Interpreting Levels of Evidence and Grades of Health Care Recommendations

Alba DiCenso and Gordon Guyatt

The following Editorial Board members also made substantive contributions to this chapter: Lazelle Benefield and Mariko Koyama.

We gratefully acknowledge the work of Robert Hayward, Scott Richardson, Lee Green, Mark Wilson, Jack Sinclair, Deborah Cook, Paul Glasziou, Alan Detsky, and Eric Bass on the original chapter that appears in the Users’ Guides to the Medical Literature, edited by Guyatt and Rennie.

In This Chapter

Health Care Recommendations

Developing Health Care Recommendations
  Health Care Recommendations and Clinical Practice Guidelines

Assessing Health Care Recommendations
  Do the Recommendations Consider All Relevant Patient Groups, Management Options, and Possible Outcomes?
  Is There a Systematic Review of Evidence Linking Options to Outcomes for Each Relevant Question?
  Is There an Appropriate Specification of Values or Preferences Associated With Outcomes?
  Do the Authors Indicate the Strength of Their Recommendations?

Relationship Between Levels of Evidence and Grades of Health Care Recommendations

Problems with Existing Systems for Grading Evidence and Health Care Recommendations

A New Grading System
  Levels of Evidence
  Grading of Recommendations

Evaluating the Trade-off Between Benefits and Risks
HEALTH CARE RECOMMENDATIONS

Each day, clinicians make dozens of patient management decisions. Some are relatively inconsequential, whereas others are important. Each one involves weighing benefits and risks, gains and losses, and recommending or instituting a course of action judged to be in the patient’s best interest. Implicit in each decision is a consideration of the relevant evidence, an intuitive integration of that evidence, and a weighing of the likely benefits and risks in light of the patient’s preferences. When making choices, nurses may benefit from structured summaries of options and outcomes, systematic reviews of the evidence on the relationship between options and outcomes, and recommendations about the best choices. This chapter outlines the process for developing recommendations, summarizes criteria for critically evaluating the methodological quality of recommendations, and introduces a taxonomy for grading evidence and recommendations that facilitates consideration of the trade-off between benefits and risks.

Traditionally, authors of original research on health care interventions include recommendations about the use of these interventions in clinical practice in the discussion sections of their papers. Authors of systematic reviews also tend to provide their impressions of the management implications derived from the summarization of evidence. Typically, however, authors of individual trials or systematic reviews do not consider all possible management options, but instead focus on comparisons of two or three alternatives. They may also fail to identify subpopulations in which the impact of the intervention may vary considerably. In addition, they may not consider some of the important outcomes associated with alternative management options. Finally, when the authors of systematic reviews provide recommendations, they typically are not grounded in an explicit presentation of societal or patient preferences.

Failure to consider these issues may lead to variability in recommendations based on the same data. For example, various recommendations emerged from different systematic reviews of adolescent pregnancy prevention interventions despite similar results.\(^1\)\(^2\) The recommendations varied from suggesting to rejecting implementation of interventions such as sex education. Varying recommendations reflect the fact that investigators reporting primary studies and systematic reviews often make their recommendations without benefit of an explicit standardized process or set of rules.

When benefits or risks are dramatic and essentially homogeneous across an entire population, intuition may provide an adequate guide to making treatment recommendations. However, such situations are unusual. In most instances, intuitive recommendations risk misleading clinicians and patients because of their susceptibility to both bias and random error. These considerations suggest that nurses should critically evaluate the methodological quality of health care recommendations before implementing them.

DEVELOPING HEALTH CARE RECOMMENDATIONS

Figure 35-1 presents the steps involved in developing a recommendation, along with the formal strategies for doing so. The first step in clinical decision making is to define the decision. This involves specifying the alternative courses of action and the possible outcomes. Interventions are often designed to delay or prevent adverse outcomes, such as
falls, pressure ulcers, urinary incontinence, or death. As usual, we refer to the outcomes that an intervention is designed to prevent as target outcomes. Interventions are associated with their own adverse outcomes: pain, worry, inconvenience, and side effects. In addition, new interventions may markedly increase—or decrease—costs. Ideally, the definition of the decision will be comprehensive: developers of health care recommendations should consider all reasonable alternatives and identify all possible beneficial and adverse outcomes. For example, in considering recommendations for prevention of falls in the elderly, options include doing nothing, performing individualized assessments, or offering an exercise program to everyone or only those who are at particularly high risk. Major outcomes include falls and fractures, the inconvenience associated with participating in the assessment or exercise program, and costs to participants, the health care system, and society.

