An Official ATS Statement: Grading the Quality of Evidence and Strength of Recommendations in ATS Guidelines and Recommendations

Holger J. Schünemann, Roman Jaeschke, Deborah J. Cook, William F. Bria, Ali A. El-Solh, Armin Ernst, Bonnie F. Fahy, Michael K. Gould, Kathleen L. Horan, Jerry A. Krishnan, Constantine A. Manthous, Janet R. Maurer, Walter T. McNicholas, Andrew D. Oxman, Gordon Rubenfeld, Gerard M. Turino, and Gordon Guyatt, on behalf of the ATS Documents Development and Implementation Committee

THIS OFFICIAL STATEMENT OF THE AMERICAN THORACIC SOCIETY (ATS) WAS ADOPTED BY THE ATS BOARD OF DIRECTORS, DECEMBER 2005

Grading the strength of recommendations and the quality of underlying evidence enhances the usefulness of clinical practice guidelines. Professional societies and other organizations, including the American Thoracic Society (ATS), should reach consensus about whether they will use one common grading system and which of the numerous grading systems they would apply across all guidelines. The profusion of guideline grading systems confuses consumers of guidelines, and undermines the value of the grading exercise in conveying a transparent message. In response to this dilemma, the international GRADE working group has developed an approach that is useful for many guideline contexts, and that several national and international organizations have adopted. The GRADE system classifies recommendations as strong or weak, according to the balance of the benefits and downsides (harms, burden, and cost) after considering the quality of evidence. The quality of evidence reflects the confidence in estimates of the true effects of an intervention, and the system classifies quality of evidence as high, moderate, low, or very low according to factors that include the study methodology, the consistency and precision of the results, and the directness of the evidence. On recommendation of the ATS Documents Development and Implementation Committee, the ATS adopted the GRADE approach for its guidelines in line with many other organizations that have recently chosen the GRADE approach. This article informs ATS guideline developers, investigators, and those interpreting future ATS guidelines that follow the GRADE approach about the methodology and applicability of ATS guidelines and recommendations.

Am J Respir Crit Care Med Vol 174. pp 605–614, 2006
DOI: 10.1164/rccm.200602-197ST
Internet address: www.atsjournals.org

INTRODUCTION

Clinical practice guidelines (CPGs) offer recommendations for the management of typical patients. These management decisions involve balancing the expected benefits and downsides (harms, burden, and costs). To make evidence-based medical decisions, clinicians also need to integrate recommendations with their own clinical judgment, and with individual patient circumstances, values, and preferences (1). A systematic approach to grading the strength of management recommendations can minimize bias and aid interpretation (2, 3). Most guideline developers, including the American Thoracic Society (ATS), recognize the need for grading, and journals are increasingly demanding such systems for publication of guidelines and recommendations. The ATS Documents Development and Implementation Committee was charged with developing, adapting or identifying, and adopting a grading system that will guide ATS panels in the development of recommendations and help clinicians interpret the recommended actions (4–6).

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 of CPGs that addresses disadvantages of prior systems (2, 7, 8). These disadvantages include the lack of separation between quality of evidence and strength of recommendation, the lack of transparency about judgments, and the lack of explicit acknowledgment of values and preferences (2, 7, 9). The aim of the independent GRADE group includes reducing confusion among guideline panels and users as a result of the existence of many, often scientifically outdated, grading systems. Following the comprehensive assessment, development, and dissemination of the work of the GRADE group, several organizations and guideline developers, including the World Health Organization, the American College of Chest Physicians (ACCP), the American Endocrine Society, and UpToDate, have adopted the GRADE system in its original format or with relatively minor modifications.

The GRADE system is based on a sequential assessment of the quality of evidence, followed by assessment of the balance between benefits versus downsides and subsequent judgment about the strength of recommendations. Because frontline consumers of recommendations will be most interested in the best course of action, the GRADE system places the strength of the recommendation first, followed by the quality of the evidence. Separating the judgments regarding the quality of evidence from judgments about the strength of recommendations is a critical and defining feature of this new grading system. The newly formed

standing ATS Documents Development and Implementation Committee agreed to adopt the GRADE approach developed by the GRADE working group based on these issues of methodology, practicality, and applicability. The ATS leadership has selected several members of the GRADE working group who are involved in disseminating the approach and collaborated with other organizations, including the ACCP, to serve on this committee (4–6, 9, 10). The first project of this committee is described in this document and informs ATS guideline developers, investigators, and those interpreting future ATS guidelines that follow the GRADE approach in greater detail than prior documents (9). Specifically, this document describes the GRADE approach and factors that influence the process of grading based on several examples. This document does not describe the way consensus is reached by a guideline panel during a guideline development process.

BOX 1. EXAMPLE: STRONG RECOMMENDATION

Thromboprophylaxis reduces the incidence of venous thromboembolism in immobile, hospitalized, severely ill medical patients. Careful thromboprophylaxis has minimal side effects and relatively low cost while being very effective at preventing deep venous thrombosis (DVT) and its sequelae. Peoples' values and preferences are such that virtually all patients admitted to a hospital would, if they understood the choice they were making, opt to receive some form of thromboprophylaxis. CPG groups can thus offer a strong recommendation for thromboprophylaxis for patients in this setting.

STRENGTH OF THE RECOMMENDATION

Guideline developers make recommendations to administer, or not administer, an intervention on the basis of tradeoffs between benefits on the one hand, and downsides (harms, burden, and cost) on the other. If benefits outweigh downsides, guideline panels will recommend that clinicians offer a treatment to appropriately chosen patients. Conversely, if downsides outweigh benefits, the guidelines will recommend against the implementation of such a treatment. The strength of a recommendation reflects the degree of confidence that the desirable effects of adherence to a recommendation outweigh the undesirable effects. Desirable effects can include beneficial health outcomes, less burden, and savings. Undesirable effects can include harms, more burden, and costs. Burdens are the demands of adhering to a recommendation that patients or caregivers (e.g., family) may dislike, such as having to take medication or the inconvenience of going to the doctor's office. Although the degree of confidence is a continuum, the GRADE approach classifies recommendations for or against treatments into two grades, strong and weak.

