, Montpellier, France; [13]Inserm UMR 780, France; [14]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

Key words: clinical practice guidelines; evidence based medicine; grading.

Holger J. Schünemann MD PhD
Department of Clinical Epidemiology & Biostatistics
McMaster University Health Sciences Centre,
Room 2C10B
1200 Main Street West Hamilton
ON L8N 3Z5
Canada

Accepted for publication 16 November 2008

What do patients want?

When offered a diagnostic procedure or a treatment option, patients ask themselves and the clinicians taking care of them about the benefits and downsides of that choice. They ask: What will I gain? Will I feel better (reduced symptoms or morbidity and improved quality of life)? Will I live longer (reduced mortality)? Patients also ask: What will I lose? Is it safe and will I dislike some aspects related to the intervention (adverse events, burden – extra time and effort)? How much will it cost me? Thus, the decision to choose among the options depends on the balance between their desirable and undesirable consequences. This balance weighs not only what patients will gain or lose but also *how much* they will gain or lose (one can estimate it based on the evidence from current research), and how important are the gains and losses for them (patients' values and preferences for the different outcomes and interventions).

The role of a clinician is not only to order a diagnostic test or prescribe a treatment, but also to advise patients – sometimes to decide for them – which of the available tests or treatments is likely to be most beneficial and which one to choose.

As we cannot predict the future, we always have to make these decisions under uncertainty about the outcomes for a particular patient. Optimal decision-making requires informing these decisions with the best available evidence (i.e. information on the past experience of the effect of similar management of similar patients). Clinical practice guidelines can help clinicians and patients make these decisions but their application is not always easy as every patient is different.

Clinical practice guidelines

Clinical practice guidelines offer recommendations for diagnostic procedures or treatment options for typical patients. They are 'systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances' (1). The purpose of guidelines is 'to make explicit recommendations with a definite intent to influence what clinicians do' (2). Clinical decisions – and also the related recommendations and their strength – depend on both the research evidence and the values and preferences of patients. For clinicians and patients to be confident that following these recommendations will do more good than harm, guidelines need to be evidence-based, transparent, and explicit about whose values and preferences were taken into account and how they influenced the final recommendations. Systematic approach, transparency, and explicitness also facilitate implementation, adaptation to local circumstances, and updating of guidelines (3).

GRADE approach

Guideline panels develop recommendations on the basis of the balance between the desirable and the undesirable consequences of the diagnostic or therapeutic options in question. They will recommend the option that results in greater net benefit and recommend against the option that results in greater net loss. The strength of their recommendation will depend on the extent to which they can be confident that desirable effects outweigh undesirable effects, or vice versa. A systematic approach to grading the strength of recommendations can minimize bias and aid interpretation (4, 5). The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) working group has conducted a review of existing grading systems and developed a system for grading the quality of evidence and strength of recommendations that addresses shortcomings of prior systems (4, 6–9). The resulting GRADE system has a number of advantages over other grading systems (Table 1). These advantages are reflected in the increasing number of professional societies and organizations endorsing or using the GRADE system – examples include the American College of Chest Physicians (ACCP) (10), the American Thoracic Society (ATS) (7), the British Medical Journal (11), the Cochrane Collaboration (12), the Endocrine Society (13), the European Respiratory Society (ERS), Infectious Disease Society of America (IDSA), Surviving Sepsis Campaign (14), UpToDate® (15), and the World Health Organization (WHO) (16) among others (a comprehensive list of endorsing organizations is available at http://www.gradeworkinggroup.org). Most recently, the Allergic Rhinitis and its Impact on Asthma

Table 1. Merits of the GRADE system for grading quality of evidence and strength of recommendations in comparison to other systems

1. Clear separation between quality of evidence and strength of recommendations*
2. Explicit and comprehensive criteria for downgrading or upgrading quality of evidence
3. Explicit consideration of the relative importance of various outcomes to patients
4. Explicit acknowledgement of values and preferences assumed when making recommendations
5. Transparent process of moving from evidence to recommendations
6. Explicit advice to make recommendations about the most appropriate course of action, even when very little evidence is available
7. Grading the strength only for recommendations about the diagnostic or therapeutic course of action, but not about prognosis or etiology
8. Clear and pragmatic interpretation of 'strong' and 'weak' recommendations
9. Balance between simplicity and methodological comprehensiveness

*In the context of clinical practice guidelines: Quality of evidence, the extent to which our confidence in an estimate of the treatment effect is adequate to support particular recommendation; Strength of recommendation, the extent to which we can, across the range of patients for whom the recommendations are intended, be confident that desirable consequences of an intervention outweigh undesirable consequences (or *vice versa*, that undesirable consequences outweigh desirable ones – in this case one would recommend against this intervention).