Having identified the options and outcomes, developers of recommendations must evaluate the links between the two. What will the alternative management strategies yield in terms of benefits and harms? How are potential benefits and risks likely to vary in different groups of patients? Once these questions are answered, making recommendations about interventions involves value judgments about the relative desirability or undesirability of possible outcomes. We use the term preferences synonymously with values or value judgments in referring to the process of trading off positive and negative consequences of alternative management strategies.

---

**Figure 35-1.** A schematic view of the process of developing a health care recommendation.
Health Care Recommendations and Clinical Practice Guidelines

Health care recommendations are found in clinical practice guidelines. Practice guidelines, or “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances,” represent an attempt to distill a large body of health care knowledge into a convenient, readily usable format. Similar to systematic reviews, they gather, appraise, and combine evidence. Guidelines, however, go beyond systematic reviews in attempting to address all of the issues relevant to a clinical decision and all of the values that might sway a clinical recommendation. Guidelines balance trade-offs between benefits and risks and make explicit recommendations, often on behalf of health organizations, with a definite intent to influence what clinicians do. Practice guideline methodology relies on the consensus of a group of decision makers, ideally including (or reflecting the views of) content experts, front-line clinicians, and patients, who carefully consider the evidence and decide on its implications. The mandate of guideline developers may focus on recommendations for a country, region, city, hospital, or clinic. Guidelines based on the same evidence may differ depending on the country (e.g., China or the United States), regional characteristics (urban or rural), size of the institution (e.g., a large teaching hospital or a small community hospital), and the population served (e.g., a poor community or an affluent one). For a detailed description of practice guidelines, see Chapter 10, Moving from Evidence to Action Using Clinical Practice Guidelines.

ASSESSING HEALTH CARE RECOMMENDATIONS

In Table 35-1, we offer four guidelines to assess the validity of a health care recommendation.

Do the Recommendations Consider All Relevant Patient Groups, Management Options, and Possible Outcomes?

Recommendations pertain to decisions, and decisions involve particular groups of patients, choices for those patients, and the consequences of the choices. Regardless of whether recommendations apply to diagnosis, prevention, treatment, or rehabilitation, they should specify all relevant patient groups, the interventions of interest, and sensible alternative options. Health care recommendations often vary for different subgroups of patients. In particular, patients at lower risk of the target outcomes that an intervention is designed to prevent are less likely to benefit from the intervention than are those at higher risk (see Chapter 33, Applying Results to Individual Patients). For instance, in a guideline on colorectal cancer screening for asymptomatic people, the Canadian Task Force on Preventive Health Care provided separate recommendations for people at normal risk and those at above-average risk because of a family history of colorectal cancer.

Guideline developers who formulate recommendations must consider not only all relevant patient groups and management options, but all important consequences of the options as well. Evidence about the effects on morbidity, mortality, and quality of life are all relevant to patients, and efficient use of resources dictates attention to costs.
Costs, regardless of whether they are based on the perspective of patients, insurers, or the health care system—or whether they consider broader issues, such as the consequences of time lost from work—can affect the recommendations (see Chapter 18, Economic Evaluation).

Making recommendations about screening requires particular attention to identifying all potential outcomes. Attempting to identify disease in asymptomatic individuals may result in several negative outcomes that clinicians do not face when diagnosing and treating symptomatic patients. Asymptomatic individuals who screen positive for a disease must live for a longer time with awareness of their illness and the associated negative psychologic consequences. This is particularly problematic if patients remain asymptomatic for long periods. For example, consider a man who screened positive for prostate cancer, but who was destined to die of heart disease before the prostate cancer became clinically manifest. Those who screen positive but ultimately are found not to have the disease, such as with false positives in breast cancer screening, may find the experience traumatic, whereas people who screen negative but ultimately are found to have the target condition may feel betrayed (see Chapter 36, Recommendations About Screening).

Is There a Systematic Review of Evidence Linking Options to Outcomes for Each Relevant Question?

Having specified options and outcomes, the next task for guideline developers is to estimate the likelihood that each outcome will occur. In effect, they have a series of specific questions. For hormone replacement therapy, the initial question may be, “What is the effect of alternative approaches on the incidence of hip fracture, breast cancer, endometrial cancer, myocardial infarction, and sudden coronary death?”

Recommendations must consolidate and combine all of the relevant evidence in an appropriate manner. In formulating recommendations, guideline developers must avoid bias that will distort the results. This requires access to, or conduct of, a systematic review of the evidence bearing on each question. Chapter 9, Summarizing the Evidence Through Systematic Reviews, provides guidelines for assessing the likelihood that collection and summarization of the evidence are free from bias.