If guideline developers are confident that the desirable effects of adherence to a recommendation outweigh the undesirable effects, they will make a strong recommendation within the context of a described intervention. This confidence arises in several ways. High-quality evidence should provide precise estimates of both benefits and downsides, and the balance should be clear (recommendations to quit smoking to prevent adverse consequences of tobacco smoke exposure or recommendation for bronchodilators in patients with known chronic obstructive pulmonary disease [COPD]). A weak recommendation is one for which a guideline panel concludes that the desirable effects of adherence to a recommendation probably outweigh the undesirable effects, but the panel is not confident. Thus, if guideline developers believe that benefits and downsides are finely balanced, or appreciable uncertainty exists about the magnitude of benefits and/or downsides, they offer a weak recommendation.

CPGs are intended for typical patients, but clinicians are becoming increasingly aware of the importance of patient values and preferences in individualized clinical decision making. One way to interpret strong and weak recommendations is in relation to patient values and preferences. For decisions in which it is clear that benefits far outweigh downsides, or downsides far outweigh benefits, almost all patients will make the same choice, and guideline developers can offer a strong recommendation (*see* Box 1).

Thus, another way for clinicians to interpret strong recommendations is, for typical patients, that they should "just take the recommended action" and offer the intervention to their patients. On the other hand, when clinicians face weak recommendations, they should more carefully consider the benefits, harms, and burden in the context of the patient before them. These situations arise when benefits and downsides are closely balanced, or because of uncertainty in benefits and/or downsides, in which appreciable numbers of patients, because of variability in values and preferences, will make different choices. In such situations, guideline developers will offer weak recommendations (Box 2).

Individualization of clinical decision making in weak recommendations remains a challenge. Although clinicians always should consider patients' preferences and values, when they face weak recommendations they may have a more detailed conversation with patients than for strong recommendations to ensure that the ultimate decision is consistent with the patient's values. For patients who are interested, a decision aid that presents patients with both benefits and downsides of therapy is likely to improve knowledge and decrease decision-making conflict, and it may promote a decision most consistent with underlying values and preferences (13). Because of time constraints and because decision aids are not universally available, clinicians cannot use decision aids in all patients and, for strong recommendations, the use of decision aids is inefficient.

Other ways of interpreting strong and weak recommendations relate to performance or quality indicators. Strong recommendations are candidate performance indicators. For weak recommendations, performance could be measured by monitoring whether clinicians have discussed recommended actions with patients or their surrogates or carefully documented the evaluation of benefits and downsides in the patient's chart. Similar interpretations follow for public policy derived from guidelines. Strong recommendations require less debate than weaker recommendations. Table 1 summarizes several ways that developers and consumers of guidelines can interpret strong and weak recommendations.

Clinicians, patients, third-party payers, institutional review committees, other stakeholders, or the courts should never view recommendations as dictates. Even strong recommendations based on high-quality evidence will not apply to all circumstances and all patients. Consumers of CPGs may reasonably conclude that following some strong recommendations based on high quality will be a mistake for some patients. No CPGs or recommendations can take into account all of the often-compelling unique features of individual clinical circumstances. Thus, nobody charged with evaluating clinicians' actions should attempt to apply recommendations in rote or blanket fashion.

Factors that Influence the Strength of A Recommendation

One of the numerous factors guideline panels should include in the grading of recommendations is the confidence in the best

BOX 2. EXAMPLE A: CONSIDERING VALUES AND PREFERENCES IN RECOMMENDATIONS

Consider a 40-year-old man who has suffered an idiopathic DVT followed by treatment with adjusted-dose warfarin for 1 year to prevent recurrent DVT and pulmonary embolism (9). Continuing on standard-intensity warfarin beyond the treatment of 1 year will reduce his absolute risk for recurrent DVT by more than 7% per year for several years (11). The burdens of treatment include taking a warfarin pill daily, keeping dietary intake of vitamin K constant, monitoring the intensity of anticoagulation with blood tests, and living with the increased risk of both minor and major bleeding. Patients who are very averse to a recurrent DVT would consider the benefits of avoiding DVT worth the downsides of taking warfarin. Other patients are likely to consider the benefit not worth the harms and burden.

EXAMPLE B: CONSIDERING VALUES AND PREFERENCES IN RECOMMENDATIONS

Lung volume reduction surgery (LVRS) for severe emphysema offers another example of an intervention in which patient preferences and values play a central role in making treatment recommendations and decisions. Results of the only large-scale randomized controlled trial (RCT) to date indicate that lung resection, when combined with medical therapy, does not affect overall survival, although exercise capacity, quality of life, and other functional outcomes at 2 years are improved compared with medical therapy alone (12). However, surgery increases the risk of short-term mortality (5.2 vs. 1.5% at 90 days). In addition, the salutatory effects of surgery on functional outcomes appear to diminish with time. Thus, whereas some patients would be enthusiastic about undergoing LVRS because of the anticipated benefit in exercise capacity and quality of life, others who fear the risk of higher mortality in the early postsurgical phase may be less so. As in the example of anticoagulation for DVT, fully informed patients who are offered LVRS for severe emphysema are likely to make different choices regarding this procedure; guideline panels should therefore offer this treatment as a weak recommendation. Recommendations for or against LVRS may further differ by subgroups because secondary analyses suggest that outcomes are highly variable across subgroups defined by the anatomic distribution of emphysema and maximal exercise capacity prior to surgery.

TABLE 1. EXAMPLES OF IMPLICATIONS OF STRONG AND WEAK RECOMMENDATIONS FOR DIFFERENT GROUPS OF GUIDELINE USERS

Strong Recommendations

- For patients: Most individuals in this situation would want the recommended course of action and only a small proportion would not. Formal decision aids are not likely to be needed to help individuals make decisions consistent with their values and preferences.
- For clinicians: Most individuals should receive the intervention. Adherence to this recommendation according to the guideline could be used as a quality criterion or performance indicator.
- For policy makers: The recommendation can be adapted as policy in most situations.

Example: Early anticoagulation in patients with deep venous thrombosis for the prevention of pulmonary embolism; antibiotics for the treatment of community-acquired pneumonia.

Weak Recommendations

- For patients: The majority of individuals in this situation would want the suggested course of action, but many would not.
- For clinicians: Decision aids may be useful in helping individuals make decisions consistent with their values and preferences. Examine the evidence or a summary of the evidence yourself.
- For policy makers: Policy making will require substantial debates and involvement of many stakeholders.

Examples: Lung volume reduction surgery in patients with severe (upper lobe predominant) emphysema and low exercise capacity; indefinite anticoagulation in patients with idiopathic venous thromboembolism (VTE).

estimates of benefit and downsides (Table 2). This confidence is best described by the rating of methodologic quality, which we describe below.

Guideline panels should, in general, make stronger recommendations for interventions that decrease adverse outcomes with high patient importance (14) (those to which, on average, patients assign greater values and preferences) than those that decrease outcomes of lesser patient importance (Box 3).