(ARIA) guideline panel decided to use the GRADE system for the 2009 revision of the guidelines and to follow the GRADE approach in the future (17).

We suggest conceptualizing GRADE as a *system* of grading quality of evidence and also as a systematic and transparent *approach* to the process of developing recommendations for clinical practice including indicating the strength of these recommendations (Table 2).

In this series of three articles, we will present the GRADE approach to transparent development of evidence-based recommendations. In this article, we will start with a brief overview of GRADE approach and we will discuss grading the quality of available evidence supporting the recommendations about therapeutic interventions. In a second article, we will present the approach to grading the quality of available evidence about diagnostic strategies. In a third article, we will present the GRADE approach to formulating the recommendations, deciding on their strength, and suggested interpretation for clinical practice.

As many guidelines are adopting GRADE, including the 2009 revision of ARIA, it is important that allergists understand the underlying concepts. Therefore, this series is intended for clinicians especially interested in allergy, who want to be able to fully interpret the recommendations in guidelines developed following the GRADE approach. Whenever possible, we will use examples specific to the field of allergy and asthma.

For clinicians interested in more in depth review of the GRADE approach and system, we recommend a series of articles recently published in the British Medical Journal (18–22) or an even more detailed series for guideline developers that will be published in the Journal of Clinical Epidemiology, and the earlier papers (4, 7).

Overview of the GRADE approach

Ask precise clinical questions

Following the GRADE approach, one begins with formulating appropriate clinical questions that the recommendations would answer. As guidelines include recommendations about the most appropriate course of action, they should answer clinical management questions about diagnosis or treatment of disease, but not about prognosis or etiology. A clinical management question should have four components: patient population, intervention (diagnostic or therapeutic), alternative intervention (comparison), and the outcomes of interest (23). For instance, consider the following: in patients with persistent allergic rhinitis (patient population) should oral H_1-antihistamines (intervention) *vs* no oral H_1-antihistamines (alternative intervention) be used to improve quality of life, reduce symptoms, and minimize the adverse effects (outcomes of interest)?

There are potential problems arising at the stage of asking a clinical question. One is the failure to consider all relevant alternatives. This may be particularly important in international guidelines where treatment options vary for patients in many diverse jurisdictions. Two other closely related mistakes in formulating questions are the failure to include all patient-important outcomes, e.g. disregarding quality of life or adverse effects, and placing excessive emphasis on surrogate outcomes with questionable importance to patients such as pulmonary function or nasal airway resistance rather than objectively measured quality of life or symptoms.

Table 2. An overview of steps followed during the development of an evidence-based clinical practice guideline

Establish the guideline panel (59)
Define the scope of the guidelines
Prioritize the problems (60)
Ask precise clinical questions
Decide on the relative importance of outcomes
Identify the existing evidence for every clinical question
Develop evidence profiles
Grade the quality of existing evidence for each outcome separately
Determine the overall quality of available evidence across outcomes
Decide on the balance between desirable and undesirable consequences
Decide on the strength of recommendation
Formulate the recommendation reflecting its strength
Write guideline

Decide on the relative importance of outcomes

The GRADE approach asks guideline developers to make explicit judgments about the importance of each outcome for making a recommendation. GRADE demands that those making recommendations classify each of the outcomes of interest as either critical for making a recommendation, important but not critical, or not important (24). Because experts, clinicians, and patients differ in their preferences and how they value particular outcomes (25), input from those affected by the recommendation (i.e. patients, their families, or members of the public) should be sought if possible. For example, outcomes such as mortality, quality of life, or exacerbations of asthma might be considered critical, nasal symptoms judged by a physician or use of rescue medications – important but not critical, and peak expiratory flow or nasal eosinophilia – not important, but perhaps informative, for making a recommendation.