The best recommendations define admissible evidence, report how it was selected and combined, make key data available for review, and report randomized trials that

### Table 35-1  Users’ Guides for the Validity of Health Care Recommendations

- Do the recommendations consider all relevant patient groups, management options, and possible outcomes?
- Is there a systematic review of evidence linking options to outcomes for each relevant question?
- Is there an appropriate specification of values or preferences associated with outcomes?
- Do the authors indicate the strength of their recommendations?
link interventions with outcomes. However, such randomized trials may be unavailable, and the authors of systematic reviews may reasonably abandon their project if there are no high-quality studies to summarize. Those making recommendations, however, must do so even in the absence of high-quality studies. For important but ethically, technically, or economically difficult questions, strong scientific evidence may never become available. Because recommendations must deal with the best (often inadequate) evidence available, guideline developers may need to consider a variety of studies (published and unpublished) and reports of expert and consumer experience. This means that the strength of the evidence in support of recommendations can vary widely, and poor-quality evidence may yield weak recommendations.

Is There an Appropriate Specification of Values or Preferences Associated With Outcomes?

Linking treatment options with outcomes is largely based on fact and science. Assigning preferences to outcomes, by contrast, is a matter of values. Consider, for example, a guideline panel examining just two of the outcomes associated with using, or not using, hormone replacement therapy (HRT). First, the panel must consider the risk of breast cancer in women using, and not using, HRT. Second, they must consider the risk of osteoporotic fracture with and without HRT. These are matters of fact and science. In making a decision, however, the panel must also trade off the increased risk of breast cancer with HRT against the decreased risk of osteoporotic fracture. This judgment is a matter of values and preferences. Because they need to make such judgments, it is important that guideline developers report the values and preferences that lie behind their decisions.

Clinicians should look for information about who was involved in assigning values to outcomes or who, by influencing recommendations, was implicitly involved in assigning values. Expert panels and consensus groups often determine what a guideline will say. You need to know who the “experts” are, bearing in mind that panels dominated by members of specialty groups may be subject to intellectual, territorial, and financial biases. Panels that include a balance of experts in research methodology, practicing clinicians, and public representatives are more likely to have considered diverse views in their deliberations than panels restricted to content area experts.

Even with broad representation, the actual process of deliberation can influence recommendations. Therefore, clinicians should look for a report of methods used to synthesize preferences from multiple sources. Informal and unstructured processes may be vulnerable to undue influence by individual panel members, particularly that of the panel chair. Explicit strategies for describing and dealing with dissent among judges, or frank reports of the degree of consensus strengthen the credibility of recommendations.

Knowing the extent to which panelists considered patient preferences is particularly important. Many guideline reports, by their silence on the matter of patient preferences, assume that guideline developers adequately represent patients’ interests. Although they are reported rarely, it would also be valuable to know which principles—such as patient autonomy, nonmaleficence, or distributive justice—were given priority in guiding decisions about the value of alternative interventions. Excellent guidelines will state whether
a guideline is intended to optimize values for individual patients, reimbursement agencies, or society as a whole. Ideally, guidelines will state the underlying value judgments on which they are based.

**Do the Authors Indicate the Strength of Their Recommendations?**

Multiple considerations should inform the strength or grade of recommendations: (1) quality of the sources contributing to the systematic review or reviews that bring together the relevant evidence, (2) magnitude and consistency of the intervention effects in different studies, (3) magnitude of adverse effects, (4) burden to patients and the health care system, (5) costs, and (6) relative value placed on different outcomes. Thus recommendations may vary from those that rely on evidence from a systematic review of randomized controlled trials that show large effects of the intervention on patient-important outcomes with minimal side effects, inconvenience, and costs (yielding a very strong recommendation) to those that rely on evidence from observational studies showing a small magnitude of treatment effect with appreciable side effects and costs (yielding a very weak recommendation).

**RELATIONSHIP BETWEEN LEVELS OF EVIDENCE AND GRADES OF HEALTH CARE RECOMMENDATIONS**

Groups developing recommendations should make sequential judgments about the quality of evidence (i.e., the extent to which one can be confident that an estimate of effect is correct) and the strength of the recommendation (i.e., the extent to which one can be confident that adherence to a recommendation will do more good than harm). These judgments are typically indicated through formal grading systems, one categorizing levels of evidence and one categorizing levels of recommendations.