Returning to the first example in Box 2, the initial choice made by the patient to accept adjusted-dose warfarin for 1 year versus shorter periods (< 3 mo) for the prevention of DVT recurrence or other adverse outcomes in patients with initial DVT illustrates several of the factors that will influence the strength of a recommendation (Box 4).

A patient's baseline risk of the adverse outcome (sometimes called control event risk) that treatment is expected to prevent may prove a key consideration (Table 2 and Box 5).

Another way of dealing with different baseline risks is to offer specific recommendations for several risk strata. For example, in the example above regarding COPD exacerbation, a guideline panel could offer a recommendation for patients with higher baseline risk and one for patients with lower baseline risk. Offering specific recommendations can help users of guidelines selecting the appropriate recommendations.

Data about patient preferences and values are often limited. Although it is ideal for clinicians to elicit patient preferences and values directly from patients, and for guideline panels to obtain values and preference estimates from population-based studies, such studies are often unavailable. When value or preference judgments are particularly important for the interpretation of recommendations, authors should describe the key values they have attributed in making recommendations. For example, providing a recommendation for use of inhaled corticosteroids in mild COPD would require a statement about the higher value assigned to the fewer exacerbations, the possible, but uncertain, slower rate of FEV_1 decline, and the questionable mortality reduction compared with avoiding the harms from thrush, reduced bone mineral density, increased fracture risk, the burden of using inhalers and the cost associated with therapy.

For a guideline panel to offer a strong recommendation, it has to be quite certain about the various factors that influence the strength of a recommendation and have the relevant information at hand that supports a clear balance toward either the benefits (to recommend an action) or the downsides (to recommend against an action) that influence a recommendation. In situations when a guideline panel is uncertain whether the balance is clear or when the relevant information is not available, a guideline panel should be more cautious and, in most instances, opt to make a weak recommendation. To achieve a balanced view when formulating recommendations, a multidisciplinary panel with broad representation, including clinicians, methodologists, generalists, patient representatives, and experienced guideline developers, should be assembled and proper group processes for reaching consensus on guidelines should be followed (20–22).

TABLE 2. FACTORS PANELS SHOULD CONSIDER IN DECIDING ON A STRONG OR WEAK RECOMMENDATION

Issue	Recommended Process
Quality of evidence	
1. Quality of evidence	Strong recommendations usually require higher quality evidence for all the critical outcomes. The lower the quality of evidence, the less likely is a strong recommendation.
Balance of benefits and downsides	
2. Relative importance of the outcomes (a) benefits of therapy (b) harm of treatment (c) burdens of therapy	Seek evidence about the relative values that patients place on outcomes and the actual value they place on them (critical, important but not critical, not important). Seek evidence about variability in preferences and values in patients and other stakeholders. It should be upfront that the relative importance of the outcomes should be included in the considerations before you make recommendations. If values and preferences vary widely, a strong recommendation becomes less likely.
3. Baseline risks of outcomes (a) benefits of therapy (b) harm of treatments (c) burdens of therapy	Consider the baseline risk for an outcome. Is the baseline risk going to make a difference? If yes, then consider making separate recommendations for different populations. The higher the baseline risk, the higher the magnitude of benefit and the more likely the recommendation is strong.
4. Magnitude of relative risk (a) benefits (reduction in RR) (b) harms (increase in RR) (c) burden	Consider the relative magnitude of the net effect. Large relative effects will lead to a higher likelihood of a strong recommendation if the balance of benefit, harms, and burden go in the same direction. If they go in opposite directions and the relative magnitude of effects is large (large benefits coming with large risk of adverse effects), the recommendation is more likely to be weak.
5. Absolute magnitude of the effect (a) benefits (b) harms (c) burden	Large absolute effects are more likely to lead to a strong recommendation.
6. Precision of the estimates of the effects (a) benefits of therapy (b) harms of treatments (c) burdens of therapy	The greater the precision, the more likely the recommendation is strong.
7. Costs	Consider that important benefits should come at a reasonable cost. The higher the incremental cost, all else being equal, the less likely that the recommendation in favor of an intervention is strong.

FORMULATING RECOMMENDATIONS

Guideline developers should offer clinicians as many indicators as possible for understanding and interpreting the strength of recommendations. For strong recommendations, the GRADE working group has suggested adopting terminology such as "We recommend . . ." or "Clinicians should . . ." When panels make a weak recommendation, they should use less definitive wording, such as "We suggest . . ." or "Clinicians might . . ." Furthermore, guideline panels should describe the population (described by the disease and other identifying factors) and intervention (as detailed as feasible) when they offer recommendations as specifically as possible.

THE QUALITY OF EVIDENCE

Before grading the quality of evidence, guideline developers and other groups making recommendations should conduct or identify a well-done systematic review and produce a transparent evidence summary on which to base judgments. One advance of the GRADE system is that, if justified by the available evidence, the judgments allow for strong recommendations in the setting of evidence from observational studies.

At the same time, the GRADE system exemplifies how high-quality evidence should allow for weak recommendations (Box 7).

In previous grading systems, grading primarily depended and focused on the quality of the underlying evidence, including the number of available studies. The severe infection examples and the lung cancer examples suggest that a separation of the strength

BOX 3. INTEGRATING PATIENT IMPORTANCE OF OUTCOMES IN RECOMMENDATIONS

Consider five patients with gastroesophageal reflux disease who need to be treated with a proton pump inhibitor so that one patient might achieve an uncertain benefit of cough reduction (15), in comparison to 10 patients with acute respiratory distress syndrome (ARDS) who need to be treated with a low tidal volume ventilation strategy to prevent one premature death (16). Despite the higher number needed to treat (NNT) in the patient with ARDS, since patients would value prolongation of life more highly than relieving cough, all else being equal, the latter intervention could warrant a stronger recommendation.

BOX 4. EXAMPLE: OTHER FACTORS INFLUENCING THE STRENGTH OF A RECOMMENDATION

A systematic review and meta-analysis describes a relative risk reduction (RRR) of approximately 80% in recurrent DVT for prophylaxis beyond 3 months up to 1 year. This large effect supports a strong recommendation for warfarin (17). Furthermore, the relatively narrow 95% confidence interval (CI; $\sim$ 74–88%) suggests that warfarin provides an RRR of at least 74%, and further supports a strong recommendation. At the same time, warfarin is associated with an inevitable burden of keeping dietary intake of vitamin K relatively constant, monitoring the intensity of anticoagulation with blood tests, and living with the increased risk of both minor and major bleeding. It is likely, however, that most patients would prefer avoiding another DVT and accept the risk of a bleeding episode (18). As a result, almost all patients with high risk of recurrent DVT would choose taking warfarin for 3 to 12 months, suggesting the appropriateness of a strong recommendation. Thereafter, there may be an appreciable number of patients who would reject life-long anticoagulation (*see* the example in Box 2).