Identify the existing evidence for every clinical question

Every clinical question should then be answered based on a systematic review of the relevant evidence (26). Guideline developers can either conduct the systematic review themselves or identify an existing high quality systematic review. This systematic review will serve to create a summary of available evidence that a guideline panel can use to inform judgments about the balance of desirable and undesirable effects in order to develop a

recommendation. GRADE suggests that this summary of evidence is prepared in a structured format of evidence profile – a table including detailed judgments about quality of evidence and estimates of the effect for each outcome.

Grade the quality of existing evidence

The recommendation and its strength depend not only on the best estimates of the expected benefits and downsides, but also on the confidence in these estimates. If we know the best estimates of the magnitude of the effects, but we have no confidence in these estimates (i.e. we do not 'believe' in them), it is very difficult to determine the balance of desirable and undesirable consequences.

One of the factors that influence our confidence in the estimates of treatment effects is the quality of supporting evidence – the higher it is, the more confidence we have in these estimates. Formal grading of the quality of evidence and its explicit consideration are essential to the process of developing recommendations. The examples of errors arising from disregarding the quality of supporting evidence are abundant in the modern history of medicine. Consider the treatment of patients with myocardial infarction. For about a decade, experts made recommendations ignoring the high quality evidence about the benefit from thrombolysis or the lack of benefit, and possibly even harm, from routine administration of antiarrhythmic agents in the early postmyocardial infarction period (27). They based their judgments on pathophysiological considerations, such as reduction in the frequency of arrhythmia that failed to recognize higher quality evidence focusing on patient important outcomes including mortality. In the field of allergy and asthma, there are less dramatic examples. As an illustration of misleading conclusions from relying on lower quality evidence, one might consider a systematic review of observational studies assessing the effect of inhaled or oral corticosteroids on height in children with asthma concluding that the use of inhaled beclomethasone dipropionate was not associated with diminished stature (28). However, a recent systematic review restricted to randomized trials found a statistically significant decrease in linear growth velocity in children with mild to moderate asthma treated with moderate doses of beclomethasone (29). A formal system of grading the quality of evidence provides a strategy to clarify how reliable is the evidence supporting the recommendations, thereby decreasing the risk of repeating the types of mistakes described above.

In the context of clinical practice guidelines, the GRADE system defines quality of evidence as the extent to which our confidence in an estimate of the treatment effect is adequate to support a particular recommendation. The GRADE system specifies four grades of evidence: high, moderate, low, and very low quality (Table 3). Acknowledging that the quality of evidence is in fact a continuum and any categorization involves arbitrariness and the possibility of oversimplification, we think that clarity, transparency, and intuitive understanding of the four categories outweigh these limitations.

Table 3. Quality of evidence and the explanation of the categories

Rank	Explanation	Examples*
High	Further research is very unlikely to change our confidence in the estimate of effect	Randomized trials without serious limitations Well-performed observational studies with very large effects (or other qualifying factors)
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate	Randomized trials with serious limitations Well-performed observational studies yielding large effects
Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate	Randomized trials with very serious limitations Observational studies without special strengths or important limitations
Very low	Any estimate of effect is very uncertain	Randomized trials with very serious limitations and inconsistent results Observational studies with serious limitations Unsystematic clinical observations (e.g. case series or case reports)

*The examples are not comprehensive. See text for criteria to downgrade or upgrade the quality of evidence.

Study design. Earlier systems of grading the quality of evidence relied almost exclusively on overall study design (e.g. randomized trials *vs* observational studies). In the GRADE system, study design remains a critical, but not a sole factor in judging the quality of evidence (Table 4). For recommendations about alternative treatment options, randomized trials provide, in general, far stronger evidence than observational studies, yet rigorous observational studies provide far stronger evidence than uncontrolled case series. Therefore, in the GRADE system, a body of evidence obtained from randomized trials is initially rated as high quality, and that obtained

Table 4. Factors influencing the quality of evidence

Study design (experimental *vs* observational)
Factors that can decrease the quality
Limitations in study design and/or execution
Inconsistency of results
Indirectness of evidence
Imprecision of results
Publication bias
Factors that can increase the quality of evidence
Large magnitude of effect
All plausible confounding may be working to reduce the demonstrated effect or increase the effect if no effect was observed
Dose-response gradient

from observational studies – as low quality. For example, a systematic review of the effect of using feather bedding in the control of asthma symptoms identified no randomized trial addressing this clinical question (30). The only available evidence indicating that more frequent wheezing is associated with nonfeather pillows comes from two case–control studies that found a 20% rise in the population prevalence odds of wheezing from 1978 to 1991, and identified an increase (from 44% to 67%) in the use of nonfeather pillows as the only domestic indoor exposure that appeared to explain this (31). The initial rating for this evidence would be low quality.