However, many current grading systems do not adequately distinguish between evidence and recommendations. Although the quality of evidence serves as the principal basis for clinical practice recommendations, the review of the evidence is a conceptually distinct process from the setting of clinical policy. Because of the health, economic, and social implications of clinical practice guidelines, the scientific evidence must be viewed within the context of the clinical practice and health care settings to which the recommendations apply. Examples of factors other than evidence that can affect the grade of a recommendation include questionable transfer of study results to other countries or practice settings, limited availability of a particular technology, demonstrated poor adherence with a procedure, potential for harm, overwhelming burden of suffering, large numbers needed to treat, costly procedures or technology, and the values and preferences that determine decisions about the balance between benefits and harms. For this reason, two rating scales are required: one for levels of evidence and one for grades of recommendations.

Appreciation of the need for these rating scales has led to a proliferation of hierarchies that use various systems of codes to communicate grades of evidence and recommendations, with the most common falling into three categories: letters (e.g., A, B, C, etc.), numbers (e.g., I, II, III, etc.), and a combination of letters and numbers (e.g., Ia,
Interpreting Levels of Evidence and Grades of Recommendations

Chapter 35

The existence of multiple classifications for evaluating and structuring evidence and the differing interpretations of grades of recommendations on the basis of this evidence pose potential problems for clinicians. Table 35-2 illustrates four evidence hierarchies. Considerable differences exist among them with respect to what counts as the highest-quality evidence and what constitutes the strongest possible recommendation.

Scientists develop evidence hierarchies and grades of recommendations to promote the use of evidence-based approaches in health care and provide clinicians with a guide to reliable knowledge. However, the large number of systems threatens to create confusion for clinicians. The various hierarchies are differentially restrictive with respect to what counts as best evidence. They give differential weight to consensus of expert opinion and evidence. To address this problem of multiple hierarchies, the GRADE Working Group has developed yet another system for grading evidence and recommendations. The GRADE Working Group consists of more than 25 scientists from around the world with extensive experience in conducting systematic reviews and developing guidelines. They have developed their system through a critical appraisal of six prominent systems, review of other systems, iterative discussions, and a pilot study. Their system has one key difference from all other systems: it includes detailed guides for working through a recommendation to arrive at a grade of methodological quality and strength of recommendation. Before describing this system, we illustrate some of the problems with existing systems.

**PROBLEMS WITH EXISTING SYSTEMS FOR GRADING EVIDENCE AND HEALTH CARE RECOMMENDATIONS**

The American Diabetes Association has published recommendations to prevent foot ulcers in people with diabetes. The guideline developers used a rating scheme for levels of evidence, which ranged from randomized controlled trials (RCTs), representing the strongest level of evidence, to expert and consensus opinion, representing the lowest level of evidence (I-A: RCT, crossover trials; I-B: controlled trial, nonrandomized; II-A: cohort, case-control; II-B: time-series, pre-post studies, repeated panel; II-C: cross-sectional population-based data; III: descriptive studies, case series, case reports; IV: expert opinion and consensus opinion). Three issues arise in review of the recommendations.

First is the use of expert opinion as level IV evidence. The concept of evidence supported by expert opinion is often ill defined. What do we mean by evidence? In the broadest definition, we mean experiences in the world. When those experiences are reported to others, they become evidence that we can all consider. Consider the following statement: I am an expert in condition \(x\), and I know that treatment \(y\) is extremely effective and provides much more benefit than harm. This statement provides no evidence. It is an opinion. Consider an alternative statement: I have administered treatment \(y\) to 1000 patients with condition \(x\). They all improved markedly with minimal side effects, and treatment \(y\) was associated with little cost and inconvenience. This statement provides evidence. It may be anecdotal and subject to a several biases, but it constitutes evidence.
### Table 35-2: Highest Level of Evidence and Strongest Grade of Recommendation in Four Evidence Hierarchies

<table>
<thead>
<tr>
<th>Source of Evidence Hierarchy</th>
<th>Highest Level of Evidence for a Treatment or Intervention</th>
<th>Conditions for Strongest Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Task Force on Preventive Health Care(^\text{14})</td>
<td>1 = at least one RCT</td>
<td>A study (including meta-analysis or systematic review) that meets all design-specific criteria well</td>
</tr>
<tr>
<td>Scientific Advisory Council of the Osteoporosis Society of Canada(^\text{15})</td>
<td>1+ = systematic review or meta-analysis of RCTs 1 = one randomized trial with adequate power</td>
<td>Supportive level 1 or 1+ evidence plus consensus</td>
</tr>
<tr>
<td>Centre for Evidence-Based Medicine(^\text{16})</td>
<td>1a = systematic review with homogeneity of RCTs 1b = one RCT with narrow confidence interval 1c = all or none*</td>
<td>Consistent level 1 studies</td>
</tr>
<tr>
<td>Scottish Intercollegiate Guidelines Network(^\text{17})</td>
<td>1++ = high-quality meta-analyses, systematic reviews of RCTs, or RCTs with very low risk of bias 1+ = well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with low risk of bias 1 = meta-analyses, systematic reviews of RCTs, or RCTs with high risk of bias</td>
<td>At least one meta-analysis, systematic review, or RCT rated as 1++ and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+ directly applicable to the target population and demonstrating overall consistency</td>
</tr>
</tbody>
</table>

RCT, Randomized controlled trial.