BOX 5. EXAMPLE: INFLUENCE OF BASELINE RISK ON THE STRENGTH OF RECOMMENDATIONS

Consider a 65-year-old patient with mild COPD and frequent exacerbations for whom inhaled corticosteroids are a treatment option. This individual's risk for suffering an exacerbation in the next year may be 20%. Considering the RR of inhaled corticosteroids for reducing exacerbations (RR, 0.76; 95% CI, 0.72–0.80) and this baseline risk, one can derive a simplified absolute magnitude of the effect (19). Inhaled corticosteroids, relative to placebo, will reduce the absolute risk by approximately 4.8% (= 20% − [0.76 × 20%]). Some patients who are very averse to experiencing an exacerbation may consider the downsides of inhaled corticosteroids (thrush, fracture risk, burden of inhalers) well worth it. Given the relatively narrow CI that follows from the CI around the RRR, one could make a strong recommendation for using inhaled corticosteroids if all patients were equally adverse to exacerbations. More patients are, however, likely to consider the benefit not worth the harms and burden of taking inhalers if their baseline risk is lower. For instance, if the baseline risk for an exacerbation is 5%, the absolute risk reduction is only 1.2% (= 5% − [0.76 × 5%]) but the possible harms and burden remain unchanged. Fewer patients with lower baseline risk would make the choice of taking inhaled steroids. When, across the range of patient values, fully informed patients are liable to make different choices, guideline panels should offer weak recommendations and explain the rationale for their recommendation.

of a recommendation from the quality of evidence (i.e., RCTs or observational studies) is important for making recommendations.

However, the basic study design remains crucial in determining our confidence in estimates of beneficial and detrimental intervention effects. In the GRADE system, the highest quality evidence comes from one or more well-designed and well-executed RCTs yielding consistent and directly applicable results. High-quality evidence can also come, under unusual circumstances, from well-done observational studies (e.g., well-conducted and controlled cohort studies) yielding very large effects.

RCTs with important limitations and well-done observational studies yielding large effects constitute the moderate-quality category. Well-done observational studies and, on occasion, RCTs with very serious limitations will be rated as low-quality evidence. The very-low-quality category includes poorly controlled observational studies and unsystematic clinical observations (e.g., case series or case reports). This grading follows the principle that all relevant clinical studies and observations provide evidence, the quality of which varies. However, the system also clarifies that expert opinion is not a category of evidence. Expert opinion represents an interpretation of evidence, including evidence ranging from observations in an expert's own practice (uncontrolled observations) to the interpretation of RCTs and meta-analyses known to the expert in the context of other experiences and knowledge. The ATS adopted the GRADE four-category system of quality of evidence (high, moderate, low, and very low quality; Table 3) where the quality of evidence reflects our confidence that estimates of an intervention's benefits and downsides generated from research are accurate.

BOX 6. EXAMPLE: STRONG RECOMMENDATIONS IN THE FACE OF OBSERVATIONAL STUDIES

The principle of administering appropriate antibiotics rapidly in the setting of severe infection or sepsis has not been tested against its alternative of no rush of delivering antibiotics in RCTs (it is unlikely that these trials will ever be performed and recommendations need to be made) (23). Yet, guideline panels that apply the GRADE approach would be very likely to make a strong recommendation for the rapid use of antibiotics in this setting on the basis of available observational studies because the benefits of antibiotic therapy clearly outweigh the downsides in most patients independent of the quality assessment.

BOX 7. EXAMPLE: WEAK RECOMMENDATIONS BASED ON HIGH-QUALITY EVIDENCE

Several RCTs compared the use of combination chemotherapy and radiotherapy versus radiotherapy alone in unresectable, locally advanced non–small cell lung cancer (stage IIIA) (24, 25). The overall quality rating for these trials could be considered high by a guideline panel. Compared with radiotherapy alone, the combination of chemotherapy and radiotherapy reduces the risk for death corresponding to a mean gain in life expectancy of few months (24), but increases harm and burden related to chemotherapy. Thus, considering the values and preferences patients would place on the small survival benefit in view of the harms and burdens, guideline panels may offer a weak recommendation (Table 1) despite the high quality of the available evidence.

Factors that Decrease the Quality of Evidence

The following limitations may decrease the quality of evidence supporting a recommendation (Table 4).

1. Limitation of methodology. Our confidence in recommendations decreases if studies suffer from major limitations that are likely to result in a biased assessment of the treatment effect. These methodologic limitations include failure to adhere to an intention-to-treat analysis (in the context of RCTs), lack of blinding with subjective outcomes highly susceptible to bias, or a large loss to follow-up.
2. Unexplained heterogeneity of results. When studies yield widely differing estimates of the treatment effect (heterogeneity or variability in results), investigators should look for explanations for that heterogeneity. For instance, drugs may have larger relative effects in sicker populations or when given in larger doses. When heterogeneity exists,

TABLE 3. DETERMINANTS OF THE QUALITY OF EVIDENCE (CONFIDENCE IN THE ESTIMATES OF BENEFITS, HARMS, BURDEN, COSTS): UNDERLYING METHODOLOGY AND QUALITY RATING

Underlying Methodology	Quality Rating
RCT	High
Downgraded RCTs or upgraded observational studies	Moderate
Well-done observational studies with control groups	Low
Others (e.g., case reports or case series)	Very low

Definition of abbreviation: RCT = randomized controlled trial.

TABLE 4. DETERMINANTS OF THE QUALITY OF EVIDENCE (CONFIDENCE IN THE ESTIMATES OF BENEFITS, HARMS, BURDEN, COSTS): FACTORS THAT MAY DECREASE OR INCREASE THE QUALITY OF EVIDENCE

Factors that may decreases the quality of evidence:
- Limitations in the design and implementation of available RCTs, suggesting high likelihood of bias (−1 or −2 categories of quality)
- Inconsistency of results (including problems with subgroup analyses) (−1)
- Indirectness of evidence (indirect population, intervention, control, outcomes, −1 or −2)
- Sparse evidence (−1)
- High probability of reporting bias (−1)

Factors that may increase the quality of evidence:
- Large magnitude of effect (direct evidence, RR > 2 or RR < 0.5 with no plausible confounders (+1); very large, with RR > 5 or RR < 0.2 and no threats to validity (+2)
- All plausible confounding would reduce a demonstrated effect (+1)
- Dose–response gradient (+1)

Numbers in parentheses indicate the levels of change in the quality of evidence.
Definition of abbreviation: RCT = randomized controlled trial; RR = relative risk.

but investigators fail to identify a plausible explanation, the quality of evidence decreases.

