ICSI
INSTITUTE FOR CLINICAL SYSTEMS IMPROVEMENT

EVIDENCE GRADING SYSTEM

The evidence grading system used in ICSI guidelines and technology assessment reports is periodically reviewed and modified. The version presented below was approved in **November, 2003**. An extended description of the development of the system and ICSI's experience and results using the evidence grading system may be found in the article: Greer N, Mosser G, Logan G, Halaas G. A practical approach to evidence grading. *Joint Commission Journal on Quality Improvement* 26:700-712, 2000.

Development

Evidence grading was introduced into ICSI guidelines and technology assessment reports in 1996. At that time, a modification of the system used in the Agency for Health Care Policy and Research (AHCPR) Unstable Angina: Diagnosis and Management Clinical Practice Guideline was used[1]. It soon became obvious that this system was too simplistic, there were objections to grading conclusions strictly on research design type, there was concern that there was no consideration for how much evidence there was, and there was concern that all design types were not adequately considered.

In 1997, ICSI assembled a work group of physicians and researchers with backgrounds in quality improvement, clinical epidemiology, and biostatistics to make recommendations for changes to the evidence grading system. The evidence grading review work group established goals for an evidence grading system. These goals were:

1. to increase the systematic use of evidence by work groups by providing a framework and a step-by-step process for reaching key conclusions;
2. to provide a method for reaching evidence-based conclusions that busy, practicing clinicians accept as practical;
3. to provide a reliable method for grading conclusions based on the strength of the underlying evidence; and
4. to convey to readers and users of the documents the strength of the underlying evidence.

The work group also reviewed many existing evidence grading systems including the system used by the United States Preventive Services Task Force[2], the system developed by Sacket[3] and modified by Cook et al.[4,5], and the system presented in the series on Users' Guides to the Medical Literature[6]. Although the ICSI work group decided that no one existing system fulfilled the goals identified above, there were features of the existing systems that could be incorporated into a new ICSI system. Specifically, the work group agreed that it was important to separate the evaluation of individual research reports from the assessment of the totality of evidence supporting a conclusion. The work group also agreed that assessing the quality of the individual research reports was important.

The System

The centerpiece of the evidence grading system is the conclusion grading worksheet. Conclusion grades are assigned to key conclusions and/or recommendations as determined by the guideline or technology assessment work group members. The worksheet, similar to an evidence table, is used to display and synthesize the evidence supporting a particular conclusion. An example of a worksheet from the Congestive Heart Failure guideline is presented in Figure 1. The work group formulates a tentative conclusion statement and, based on a literature search done by a medical librarian using keywords suggested by the work group, identifies the key references to include on the worksheet. The work group is encouraged to identify the strongest possible evidence (based on design type, sample size, patient population, etc.) that supports or disputes the conclusion statement. The worksheet is then prepared by ICSI staff and includes, for each reference, the citation, design type, class of research report, quality score, information about the population studied, results of the study, and the authors' conclusions. The conclusion grading worksheet is reviewed by a designated member of the work group and a tentative conclusion grade is selected. The designated work group member then presents the worksheet to the rest of the work group. There is discussion of the individual research reports and comments from the work group may be added to the worksheet. There is also discussion of the proposed conclusion grade and a final decision is made on the appropriate grade. Involvement of a member of the work group in the development of the worksheet and the deliberation by the work group in determining the final conclusion grade are considered strengths of the system. Further information about the classes of research reports, quality scores, and conclusion grades is presented in Tables 1-3.

Figure 1. Conclusion Grading Worksheet

Work Group's Conclusion: Digoxin improves symptoms, exercise tolerance, and quality of life, but neither increases or decreases mortality.