*The Center for Evidence-Based Medicine explains “all or none” as follows: “met when all patients died before the treatment became available but some now survive on it; or when some patients died before the treatment became available but none now die on it.”

Modified and reproduced with permission of the Canadian Medical Association from Upshur RE. Are all evidence-based practices alike? Problems in the ranking of evidence. CMAJ. 2003;169:672-673.
Systems of grading recommendations that specify “expert opinion” as a level of evidence ignore the distinction between opinion and evidence. Perhaps such systems are implicitly suggesting that expert members of the panel bring their individual clinical experience to bear on the decision and are therefore providing evidence. If so, an explicit and clear statement of that assumption would be helpful. Moreover, if that is what they are saying, such systems need to confront the fact that many experts become increasingly distant from clinical practice with increasing years of widely acknowledged expertise and that evidence available only within the panel (in contrast to, for example, published case series) lacks transparency. Clearly specifying that evidence, and not opinion, bears on grading the strength of evidence represents a superior approach to this issue.

Second, although the guideline developers outlined the rating scheme for the strength of evidence, they did not link the strength of evidence to each recommendation. Readers are left with an undifferentiated list of recommendations with no accompanying information about the strength of evidence for each recommendation.

Finally, the guideline developers did not provide information about the strength of the recommendations. With the absence of grades of recommendations, readers have no information about the balance between benefits and harm, adherence problems, and cost implications of implementing the recommendations.

The New Zealand Guidelines Group (NZGG) developed an evidence-based best practice guideline for health professionals working in cardiac rehabilitation in hospitals and communities in New Zealand.22 The group rated the evidence using an adapted version of the Scottish Intercollegiate Guidelines Network (SIGN) grading system for recommendations in evidence-based guidelines.17 Table 35-3 outlines the levels of evidence and grades of recommendations. The highest level of evidence (1++) is derived from high-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias, and the lowest level of evidence (4) is derived from expert opinion. A recommendation is assigned the highest grade of A if it is substantiated by at least one meta-analysis, systematic review, or RCT rated as 1++, and if it is directly applicable to the target population; or if it is substantiated by a body of evidence consisting principally of studies rated as 1+, and if it is directly applicable to the target population and demonstrates overall consistency of results. A recommendation is assigned the lowest grade of D if it is based on level 3 or 4 evidence or is extrapolated from studies rated as 2+, or expert opinion.

These guideline developers created separate levels of evidence and grades of recommendation, but only the grades of recommendation (which incorporate the levels of evidence) accompany the recommendation statements. As with the guideline on foot ulcers described previously, they included expert opinion as the weakest level of evidence. The grades of recommendation are based largely on the levels of evidence, although the guideline developers also stipulate direct applicability to the target population for levels A, B, and C.

The biggest problem with this approach is that it ignores the fact that one can have very high-quality evidence accompanying a weak recommendation. Consider a series of large, well-planned, and brilliantly executed randomized trials of the highest methodological quality summarized in a rigorous systematic review and meta-analysis. All of us
Table 35-3 Grading System Used by New Zealand Guidelines Group in Cardiac Rehabilitation Guideline

### Levels of Evidence

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1++</td>
<td>High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias</td>
</tr>
<tr>
<td>1+</td>
<td>Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias</td>
</tr>
<tr>
<td>1−</td>
<td>Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias</td>
</tr>
<tr>
<td>2++</td>
<td>High-quality systematic reviews of case-control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal</td>
</tr>
<tr>
<td>2+</td>
<td>Well-conducted case-control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal</td>
</tr>
<tr>
<td>2−</td>
<td>Case-control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal</td>
</tr>
<tr>
<td>3</td>
<td>Nonanalytic studies (case reports, case series)</td>
</tr>
<tr>
<td>4</td>
<td>Expert opinion</td>
</tr>
</tbody>
</table>