In the GRADE system, 'expert opinion' is not a category of quality of evidence, but an interpretation of existing evidence. Therefore, expert opinion is nearly always necessary to integrate and contextualize evidence, either from a clinical or from a methodological viewpoint.

A well-designed and executed randomized trial or observational study provides different quality evidence than the one that was poorly conducted. Therefore, relying on study design alone has apparent limitations. GRADE provides additional quality criteria that serve to overcome this shortcoming. We have identified five factors that can reduce the quality of evidence for each study design and three that can increase it (Table 4).

Limitations in study design and/or execution (risk of bias). Quality of evidence initially rated based on study design decreases when studies suffer from major methodological limitations that can bias their estimates of the treatment effect. These limitations include lack of allocation concealment, lack of blinding – particularly if outcomes are subjective and their assessment is highly susceptible to bias, lack of accounting for a large proportion of patients who started the study (large loss to follow-up or outcome not measured in a large proportion of patients), failure to adhere to the intention-to-treat principle during the analysis, stopping early for benefit, or selectively reporting outcomes that show an apparent treatment effect and failing to report other outcomes that show no evident effect (32–36). For example, the evidence for the effect of sublingual immunotherapy in children with allergic rhinitis on the development of asthma, comes from a single randomized trial with no description of randomization, concealment of allocation, type of analysis, no blinding, and 21% of children lost to follow-up (37). These very serious limitations would warrant downgrading the quality of evidence by two levels (i.e. from high to low). In another example, a systematic review showed that the family therapy for children with asthma improved outcomes such as daytime wheeze and the number of functionally impaired days. However, allocation was clearly not concealed in one of the two included trials and unclear whether it was concealed in the second trial (38). This limitation might warrant downgrading the quality of evidence by one level.

Inconsistency of results. Widely differing estimates of the treatment effect across individual studies (variability or heterogeneity of results) suggest true differences in underlying treatment effects (39). Authors of a systematic review should try to identify plausible explanations for inconsistent results but, if they do not succeed, the quality of evidence decreases. Variability in individual study results may arise from clinical differences in populations (e.g. drugs may have larger relative effects in sicker patients), interventions (e.g. larger effects or larger side-effects with higher drug doses), and outcome measures (e.g. differences in the definition of 'response to treatment'), or from methodological limitations such as problems with randomization, early termination of trials, or publication bias (40, 41). For example, a systematic review of subcutaneous allergen-specific immunotherapy in adults with allergic rhinitis found inconsistency in the effect of treatment on nasal symptoms that was suggested both by visual examination of forest plot and statistical tests. Despite the effort to find a reason for this heterogeneity, authors of the systematic review were not able to explain it (42). In another example, a systematic review showed that ketotifen reduced the use of bronchodilators in children with mild to moderate asthma. However, there was significant heterogeneity among the results of individual trials (I^2 = 76.1%) (43). Subgroup and sensitivity analyses explained this heterogeneity – the effect was stronger in school children than in infants or preschool children (differences in populations) and it disappeared in trials with adequate blinding (differences in study limitations).

Indirectness of evidence. GRADE distinguishes two types of indirectness – indirect comparisons and differences in populations, interventions, and outcomes of interest between the studies (existing evidence) and the scope of the recommendation (clinical question).