3. Indirectness of evidence (i.e., the question being addressed in the guideline is quite different from the available evidence regarding the population, intervention, comparison, or outcome). Investigators may have undertaken studies in similar, but not identical, populations to those under consideration for a recommendation. Guideline panels should consider this as indirect evidence and, to the extent they are uncertain about the applicability to their relevant population, downgrade the quality rating. For instance, although a trial of intensive insulin therapy in primarily postsurgical patients demonstrated an impressive survival benefit (26), uncertainty regarding direct applicability of this evidence to critically ill medical patients led to further RCTs (27).

   Indirectness may also apply to the intervention (e.g., RCTs of related but not identical interventions, such as different doses and regimens of inhaled corticosteroids) and outcomes (e.g., measuring FEV_1 when exacerbations or mortality may be most important, a surrogate outcome).

4. Lack of precision. When studies include very few patients and very few events, a guideline panel may judge the quality of the evidence lower than it otherwise would. The following statements may be offered to guide judgments when imprecise results warrant quality downgrading: Data are sparse if the results include so few events, that they are uninformative. Data are imprecise if the confidence intervals are sufficiently wide that an estimate is consistent with either important net benefits or harms and thus consistent with divergent recommendations.

The factors influencing the quality of evidence may be additive such that the presence of several of these factors, if judged important, would lower the quality of evidence by more than one category. Each of these factors (methodologic limitations, indirectness, heterogeneity, and imprecision) may also decrease the quality of evidence associated with observational studies (moving the categorization of such evidence from low to very low quality).

Factors that Increase the Quality of Evidence

Although well-done observational studies will generally yield low-quality evidence, there may be unusual circumstances in which guideline panels classify such evidence as moderate or even high quality (Box 6).

1. On rare occasions, when controlled, methodologically strong observational studies yield large or very large and consistent estimates of the magnitude of a treatment effect, we may be confident about the results. In those situations, although the observational studies are likely to have provided an overestimate of the true effect, the weak study design may not explain all of the apparent benefit. Thus, despite reservations based on the observational study design, we are confident that the effect exists. Table 4 shows how the magnitude of the effect in these trials may move the assigned quality of evidence from low to moderate (if the effect is large in the absence of other methodologic limitations), or make the quality rating high (if the effect is very large in the absence of other methodologic limitations).
2. On other occasions, all plausible biases from observational studies may be working to underestimate an apparent treatment effect. For example, if only sicker patients receive an experimental intervention, yet they still fare better, it is likely that the actual treatment effect is larger than the data suggest.
3. The presence of a dose–response gradient may also increase our confidence in the findings of observational studies and thereby enhance the assigned quality of evidence.

BOX 8. EXAMPLE: RATING THE OVERALL QUALITY OF EVIDENCE WHEN THE QUALITY DIFFERS ACROSS OUTCOMES

Consider, for instance, administration of selective digestive decontamination (SDD) in intensive care unit patients. Several meta-analyses of high-quality RCTs suggested a decrease in the incidence of infections and, likely, the mortality of ventilated patients with SDD (23). The quality of evidence on the effect of SDD on the emergence of bacterial antibiotic resistance and its clinical relevance is much less clear. One might reasonably rate the evidence about this feared potential adverse effect as low quality. Should the overall quality of evidence for use of SDD therefore be considered high, moderate, or low? In such instances, we suggest that authors should consider whether downsides of therapy are critical to the decision regarding the optimal management strategy. If they are, one must rate the overall quality of the evidence according to the studies that address adverse effects. If not, the judgment on the overall rating of the evidence is based on the evidence regarding benefit. Thus, guideline panels have to decide *a priori* which outcomes are critical for the decision-making process.

WHAT TO DO WHEN QUALITY OF EVIDENCE DIFFERS ACROSS OUTCOMES

Guideline panels usually provide a single rating of quality of evidence for every recommendation. Recommendations, however, depend on evidence regarding a number of outcomes. Thus, it may be necessary to report a single evidence grade when the quality of evidence differs across important outcomes. Guideline panels should determine the quality of evidence for each outcome, but in terms of overall quality of evidence, the lowest quality of data available for any one of the critical outcomes determines the overall quality of evidence (Box 7).

Guideline panels may refer to the checklist provided in Table 5 while developing and grading recommendations. The example (Box 8) from the management of impending sepsis shows how panelists might work through the issues.

ADDITIONAL CONSIDERATIONS

The ATS has produced numerous guidelines, many of them in collaboration with other guideline developers or organizations. Although widely recognized, the guidelines have been variable in the extent to which they have adhered to methodologic standards (28), and they have applied a variety of approaches to grading the quality of evidence. For example, some collaborative efforts involve grading systems that rate the quality of evidence but do not provide a grade for the strength of a recommendation.

The following example provides an additional reason for a new, sensible grading system for the ATS. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines state (29): "The most common causes of an exacerbation are infection of the tracheobronchial tree and air pollution, but the cause of about one-third of severe exacerbations cannot be identified (Evidence B)." Grading the evidence for etiologic questions in guidelines presents challenges because clinicians typically do not make recommendations about prognostic or etiologic factors (information about etiology) and the evidence does not come from randomized comparisons of one risk factor versus another. As a result, randomized designs do not provide higher quality evidence or information about etiologic factors than observational studies and generic grading systems therefore do not apply to such statements. Thus, grading of etiologic information is irrelevant for guidelines and recommendations because action follows from knowing that modifying etiologic factors influences outcomes. Grading in guidelines therefore should be restricted to recommended actions. Because of the need to clarify methodologic issues around grading the quality of evidence and recommendations and to unify and improve the existing grading methodology applied by ATS guideline developers, we proposed the use of the GRADE approach. The framework summarized in Table 6 generates recommendations ranging from a strong recommendation based on high-quality evidence to weak recommendations based on very-low-quality evidence.