### Grades of Recommendation

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population, OR a body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results</td>
</tr>
<tr>
<td>B</td>
<td>A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results, OR extrapolated evidence from studies rated as 1++ or 1+</td>
</tr>
<tr>
<td>C</td>
<td>A body of evidence including studies rated as 2+, directly applicable to the target population, and demonstrating overall consistency of results, OR extrapolated evidence from studies rated as 2++</td>
</tr>
<tr>
<td>D</td>
<td>Evidence levels 3 or 4, OR extrapolated evidence from studies rated as 2+, or expert opinion</td>
</tr>
</tbody>
</table>


would classify this as highest-quality evidence. Now, what would we recommend if the intervention was administered to 1000 people and prevented a single premature death. All 1000 people, however, became moderately to severely nauseated for 6 months, and 100 of these people lost all their hair. Would we recommend this intervention? Some of us might, whereas others would not. But whatever our decision, none of us would claim that our recommendation is a strong one. Thus, systems such as SIGN that assume that high-quality evidence leads to strong recommendations are problematic.
The Canadian Task Force on Preventive Health Care developed recommendations on the effectiveness of specific screening techniques for colorectal cancer in asymptomatic patients. Table 35-4 outlines the levels of evidence and grades of recommendations used by the Task Force. The strongest level of evidence (I) is derived from at least one well-designed randomized controlled trial, and the weakest level of evidence (III) is derived from opinions of respected authorities, clinical experience, or descriptive

Table 35-4  Canadian Task Force on Preventive Health Care Levels of Evidence and Grades of Recommendations*

Levels of Evidence

<table>
<thead>
<tr>
<th>Levels</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Evidence from at least one well-designed randomized controlled trial</td>
</tr>
<tr>
<td>II-1</td>
<td>Evidence from well-designed controlled trials without randomization</td>
</tr>
<tr>
<td>II-2</td>
<td>Evidence from well-designed cohort or case-control analytic studies, preferably from more than one center or research group</td>
</tr>
<tr>
<td>II-3</td>
<td>Evidence from comparisons between times or places with or without the intervention: dramatic results from uncontrolled studies could be included here</td>
</tr>
<tr>
<td>III</td>
<td>Opinions of respected authorities, based on clinical experience; descriptive studies or reports of expert committees</td>
</tr>
</tbody>
</table>

Grades of Recommendations

<table>
<thead>
<tr>
<th>Grades</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Good evidence to support the recommendation that the condition or maneuver be specifically considered in a periodic health examination (PHE)</td>
</tr>
<tr>
<td>B</td>
<td>Fair evidence to support the recommendation that the condition or maneuver be specifically considered in a PHE</td>
</tr>
<tr>
<td>C</td>
<td>Insufficient evidence regarding inclusion or exclusion of the condition or maneuver in a PHE, but recommendations may be made on other grounds</td>
</tr>
<tr>
<td>D</td>
<td>Fair evidence to support the recommendation that the condition or maneuver be specifically excluded from a PHE</td>
</tr>
<tr>
<td>E</td>
<td>Good evidence to support the recommendation that the condition or maneuver be specifically excluded from a PHE</td>
</tr>
</tbody>
</table>

*Since this systematic review was published, the Canadian Task Force on Preventive Health Care (CTFPHC) has made revisions to their grades of recommendations. Grade C now means the following: “Existing evidence is conflicting and does not allow making a recommendation for or against use of the clinical preventive action.” They have also added a Grade I recommendation indicating “insufficient evidence (in quantity and/or quality) to make a recommendation.” For more detailed information about the most recent levels of evidence and grades of recommendations, please refer to the CTFPHC History/Methodology section on their Web site at www.ctfphc.org. Modified and reproduced with permission of the Canadian Task Force on Preventive Health Care from McLeod R, with the Canadian Task Force on Preventive Health Care. Screening Strategies for Colorectal Cancer: Systematic Review & Recommendations. CTFPHC Technical Report #01-2. February 2001. London, Ontario: Canadian Task Force. Available at: http://www.ctfphc.org
studies or reports of expert committees. With respect to this weakest level of evidence, parts of it constitute evidence (clinical experience and descriptive studies), but parts of it do not (opinions of respected authorities and reports of expert committees).

A recommendation is assigned the highest grade of A if good evidence suggests that the condition or maneuver should be specifically considered in a periodic health examination; a grade of C if there is insufficient evidence; and a grade of E if good evidence suggests that the condition or maneuver should be specifically excluded from the periodic health examination.

In this system, Grade C is not a recommendation at all. For example, a man asks a nurse practitioner if he should be tested with prostate-specific antigen (PSA). The NP refers to a guideline for information about what to advise the patient. Using the Canadian Task Force system, a guideline tells the NP that there is insufficient evidence to decide (which is, at the time of this writing, the case for PSA screening). The NP is no further ahead in knowing what to advise the patient because there is no recommendation. For clinicians seeking guidance about how to proceed in the absence of adequate evidence (and, after all, the patient and NP must proceed one way or the other, screen or not screen), the guideline has not been helpful.