Indirect comparison arises when, for instance, the recommendation addresses the choice between two active drugs: A or B, but the available studies compared A *vs* placebo and B *vs* placebo. Such trials allow indirect comparison of the magnitude of the effect of A *vs* B. Such an indirect comparison provides lower quality evidence than a head-to-head comparison of A *vs* B would provide. This type of indirectness is common when choosing between the drugs within the same class (e.g. long acting β-agonists, oral or topical H_1-antihistamines, allergen extracts for immunotherapy, etc.). As an illustration, one might consider allergen-specific immunotherapy in patients with severe allergic rhinitis. Systematic review of the studies in seasonal allergic rhinitis showed a consistent small to large effect of subcutaneous allergen-specific immunotherapy (SCIT) compared with placebo on symptoms of allergic rhinitis, ocular symptoms, and quality of life (42). Another systematic review showed that sublingual allergen-specific immunotherapy (SLIT) is also effective in reducing symptoms and medication

Table 5. Sources of likely indirectness of evidence

Source of indirectness	Question of interest	Example
Indirect comparison	Early administration of systemic corticosteroids in the emergency department to treat acute exacerbations in adult patients with asthma	Both oral and intravenous routes are effective but there is no direct comparison of these two routes of administration in adults
Differences in populations	Oral H_1-antihistamines for improving quality of life in adults with asthma and concomitant allergic rhinitis	In the only study that measured quality of life, 60% of patients had a past history of asthma but no symptoms of asthma at the beginning of a trial
	Ketotifen for long-term control of symptoms and wheeze in children with asthma	Inhaled corticosteroids, the mainstay of therapy of asthma nowadays, were allowed as an additional intervention in 60% of trials assessing ketotifen. There was no enough information to assess the effect of ketotifen as an add-on therapy in children with asthma on inhaled corticosteroids
	Anti-leukotrienes plus inhaled glucocorticosteroids *vs* inhaled glucocorticosteroids alone to prevent asthma exacerbations and nighttime symptoms in patients with chronic asthma and allergic rhinitis	Trials that measured asthma exacerbations and nighttime symptoms did not include patients with allergic rhinitis
Differences in intervention	Avoidance of pet allergens in nonallergic infants or preschool children to prevent development of allergy	Available studies used multifaceted interventions directed at multiple potential risk factors in addition to pet avoidance
	Oral decongestant as a rescue medication in patients with allergic rhinitis	Available studies used oral decongestants administered regularly, but none investigated their use as a rescue medication for quick alleviation of the symptoms
Differences in outcomes of interest	Intranasal glucocorticosteroids *vs* oral H_1-antihistamines in children with seasonal allergic rhinitis	In the available study, parents were rating the symptoms and quality of life of their teenage children, instead of the children themselves

requirements in these patients (44); however, the magnitude of the benefit achieved with SLIT compared with that of SCIT is not clear, because they have been compared directly in only very few studies (45, 46).

Evidence supporting the recommendation is also indirect when it comes from studies in which population, intervention, alternative intervention, or outcomes of interest were different from those that the recommendation refers to (Table 5).

Imprecision of results. When studies include relatively few patients and few events occur, estimates of the effect usually have wide confidence intervals that include both important benefits or no important effects (or even important harm). With such indeterminate results, one can judge the quality of the evidence lower than one otherwise would, because of resulting uncertainty in the effect. For instance, observational studies examining the impact of exclusive breast feeding on the development of allergic rhinitis in high-risk infants showed a relative risk of 0.87 (95% CI: 0.48–1.58) that rules out neither important benefit nor important harm (47).

Publication bias. The quality of evidence will be reduced when investigators fail to report (publish) studies they have undertaken – typically those that showed no effect. Unfortunately, one must often guess about the likelihood of publication bias. The risk of publication bias is higher when only few small studies are available (48–50). For example, a systematic review of topical treatments for seasonal allergic conjunctivitis showed that patients using topical sodium cromoglicate were more likely to perceive benefit than those using placebo. However, only small trials reported clinically and statistically significant benefits of active treatment, while a larger trial showed a much smaller and a statistically not significant effect (51). These findings suggest that smaller studies demonstrating smaller effects might not have been published.

Evidence supporting a particular recommendation can suffer from more than one of the above limitations, and the more serious they are, the lower the quality of the evidence is. For example, randomized trials of high efficiency particulate air filters in patients with perennial allergic rhinitis suffered very serious limitations in design (warranting downgrading by 2 levels) and the results were imprecise (52).

The GRADE system offers three criteria that, when fulfilled, can increase the quality of evidence. They are infrequently applicable, but are the most common reason for upgrading the quality of evidence from well-performed observational studies that without these additional merits would provide only low quality evidence.