Conclusion Grade: I

Author/Year	Design Type	Class	Quality +,–,ø	Population Studied/Sample Size	Primary Outcome Measure(s)/Results (e.g., p-value, confidence interval, relative risk, odds ratio, likelihood ratio, number needed to treat)	Authors' Conclusions/ *Work Group's Comments (italicized)*
Captopril-Digoxin Research Group (1988)	RCT	A	ø	-Patients (<75 years) with sinus rhythm and heart failure secondary to ischemic heart disease, primary myocardial disease, or in heart failure without significant valvular regurgitation after valvular surgery (receiving diuretic therapy if needed) -Randomized to 1 of 3 groups (after withdrawal from therapy and stabilization of diuretic dose): captopril (25 mg 3x/day increased to 50 mg 3x/day after 1 wk if tolerated), digoxin (0.125-0.375 mg daily based on trough serum levels), or placebo -Included: ejection fraction ≤40%, treadmill time >4 min but < age- and sex-predicted average maximum -Excluded: MI within preceding 2 mos, unstable angina, hypertension (SBP>160 mmHg, DBP >95 mmHg) despite diuretic therapy, pulmonary disease (FEV^1/FVC ratio <60%) -Concomitant therapy with inotropic agents, vasodilators, β-adrenergic blockers, calcium antagonists, immunosuppressive agents, or other investigational drugs was prohibited	-Follow-up at 1, 2, & 6 mos after randomization -300 patients randomized (104 to captopril, 96 to digoxin, and 100 to placebo); baseline characteristics were similar (captopril group younger, p=0. 02) -Mean changes from baseline (analysis while adhering to assigned therapy): Variable / Captopril / Digoxin / Placebo Exer. time (s) / 82* n=101 / 54 n=95 / 35 n=97 NYHA class / -0.20** n=100 / -0.09 n=95 / 0.02 n=98 Eject. fraction (%) / 1.8 n=87 / 4.4*** n=82 / 0.9 n=78 Premature beats per hour# / -29.4**** n=55 / 2.3 n=47 / -16.1 n=45 *different from placebo (p<0.05); ** different from placebo with respect to proportion of patients improved (p<0.01) (see below); *** different from placebo (p<0.01) and captopril (p<0.05) groups; **** different from digoxin (p<0.05); #only patients with >10 ventricular premature beats/hr at baseline -NYHA class improved for 41% of captopril, 31% of digoxin, and 22% of placebo groups -Withdrawal from study because of treatment failure occurred with 15% of placebo group (vs. 5.8% of captopril group and 4.2% of digoxin group; p<0.05); more patients in placebo group required increase in diuretic dose (p<0.005) and hospitalization (or emergency visits) (p<0.05) than in other groups -Similar trends seen in intention-to-treat analysis -Rate of discontinuation due to adverse drug reactions: 2.9% captopril, 4.2% digoxin, 0% placebo -More possible adverse drug effects attributed to captopril (44.2%) during blinded portion of study than to other treatments (30.2% digoxin, 24% placebo) (usually mild and transient dizziness and light-headedness) -21 deaths (8 captopril, 7 digoxin, 6 placebo)	-Captopril therapy is significantly more effective than placebo and is an effective alternative to digoxin treatment in patients with mild to moderate heart failure who are undergoing maintenance diuretic therapy. Significant improvements in exercise tolerance and functional class compared to the placebo group were seen in the captopril group but not the digoxin group. Captopril also significantly reduced ventricular premature beat rates compared with digoxin in patients with more than 10 premature beats/hour at baseline. Digoxin significantly increased left ventricular ejection fractions compared with both placebo and captopril. Patients receiving placebo had a greater incidence of treatment failure and required significantly more diuretics, hospitalizations, and/or emergency department visits for heart failure than did patients receiving captopril or digoxin. NOTES: trial was double-blind; did intention-to-treat analysis (as well as analysis while patients adhered to assigned therapy); most patients were NYHA functional class II;