Unlike the previous two examples, each recommendation statement has both a level of evidence and a grade of recommendation. This is useful because it is possible, as noted previously, for any one recommendation statement to have strong evidence and a weak recommendation. For example, although administration of anticoagulant therapy in atrial fibrillation is supported by strong evidence, it might warrant only a weak recommendation for implementation in rural settings, where travel distances are large and anticoagulant intensity monitoring is difficult, or in more severely resource-constrained settings where there is a direct inverse relationship between the resources available for purchase of antibiotic drugs and those allocated to monitoring levels of anticoagulation.

It is also possible to have weak evidence and a strong recommendation. Evidence for the long-term health benefits of exercising in one’s youth comes from weak observational studies. We may nevertheless provide a strong recommendation for exercise on the basis that costs and side effects are negligible.

These three examples of health care recommendations illustrate the diversity in the presence and nature of hierarchies that rank levels of evidence and grades of recommendations. Given the existence of numerous and varied systems, readers must look for and review the specific hierarchy for levels of evidence and grades of recommendations so that the strength of the recommendations can be accurately interpreted.

**A NEW GRADING SYSTEM**

The Grades of Recommendation Assessment, Development, and Evaluation (GRADE) Working Group has developed a new system for evaluating evidence and recommendations. We propose consideration and adoption of this system because it has been developed by many experienced scientists from around the world based on critical appraisal of other systems, iterative discussions, and a pilot study. This system facilitates consideration of the trade-off between benefits and risks and includes detailed guides...
for working through a recommendation to arrive at a grade of methodological quality and strength of recommendation.

**Levels of Evidence**

With respect to quality of evidence, the group suggests that systematic reviews guide judgments. Reviewers should consider four key elements of quality: study design, study quality, consistency, and directness. *Study designs* are broadly categorized as randomized controlled trials, quasi-randomized trials (trials using nonrandom allocation such as alternation), and observational studies. Some study designs are more subject to bias than others, and therefore provide weaker justification for clinical decisions.18 As a result, observational studies do not always predict the findings of subsequent randomized trials, and they frequently overestimate treatment effects.19 A dramatic example of such a discrepancy occurred when the results of observational studies that suggested HRT decreased the risk of coronary heart disease were contradicted by subsequent randomized trials that found no such reduction in risk and the possibility of an increased risk.26,27

*Study quality* refers to the detailed study design and execution. Reviewers should use appropriate criteria to assess quality for each important outcome.19 For randomized trials, for example, reviewers might use criteria such as the adequacy of allocation concealment, blinding, and follow-up (see Chapter 4, Health Care Interventions). Reviewers should make explicit their reasons for downgrading a quality rating. For example, they may state that failure to blind patients and clinicians reduced the quality of evidence for an intervention’s impact on pain severity and that they considered this to be a serious flaw.19

*Consistency* refers to the similarity of estimates of effect across studies. Differences in the direction of effect, the size of differences, and the statistical significance of differences decrease our confidence in the estimate of effect for a particular outcome.19 *Directness* refers to the extent to which the people, interventions, and outcome measures are similar to those of interest. For example, there may be uncertainty about the directness of the evidence if the patients of interest are older, sicker, or have more comorbidity than those included in the studies (see Chapter 33, Applying Results to Individual Patients).19 To determine whether important uncertainty exists, one can ask whether there is a compelling reason to expect important differences in effect sizes. Use of surrogate outcomes for outcomes that are important to patients (e.g., bone density as a surrogate for fractures) often results in appropriately increased uncertainty (see Chapter 13, Surrogate Outcomes).19

To grade levels of evidence, the Working Group suggests first categorizing the evidence based on study design (Table 35-5) and subsequently considering whether the studies have serious flaws, important inconsistencies in the results, or whether uncertainty about the directness of the evidence is warranted. They suggest the following definitions in grading the quality of evidence19:

**High:** Further research is very unlikely to change our confidence in the estimate of effect.

**Moderate:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Table 35-5  Criteria for Assigning Grade of Evidence

Type of Evidence
Randomized trial = high
Observational study = low
Any other evidence = very low

Decrease Grade if:
• Serious (−1) or very serious (−2) limitation to study quality
• Important inconsistency (−1)
• Some (−1) or major (−2) uncertainty about directness
• Imprecise or sparse data (−1)
• High probability of reporting bias (−1)

Increase Grade if:
• Strong evidence of association—significant relative risk of > 2 (< 0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1)
• Very strong evidence of association—significant relative risk of > 5 (< 0.2) based on direct evidence with no major threats to validity (+2)
• Evidence of a dose-response gradient (+1)
• All plausible confounders would have reduced the effect (+1)

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.