Large magnitude of effect. On rare occasions, when studies yield large or very large estimates of the magnitude of the effect, one may be more confident about the results. Based on modeling studies that provide estimates of the magnitude of effect that is very unlikely to be explained by bias (53, 54), the GRADE system defines a large effect as a relative risk (RR) of > 2.0 or < 0.5 (based

on consistent evidence from at least two studies, with no plausible confounders) and a very large effect as a RR of >5.0 or <0.2 (based on direct evidence with no major threats to validity). For example, the extremely large and consistent effect of epinephrine injections in anaphylactic shock leaves us convinced of the benefits of the intervention.

All plausible confounding would reduce the demonstrated effect or increase it if no effect was observed. On rare occasions, all plausible biases may be working to underestimate the true treatment effect. For instance, if only sicker patients receive an experimental intervention, yet they still fare better than patients not receiving it do, it is likely that the actual effect may be larger than the data suggest. There are few examples so far in the literature and we were not able to identify one in the field of asthma and allergy. However, consider a systematic review of observational studies that included 38 million patients, which demonstrated higher death rates in private for-profit *vs* private not-for-profit hospitals (55). Biases related to different disease severity in patients in the two hospital types, and the spill-over effect from well-insured patients would both lead to estimates in favor of for-profit hospitals (56). One might therefore consider the evidence from these observational studies higher than low quality. Because the plausible biases would all diminish the demonstrated intervention effect, one might consider the evidence from these observational studies as moderate rather than low quality. A parallel situation exists when observational studies have failed to demonstrate an association but all plausible biases would have increased an intervention effect. This situation will usually arise in the exploration of apparent harmful effects. For example, because the hypoglycemic drug phenformin causes lactic acidosis, the related agent metformin is under suspicion for the same toxicity. Nevertheless, very large observational studies have failed to demonstrate an association (57). Given the likelihood that clinicians would be more alert to lactic acidosis in the presence of the agent and over-report its occurrence, one might consider this moderate or even high quality evidence refuting a causal relationship between typical therapeutic doses of metformin and lactic acidosis.

Dose-response gradient. The presence of a dose-response gradient may also increase one's confidence in the findings and thereby increase the quality of evidence. Most evidence for dose-response gradient in the treatment of allergic diseases comes from well-performed randomized trials that do not require upgrading. However, consider the following example: there are no studies of interventions aimed at reduction of second-hand tobacco smoke exposure that examined development of asthma or wheeze in children. On the other hand, observational studies found an increased risk of developing wheeze illnesses in early childhood when exposed to second-hand smoke from parental smoking. Moreover, the greater the exposure, the higher was the risk. While grading the quality of available evidence supporting the recommendation to reduce second-hand smoke in children, one might consider these results as indirect evidence of benefit from reducing the second-hand tobacco smoke exposure and initially rate it as low quality evidence from observational studies that is downgraded to very low because of indirectness (evidence of increased risk with increased exposure rather than of benefit with reduced exposure). The observed dose-response gradient would justify upgrading the quality of evidence back to low.

Determine the overall quality of evidence across outcomes

Each recommendation depends on the evidence about outcomes identified when asking clinical questions and regarded as important to patients. Following the GRADE process, those making recommendations first grade the quality of available evidence supporting each outcome separately. Subsequently, they specify the overall quality of evidence across these multiple outcomes, because guidelines provide a single grade of quality of evidence for each recommendation. For any recommendation, when the quality of evidence differs across outcomes, the GRADE system demands that the lowest grade of quality of available evidence for any of the outcomes deemed critical determines the overall quality of evidence supporting this recommendation. For example, based on a systematic review of monoclonal anti-IgE for chronic asthma in adults and children (58), one might grade the quality of evidence about asthma symptoms, exacerbations, and quality of life as high, but the quality of evidence about adverse effects as moderate. Consequently, the overall quality of evidence supporting the recommendation about the use of this treatment would be moderate.

Conclusions

The GRADE approach provides a comprehensive, explicit, and transparent methodology for grading the quality of evidence and strength of recommendations about the management of patients. GRADE classifies the quality of available evidence as high, moderate, low or very low. Although judgments are required at every step of guideline development, a systematic and explicit approach to grading the quality of evidence facilitates scrutiny and transparency of these judgments.