TABLE 5. A CHECKLIST FOR DEVELOPING AND GRADING RECOMMENDATIONS

Define the population, intervention and alternative, and the relevant outcomes
Summarize the relevant evidence (relying on systematic reviews)
If randomized trials available, start by assuming high quality; if well-done observational studies are available, assume low quality, but then check for the following: • Serious methodologic limitations (lack of blinding, concealment, high loss to follow-up, stopped early) • Indirectness in population, intervention, or outcome (use of surrogates) • Inconsistency in results • Imprecision in estimates
Grade RCTs down from high to moderate, low, or very low depending on limitations, or observational studies to very low.
If no randomized trials are available but well-done observational studies are available (including indirectly relevant trials and well-done observational study), start by assuming low quality, but then check for the following: • Large or very large treatment effect • All plausible confounders would diminish effect of intervention • Dose–response gradient
Grade up to moderate or even high depending on special strengths or weaknesses
Studies starting at very low will not be upgraded. Observational studies with limitations will not be upgraded. Only observational studies with no threats to validity can be upgraded.
Decide on best estimates of benefits, harms, burden, and costs for relevant population
Decide on whether the benefits are, overall, worth the harms, burden, and costs for relevant population and decide how clear and precise this balance is

For definition of abbreviation, *see* Table 3.

BOX 9. A CHECKLIST AND AN EXAMPLE FOR THE PROCESS OF MAKING RECOMMENDATIONS

Question:
Should patients with pancreatic necrosis in acute pancreatitis receive antibiotic prophylaxis?
Patients:
Patients with pancreatic necrosis in the course of acute pancreatitis
Intervention:
Systemic prophylactic antibiotic
Outcomes:
All-cause mortality, pancreatic sepsis, fungal superinfection
Evidence Summary:
Systematic review of four randomized trials. Among 218 included patients, analysis showed statistically significant reduction of all-cause mortality (RRR, 66%; 95% CI, 16–85%; NNT, 10; 95% CI, 6–34) and pancreatic sepsis (RRR, 36%; 95% CI, 1–58%; NNT, 9; 95% CI, 5–100). No increase in fungal infections. No data on the incidence of resistant organism selection.
Quality of Evidence:
Randomized trials without serious limitations provide direct and consistent evidence pointing towards large effect size. At the same time, all studies were unblinded, and the total number of patients was relatively few. In balance, the evidence may be considered moderate rather than high. If the outcome resistance pattern is considered critical for making a decision, it may even be considered low.
Best Estimates:
Reduction of mortality and pancreatic sepsis.
Judgment of Benefits versus Risks, Burden, and Cost:
Information available suggests benefits of prophylaxis, but the balance is not clear.
Grade of Recommendation:
Quality of evidence only moderate for outcomes available and minimal evidence for some other important outcomes leaves uncertainty. The recommendation could be expressed as "For patients with pancreatic necrosis in the course of acute pancreatitis, we suggest systemic prophylactic antibiotic (weak recommendation based on moderate-quality evidence)."

Strengths and Limitations

One of the major merits of GRADE is the simplicity of its two-category system of grading recommendations. The behavioral implications of strong and weak recommendations provide practical guidance to clinicians and other users (Table 1). The definition of categories of methodologic problems and merits allows an explicitness and transparency that other systems lack.

The ATS makes no official recommendation to others for using the GRADE approach, but guideline panels considering using GRADE can anticipate support by the GRADE working group that is not available for other systems. Independent of ATS efforts, the large group of methodologists involved in GRADE

TABLE 6. GRADING RECOMMENDATIONS

Grade of Recommendation	Clarity of Risk/Benefit	Quality of Supporting Evidence	Implications
Strong recommendation High-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Consistent evidence from well-performed randomized controlled trials or exceptionally strong evidence from unbiased observational studies	Recommendation can apply to most patients in most circumstances. Further research is very unlikely to change our confidence in the estimate of effect.
Strong recommendation Moderate-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Evidence from randomized controlled trials with important limitations (inconsistent results, methodologic flaws, indirect or imprecise), or unusually strong evidence from unbiased observational studies	Recommendation can apply to most patients in most circumstances. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Strong recommendation Low-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Evidence for at least one critical outcome from observational studies, from randomized controlled trials with serious flaws or indirect evidence	Recommendation may change when higher quality evidence becomes available. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Strong recommendation Very-low-quality evidence (very rarely applicable)	Benefits clearly outweigh harms and burdens , or vice versa	Evidence for at least one of the critical outcomes from unsystematic clinical observations or very indirect evidence	Recommendation may change when higher quality evidence becomes available; any estimate of effect, for at least one critical outcome, is very uncertain.
Weak recommendation High-quality evidence	Benefits closely balanced with harms and burdens	Consistent evidence from well-performed randomized controlled trials or exceptionally strong evidence from unbiased observational studies	The best action may differ depending on circumstances or patients or societal values. Further research is very unlikely to change our confidence in the estimate of effect.
Weak recommendation Moderate-quality evidence	Benefits closely balanced with harms and burdens	Evidence from randomized, controlled trials with important limitations (inconsistent results, methodologic flaws, indirect or imprecise), or unusually strong evidence from unbiased observational studies	Alternative approaches likely to be better for some patients under some circumstances. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Weak recommendation Low-quality evidence	Uncertainty in the estimates of benefits, harms, and burdens; benefits may be closely balanced with harms and burdens	Evidence for at least one critical outcome from observational studies, from randomized controlled trials with serious flaws, or indirect evidence	Other alternatives may be equally reasonable. 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.
Weak recommendation. Very low quality of evidence	Major uncertainty in the estimates of benefits, harms, and burdens; benefits may or may not be balanced with harms and burdens	Evidence for at least one critical outcome from unsystematic clinical observations or very indirect evidence	Other alternatives may be equally reasonable. Any estimate of effect, for at least one critical outcome, is very uncertain.

conduct regular workshops around the world and have acted as resources for any group considering to use GRADE. The approach offers the possibility of working electronically and making guideline material available on the World Wide Web (www.gradeworkinggroup.org). Evidence tables and recommendations could form the sole publication in print, whereas information required for decision making by guideline panels and for those clinicians who require an in-depth understanding of all the evidence could be deposited in electronic format and connected via hyperlinks. Additional advantages include that the GRADE system applies to diagnostic recommendations similarly to how it applies to questions about therapy. The final recommendation from a diagnostic question depends on the balance between benefits and downsides of the diagnostic strategy in terms of patient important outcomes, although, until recently, these outcomes have been measured infrequently. Finally, the novel approach to grading the quality of evidence for each important outcome and applying the quality of all critical outcomes to the final quality grade provides increased transparency about evidence supporting recommendations that help in responding to health care questions. The health care questions are often complex and associated with finely balanced benefits and downsides.