Very low: Any estimate of effect is very uncertain.

Limitations in study quality, important inconsistency of results, or uncertainty about the directness of evidence can lower the grade of evidence. For instance, if all available studies have serious flaws, the grade will drop by one level, and if all studies have very serious flaws, the grade will drop by two levels. Fatally flawed studies may be excluded from consideration. Limitations act cumulatively. For example, if randomized trials have both serious flaws and uncertainty exists about the directness of the evidence, the grade of evidence drops from high to low.

Grading of Recommendations
Recommendations involve a trade-off between benefits and harms. Making that trade-off involves placing a value on each outcome. Different people will have different values and, therefore, it is difficult to judge how much weight to give to different outcomes. Grading recommendations entails making explicit judgments about the balance between the main health benefits and harms. Recommendations must apply to specific settings and groups of patients whenever the benefits and harms differ across settings or patient groups. The Working Group suggests a simple scheme to, in effect, categorize...
recommendations as strong (the trade-offs are clear) or weak (benefits and downsides are closely balanced). The final grade of recommendation becomes very explicit: “Do it” or “Don’t do it”: Indicating a judgment that most well-informed people will make the same choice.

“Probably do it” or “Probably don’t do it”: Indicating a judgment that most well-informed people will make the same choice, but a substantial minority will not.

In some instances, it may not be appropriate to make a recommendation because of unclear trade-offs or lack of agreement (although, as pointed out earlier, this is unhelpful for clinicians looking for guidance).

**EVALUATING THE TRADE-OFF BETWEEN BENEFITS AND RISKS**

Depending on the balance between benefits and risks, methodologically strong studies that suggest a benefit of one intervention over another intervention may lead to varying recommendations. When side effects are minimal, risk reductions are large, or a patient’s risk of the target event that the intervention will prevent is very high, guideline developers may make a strong recommendation to administer the more effective intervention to patients with compatible values and preferences. When benefits and risks are closely balanced, we may see conflicting recommendations and practice. When risk reductions are small and toxicity is high, guideline developers may even recommend the less effective intervention or recommend not to treat at all. As the magnitude of benefits and risks become more closely balanced, decisions about administration of effective interventions also become more cost sensitive.

The categories for recommendations suggested by the GRADE Working Group reflect the balance between benefits and risks of health care interventions. If the benefits clearly outweigh the risks (or vice versa) and virtually all patients would make the same choice, the recommendation is designated as “do it” or “don’t do it.” An example is the prophylaxis of deep venous thrombosis after hip fracture surgery, in which heparin reduces the risk of deep venous thrombosis by approximately 40%. Here, because sample sizes of the studies are relatively large and confidence intervals are narrow, and because prophylaxis is associated with low costs and complications, benefits clearly outweigh the downsides of treatment, and the recommendation is strong. If they understand the benefits and risks, virtually all patients will comply with prophylaxis to reduce thromboembolism after hip fracture. Thus, one way of thinking about a “do it” or “don’t do it” recommendation is that variability in patient values or individual clinician values is unlikely to influence treatment choice in typical patients.

When the balance is less certain and different patients may make different choices, they designate the recommendation as “probably do it” or “probably don’t do it.” Several factors may create uncertainty in the balance between benefits and risks, including marked variation in patient values, a wide range of confidence intervals around estimates of benefits and risks, high costs, or small effect sizes (see Chapter 10, Moving From Evidence to Action Using Clinical Practice Guidelines).

If the balance between benefits and risks is uncertain, we may have methodologically rigorous studies but recommendations may still be weak (“probably do it” or “probably
don’t do it”). The more balanced the trade-off between benefits and risks, the greater is the influence of individual patient values in decision making. When the trade-off between benefits and risks is less clear, clinicians will want to ensure that individual patient values bear strongly on the final decision (see Chapter 34, Incorporating Patient Values). In considering the duration of anticoagulation after an episode of idiopathic deep venous thrombosis, patients may make different choices depending on the relative value they place on avoiding a fatal pulmonary embolus, avoiding bleeding, and the inconvenience and worry associated with repeated testing to determine the intensity of anticoagulation. “Probably do it” or “probably don’t do it” recommendations are those in which variation in patient values or individual clinician values will often mandate different treatment choices, even among typical patients.

CONCLUSION

In this chapter, we outlined the process of developing recommendations, summarized criteria for critically evaluating the methodological quality of recommendations, and introduced a taxonomy for grading evidence and recommendations that facilitates consideration of the trade-off between benefits and risks. We have illustrated the importance and added value of having information on both the level of evidence and the grade of recommendation and how the two are not always entirely dependent on one another.

REFERENCES