In the next article in this series, we will discuss the GRADE approach to making recommendations about diagnostic methods in more detail and we will highlight the differences in grading the quality of available evidence between therapeutic and diagnostic interventions.

References

1. Field MJ, Lohr KN. Clinical practice guidelines: directions for a new program. Washington, DC: National Academy Press, 1990.
2. Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. JAMA 1995;**274**:570–574.
3. Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 13. Applicability, transferability and adaptation. Health Res Policy Syst 2006;**4**:25.
4. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S et al. Grading quality of evidence and strength of recommendations. BMJ 2004;**328**:1490.
5. Schünemann HJ, Vist GE, Jaeschke R, Kunz R, Cook DJ, Guyatt GH. Grading recommendations. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd edn. US: The McGraw-Hill Professional, 2008: 679–701.
6. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004;**4**:38.
7. Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 2006;**174**:605–614.
8. Schünemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;**169**:677–680.
9. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations? ACP J Club 2006;**144**: A8–A9.
10. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B et al. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians task force. Chest 2006;**129**:174–181.
11. BMJ Publishing Group Ltd. Article requirements. 2006 [cited 2008 February 11]; Available at: http://resources.bmj.com/bmj/authors/article-submission/article-requirements.
12. Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P et al. 12. Interpreting results and drawing conclusions [preliminary draft, 16 October 2007]. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1 [Updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.
13. The Endochrine Society. Evidence-based guidelines chart new course for endochrinology. Endochrine News, 2005;**30**:10.
14. Dellinger RP, Levy MM, Carlet JM, Bion J, Parker MM, Jaeschke R et al. Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock: 2008. Crit Care Med 2008;**36**:296–327.
15. UpToDate. Editorial Policy. 2006 [cited 2008 February 11]; Available at: http://www.uptodate.com/service/editorial_policy.asp.
16. World Health Organization. Guidelines for WHO guidelines [EIP/GPE/EQC/2003.1]Geneva, Switzerland: Global Programme on Evidence for Health Policy, WHO, 2003.
17. Brozek JL, Baena-Cagnani CE, Bonini S, Canonica GW, Rasi G, van Wijk RG et al. Methodology for development of the Allergic Rhinitis and its Impact on Asthma guideline 2008 update. Allergy 2008;**63**:38–46.
18. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A et al. Going from evidence to recommendations. BMJ 2008;**336**:1049–1051.
19. Guyatt GH, Oxman AD, Kunz R, Jaeschke R, Helfand M, Liberati A et al. Incorporating considerations of resources use into grading recommendations. BMJ 2008;**336**:1170–1173.
20. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is ''quality of evidence'' and why is it important to clinicians? BMJ 2008;**336**:995–998.
21. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;**336**:924–926.
22. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;**336**:1106–1110.
23. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. CMAJ 1988;**138**:697–703.
24. Schünemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L et al. Transparent development of the WHO rapid advice guidelines. PLoS Med 2007;**4**:e119.
25. Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 2001;**323**:1218–1222.
26. Guyatt GH, Jaeschke R, Prasad K, Cook DJ. Summarizing the evidence. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd edn. US: The McGraw-Hill Professional, 2008:523–542.
27. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 1992;**268**:240–248.
28. David BA, MaryLou M, Brian M. A meta-analysis of the effect of oral and inhaled corticosteroids on growth. The Journal of allergy and clinical immunology 1994;**93**:967–976.
29. Sharek PJ, Bergman DA, Ducharme F. Beclomethasone for asthma in children: effects on linear growth. Cochrane Database of Systematic Reviews 1999, Issue 3. Art. No.: CD001282. DOI: 10.1002/14651858.CD001282.
30. Campbell F, Gibson P. Feather versus non-feather bedding for asthma. Cochrane Database of Systematic Reviews 2000, Issue 4. Art. No.: CD002154. DOI: 10.1002/14651858.CD002154.
31. Butland BK, Strachan DP, Anderson HR. The home environment and asthma symptoms in childhood: two population based case-control studies 13 years apart. Thorax 1997;**52**:618–624.