Adopting the GRADE approach also has some disadvantages that are inherent to any grading system. Systems currently used to analyze scientific data for the purpose of creating CPGs have not been tested rigorously for validity and reproducibility. GRADE is no exception. Establishing criteria for the validity of rating of quality of evidence is extremely challenging. Establishing criteria for the validity of the direction and strength of recommendation is even more problematic because it depends on underlying values and preferences that would have to be precisely specified. The "impact" of a CPG—the degree to which it affects behaviors—does not qualify as a measure of outcome because impact does not necessarily reflect a guideline's internal validity (i.e., the extent to which the process of development has produced an approximation of "scientific truth"). Even if the process yields truth, it may not necessarily convince sufficiently to alter clinicians' behaviors. Many factors influence many behaviors, including the respect for the methods used to create a guideline, or reputations of panelists enlisted to draft it, the societies supporting it, or of the journal that publishes a guideline. Thus, the impact of a guideline on behaviors is a poor measure of the validity of the processes used to create it. These challenges result in a situation in which none of the competing systems have been validated; thus, validity was not a criterion that the ATS document and implementation committee applied when making its choices.

Consumers of grading systems often raise concerns about the reproducibility of the grading process. When the GRADE group assessed an early version of its grading system across medical specialty areas, there was varying agreement about the quality

of evidence for the rated outcomes (κ coefficients for agreement beyond chance ranged from 0 to 0.82) (30). However, there was fair agreement about the relative importance of each important or critical clinical outcome. There was poorer agreement about the balance of benefits and downsides in recommendations. Given the inevitably varying values and preferences of the raters, and the difficulty in precisely specifying underlying values and preferences (including risk aversion), one might anticipate such variability.

Lack of reproducibility is of less concern if judgments are made transparent and consumers can track reasons for decisions about the grading by guideline panels. If guideline developers applied consistent approaches to evaluating quality of evidence and grading recommendations, differences in judgments could be more easily understood. One merit of the GRADE system is the transparency of the judgments, which is strongest for rating the quality of the evidence. Table 5 presents a summary of the sequential judgments a guideline panel would make following the GRADE approach. Readers of a graded guideline or recommendation should be aware that judgments about the quality of evidence, including those following the GRADE approach, require experience and expertise by guideline panels about the addressed health care question and research methodology. As described above, however, expert opinion does not constitute a form of evidence but an interpretation of existing evidence.

Other disadvantages of adopting the GRADE approach include the requirement for resources to conduct detailed assessment of the evidence and the requirement of consumers to develop some basic understanding of the system. The latter is of concern for any grading system, and GRADE's choice of a simple two-category approach to the strength of the recommendation facilitates ease of understanding.

Some users of recommendations may find a two-category rating of the strength of recommendation too simplistic for the problems clinicians encounter in daily practice. However, there are several issues to consider that speak for the simpler choice of a two-category grading of the strength of recommendations over more categories. First, there are several ways (Table 1) consumers of guidelines can interpret these recommendations. Second, when balancing the continuum of benefits and downsides—often a very challenging process—guideline panels can choose between two categories of strength against an action and two categories for an action. Third, the GRADE system explicitly asks for a detailed and transparent description of the underlying judgments and values that influence a recommendation. Thus, consumers of guidelines have the option of making different choices (predominantly in the case of weak recommendations) when they have information that leads them to disagree with the judgments and have evidence that the values of their patients differ. The ATS Documents and Implementation Committee recognizes the limitations of GRADE, particularly with respect to validity and reproducibility. There are, however, no competing systems that are superior in this regard, and GRADE has many strengths. Because we see compelling arguments for adopting a single, uniform approach to grading recommendations that is consistent or nearly consistent with systems adopted by other leading organizations (9), the ATS Documents Committee has chosen GRADE as the preferred current methodology for rating the quality of evidence and strength of recommendations. The ATS adopted the original GRADE four-category grading system for the quality of evidence. The latter represents an important distinction to the GRADE approach adapted by the ACCP that combines the low and very low quality of evidence (9). The ACCP refrained from using the very-low-quality category in part because, for many of the therapeutic areas that ACCP guidelines focus on, such as antithrombotic guidelines, higher quality primary evidence exists (31).

CONCLUSIONS

In the grading system the Documents Development and Implementation Committee adopted for the ATS, the strength of any recommendation depends on two factors: the quality of the evidence regarding treatment effect and the tradeoff between benefits and downsides of an intervention. The system classifies methodologic quality in four categories: randomized trials that show consistent results, or observational studies with very large treatment effects (high quality); randomized trials with methodologic limitations, or observational studies with large effect (moderate quality); and observational studies without exceptional strengths, or randomized trials with very serious limitations (low quality). We classify unsystematic clinical observations (e.g., case reports and case series) as evidence of very-low-quality evidence (very low quality). The balance between benefits and downsides falls into one of two categories. Recommendations are either strong, defined as being "confident that adherence to the recommendation will do more good than harm or that the net benefits are worth the costs," or weak, defined as being "uncertain that adherence to the recommendation will do more good than harm OR that the net benefits are worth the costs." Panels can make recommendations for or against a given intervention. The language of strong recommendations (worded as "we recommend" or "should" in the actual recommendation) reflects the following clinical message: the recommendation applies to most patients under most circumstances. The language of weak recommendations (worded as "we suggest" or "might") reflects a different clinical message: the need to consider more carefully than usual individual patients' circumstances, preferences, and values. The uncertainty associated with weak recommendations follows either from poor-quality evidence (if we are uncertain of benefits and downsides, it is not wise to make a strong recommendation for or against), or from closely balanced benefits versus downsides.

This statement was prepared by the ATS Documents Development and Implementation Committee.