32. Guyatt GH, Straus S, Meade MO, Kunz R, Cook DJ, Devereaux PJ, et al.. Therapy. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. US: The McGraw-Hill Professional, 2008:67–86.
33. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;**294**:2203–2209.
34. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005;**330**:753.
35. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;**291**:2457–2465.
36. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2004;**171**:735–740.
37. Novembre E, Galli E, Landi F, Caffarelli C, Pifferi M, De Marco E et al. Coseasonal sublingual immunotherapy reduces the development of asthma in children with allergic rhinoconjunctivitis. J Allergy Clin Immunol 2004;**114**:851–857.
38. Yorke J, Shuldham C. Family therapy for asthma in children. Cochrane Database of Systematic Reviews 2005, Issue 2. Art. No.: CD000089. DOI: 10.1002/14651858.CD000089.pub2.
39. Montori V, Hatala R, Ioannidis JP, Meade MO, Wyer P, Guyatt GH. Advanced topics in systematic reviews. Making sense of variability in study results. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd edn. US: The McGraw-Hill Professional, 2008:563–570.
40. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;**327**:557–560.
41. Fletcher J. What is heterogeneity and is it important? BMJ 2007;**334**:94–96.
42. Calderon M, Alves B, Jacobson M, Hurwitz B, Sheikh A, Durham S. Allergen injection immunotherapy for seasonal allergic rhinitis. Cochrane Database Syst Rev 2007;**1**:CD001936.
43. Bassler D, Mitra A, Ducharme FM, Forster J, Schwarzer G. Ketotifen alone or as additional medication for long-term control of asthma and wheeze in children. Cochrane Database Syst Rev 2004;**1**:CD001384.
44. Wilson DR, Torres LI, Durham SR. Sublingual immunotherapy for allergic rhinitis. Cochrane Database Syst Rev 2003;**2**:CD002893.
45. Mungan D, Misirligil Z, Gurbuz L. Comparison of the efficacy of subcutaneous and sublingual immunotherapy in mite-sensitive patients with rhinitis and asthma – a placebo controlled study. Ann Allergy Asthma Immunol 1999;**82**:485–490.
46. Quirino T, Iemoli E, Siciliani E, Parmiani S, Milazzo F. Sublingual versus injective immunotherapy in grass pollen allergic patients: a double blind (double dummy) study. Clin Exp Allergy 1996;**26**:1253–1261.
47. Mimouni Bloch A, Mimouni D, Mimouni M, Gdalevich M. Does breastfeeding protect against allergic rhinitis during childhood? A meta-analysis of prospective studies. Acta Paediatr 2002;**91**:275–279.
48. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989;**81**:107–115.
49. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj 1997;**315**:629–634.
50. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;**2**:e124.
51. Owen CG, Shah A, Henshaw K, Smeeth L, Sheikh A. Topical treatments for seasonal allergic conjunctivitis: systematic review and meta-analysis of efficacy and effectiveness. Br J Gen Pract 2004;**54**:451–456.
52. Sheikh A, Hurwitz B. House dust mite avoidance measures for perennial allergic rhinitis. Cochrane Database Syst Rev 2001;**4**:CD001563.
53. Bross ID. Pertinency of an extraneous variable. J Chronic Dis 1967;**20**:487–495.
54. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007;**334**:349–351.
55. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schunemann HJ, Haines T et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 2002;**166**:1399–1406.
56. Devereaux PJ, Schunemann HJ, Ravindran N, Bhandari M, Garg AX, Choi PT et al. Comparison of mortality between private for-profit and private not-for-profit hemodialysis centers: a systematic review and meta-analysis. JAMA 2002;**288**:2449–2457.
57. Salpeter S, Greyber E, Pasternak G, Salpeter E. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. Cochrane Database Syst Rev 2006;**1**:CD002967.
58. Walker S, Monteil M, Phelan K, Lasserson TJ, Walters EH. Anti-IgE for chronic asthma in adults and children. Cochrane Database Syst Rev 2006;**2**:CD003559.
59. Fretheim A, Schünemann HJ, Oxman AD. Improving the use of research evidence in guideline development: 3. Group composition and consultation process. Health Res Policy Syst 2006;**4**:15.
60. Oxman AD, Schünemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 2. Priority setting. Health Res Policy Syst 2006;**4**:14.