Members of the committee are as follows:

Holger J. Schünemann, M.D., Ph.D. (*chair*), Rome, Italy
Roman Jaeschke, M.D., M.Sc., Hamilton, Canada
Deborah J. Cook, M.D., M.Sc., Hamilton, Canada
William F. Bria, M.D., Ann Arbor, Michigan
Ali A. El-Solh, M.D., M.P.H., Buffalo, New York
Armin Ernst, M.D., Boston, Massachusetts
Bonnie F. Fahy, R.N., M.S.N., Phoenix, Arizona
Richard L. Gelula, M.S.W., Washington, D.C.
Michael K. Gould, M.D., M.S., Stanford, California
Kathleen L. Horan, M.D., Stanford, California,
Jerry A. Krishnan, M.D., Ph.D., Baltimore, Maryland
Constantine A. Manthous, M.D., Providence, Rhode Island
Janet R. Maurer, M.D., Anthem, Arizona
Walter T. McNicholas, M.D., Dublin, Ireland
Andrew D. Oxman, M.D., M.Sc., Oslo, Norway
Gordon Rubenfeld, M.D., Seattle, Washington
Gerard M. Turino, M.D. (*vice-chair*), New York, New York
Gordon Guyatt, M.D., M.Sc., Hamilton, Canada
Jeffrey S. Wagener, M.D., Denver, Colorado

Conflict of Interest Statement: H.J.S., R.J. and G.G. are members of the GRADE working group that developed the GRADE grading system. The GRADE working group is an informal group of methodologists and guideline developers with interest in improving guideline methodology. H.J.S., R.J., and G.G. participate in developing a freely available software (GRADEpro) for applying the GRADE approach. They have no direct financial interest in the GRADE approach or the GRADEpro software. D.J.C., W.F.B., A.A. E.-S., A.E., B.F.F., M.K.G., K.L.H., J.A.K.,

C.A.M., J.R.M., W.T.M., A.D.O., G.R., and G.M.T. do not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript.

References

1. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *ACP J Club* 2002;136: A11.
2. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, *et al.* Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
3. Guyatt G, Cook D, Jaeschke R, Schünemann H, Pauker S. Grading recommendations: a qualitative approach. In: Guyatt G, Rennie D, editors. Users' guides to the medical literature: a manual for evidence-based practice. Chicago, IL: AMA Press; 2002. p. 599–608.
4. Schünemann H, Heffner JE. A new ATS committee: competing in the marketplace of ideas. *Proc Am Thorac Soc* 2005;2:249–250.
5. Schünemann H, Heffner JE. A new ATS committee: competing in the marketplace of ideas. *Am J Respir Crit Care Med* 2005;172:1067–1068.
6. Schünemann H, Heffner JE. A new ATS committee: competing in the marketplace of ideas. *Am J Respir Cell Mol Biol* 2005;33:423–424.
7. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B, *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.
8. Schunemann 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, Gutterman D, Baumann M, Addrizzo-Harris D, Hylek E, Phillips B, Raskob G, Lewis S, Schünemann H. Grading strength of recommendations and quality of evidence in clinical guidelines. *Chest* 2006;129:174–181.
10. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations [editorial]? *ACP J Club* 2006;144:A08.
11. Büller HR, Agnelli G, Hull RD, Hyers TM, Prins MH, Raskob GE. Antithrombotic therapy for venous thromboembolic disease. *Chest* 2004;126:401S–428S.
12. Fishman A, Martinez F, Naunheim K, Piantadosi S, Wise R, Ries A, Weinmann G, Wood DE. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003;348:2059–2073.
13. O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, Tait V, Tetroe J, Fiset V, Barry M, *et al*. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2003;2:CD001431.
14. Guyatt G, Moutori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the centre: in our practice, and in our use of language. *ACP Journal Club* 2004;140:A11.
15. Chang AB, Lasserson TJ, Kiljander TO, Connor FL, Gaffney JT, Garske LA. Systematic review and meta-analysis of randomised controlled trials of gastro-oesophageal reflux interventions for chronic cough associated with gastro-oesophageal reflux. *BMJ* 2006;332:11–17.
16. Petrucci N, Iacovelli W. Ventilation with lower tidal volumes versus traditional tidal volumes in adults for acute lung injury and acute respiratory distress syndrome. *Cochrane Database Syst Rev* 2004;2: CD003844.
17. Hutten B, Prins MH. Duration of treatment with vitamin K antagonists in symptomatic venous thromboembolism. *Cochrane Database Syst Rev* 2000;3:CD 001367.
18. Buller HR, Agnelli G, Hull RD, Hyers TM, Prins MH, Raskob GE. Antithrombotic therapy for venous thromboembolic disease: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004;126:401S–428S.
19. Sin DD, McAlister FA, Man SF, Anthonisen NR. Contemporary management of chronic obstructive pulmonary disease: scientific review. *JAMA* 2003;290:2301–2312.
20. Fretheim A, Schünemann HJ, Oxman AD. Improving the use of research evidence in guideline development: 3. Group composition. *Health Res Policy Syst* (In press)
21. Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. *BMJ* 1999;318:593–596.
22. Fretheim A, Schünemann HJ, Oxman AD. Improving the use of research evidence in guideline development: 5. Group process. *Health Res Policy Syst* (In press)
23. Garanch-Montera J, Garcia-Garmendia JL, Barrero-Almovodar A, Jimenez-Jimenez FJ, Perez-Paredes C, Ortiz-Leyba C. Impact of adequate antibiotics therapy on the outcome of patients admitted to the intensive care unit with sepsis. *Crit Care Med* 2003;12:2742–2751.
24. Pritchard RS, Anthony SP. Chemotherapy plus radiotherapy compared with radiotherapy alone in the treatment of locally advanced, unresectable, non-small-cell lung cancer: a meta-analysis. *Ann Intern Med* 1996;125:723–729.
25. Robinson LA, Wagner H Jr, Ruckdeschel JC. Treatment of stage IIIA non-small cell lung cancer. *Chest* 2003;123:202S–220S.
26. Van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Bouillon R. Intensive insulin therapy in critically ill patients. *N Engl J Med* 2001;345:1359–1367.
27. Meijering S, Corstjens AM, Tulleken JE, Meertens JH, Zijlstra JG, Ligtenberg JJ. Towards a feasible algorithm for tight glycaemic control in critically ill patients: a systematic review of the literature. *Crit Care* 2006;10:R19.
28. American Thoracic Society. Attributes of ATS documents that guide clinical practice: recommendations of the ATS Clinical Practice Committee. *Am J Respir Crit Care Med* 1997;156:2015–2025.
29. Pauwels, R., Buist AS, Calverley PMA, Jenkins CR, Hurd SS; GOLD Scientific Committee. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease NHLBI/ WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) workshop summary. *Am J Respir Crit Care Med* 2001;163: 1256–1276.
30. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, Hill S, Jaeschke R, Liberati A, Magrini N, *et al.* Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. *BMC Health Serv Res* 2005;5:25.
31. Hirsh J, Guyatt G, Albers GW, Schünemann HJ. The Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy: evidence-based guidelines. *Chest* 2004;126:172S–173S.