Digitalis Investigation Group (1997)	RCT	A	ø	-6800 patients (302 clinical centers) with heart failure and left ventricular ejection fraction < 0.45 in normal sinus rhythm -988 patients with heart failure and ejection fraction >0.45 were enrolled in ancillary trial -May have been already receiving digoxin -Randomly assigned to digoxin or placebo (digoxin dose varied) -Other therapy used if patient had worsening symptoms of heart failure; if remained symptomatic, allowed open-label treatment with digoxin	-Follow-up visits at 4 and 16 wks then every 4 mos (mean duration 37 mos, range 28-58 mos) -No significant differences between groups (baseline) -1181 deaths in digoxin group (34.8%), 1194 in placebo group (35.1%) (RR=0.99, 95%CI: 0.91-1.07) -1016 deaths from cardiovascular causes in digoxin group (29.9%), 1004 in placebo group (29.5%) (RR=1.01; 95%CI: 0.93-1.10) -Trend toward lower risk of mortality attributable to worsening heart failure in digoxin group (p=0.06) -910 patients hospitalized for worsening heart failure in digoxin group and 1180 in placebo group (RR=0.72; 95%CI: 0.66-0.79) -Risk of death from any cause or hospitalization for worsening heart failure was lower in digoxin group (RR=0.85; 95% CI: 0.79-0.91); similar results for death due to worsening heart failure or hospitalization related to worsening heart failure -Fewer hospitalizations for any cause (per patient) in digoxin group (p=0.01) and for cardiovascular causes (p<0.001) -Benefit of digoxin appeared to be greater among patients at high risk (lower ejection fraction, enlarged heart, or NYHA III or IV) -At 1 yr, 85.6% of digoxin group patients were taking study drug and 82.9% of placebo group were taking placebo; at final study visit 70.8% of surviving patients in digoxin group were taking study drug and an additional 10.3% were taking open-label digoxin; 67.9% of surviving placebo group patients were taking placebo and 15.6% were taking open label digoxin. -Suspected digoxin toxicity greater in digoxin group -In ancillary trial, no difference in number of deaths or combined outcome of death or hospitalization due to worsening heart failure	-In patients with left ventricular ejection fractions ≤0.45 digoxin had no effect on overall mortality when added to diuretics and ACE inhibitors; the risk of hospitalization was reduced and the combined outcome of death or hospitalization attributable to worsening heart failure was also reduced. In clinical practice, digoxin therapy is likely to decrease the frequency of hospitalization but not survival. NOTES: exclusion criteria were not given in this publication (previously published); trial was double-blind; did intention-to-treat analysis; physicians were strongly encouraged to give patients ACE inhibitors; patients receiving digoxin at entry were randomly assigned with no washout period; vital status of 47 patients in digoxin group and 46 in placebo group (1.4% of total) were unknown (a sensitivity analysis assuming that either all placebo or all digoxin patient died did not change overall result)

Figure 1. This is an example of a worksheet included in the ICSI guideline on Congestive Heart Failure in Adults. The work group identifies the key research articles to be summarized on the worksheet and, once the information is entered on the worksheet, determines the appropriate conclusion statement and conclusion grade. The worksheets appear as an Appendix to the guideline.

Classes of Research Reports

Each individual research report cited in a guideline or technology assessment report is assigned a class by ICSI staff (see Table 1). Primary reports of new data collection are assigned a letter A, B, C, or D based on the design type. The hierarchy of design types (with "A" representing randomized, controlled trials etc.) is fairly consistent among evidence grading systems and reflects the fact that different study design types vary in the likelihood that an individual study will be biased[7]. Secondary reports (reports that synthesize or reflect upon collections of primary reports) are assigned an M, an R, or an X. The definitions of the various design types are those found in epidemiology textbooks[8,9,10].

Table 1. Classes of Research Reports

Primary Reports of New Data Collection	
A	randomized, controlled trial
B	cohort study
C	nonrandomized trial with concurrent or historical controls case-control study study of sensitivity and specificity of a diagnostic test population-based descriptive study

D	cross-sectional study
	case series
	case report

Reports that Synthesize or Reflect Upon Collections of Primary Reports	
M	meta-analysis systematic review decision analysis cost-effectiveness analysis
R	consensus statement consensus report narrative review
X	medical opinion

Research Report Quality Categories

The quality of individual research reports (either primary reports or systematic reviews) is designated as plus (+), minus (-), or neutral (ø) based on the questions presented in Tables 2a and 2b. The quality considerations reflected in the tables are considerations standardly addressed in textbooks of clinical epidemiology[9,10]. The assessment of quality is completed by ICSI staff.

Table 2a. Primary Research Report Quality Categories

PLUS (+)

Y N 1. Were the inclusion and exclusion criteria exceptionally well-defined and adhered to?

Y N 2. Were no serious questions of bias introduced in the study (e.g., through the processes of subject selection, end point selection, and observation or data collection)?

Y N 3. Does the report show a statistically significant and clinically important treatment effect or, for a negative conclusion, have high power?

Y N 4. Are the results widely generalizable to other populations?

Y N 5. Were other characteristics of a well-designed study clearly addressed in the report (e.g., treatment and control groups comparable at baseline, compliance with the intervention, use of intention to treat analysis, all important outcomes measured, statistics appropriate for study design)?

If the answer to 2 or more of the above questions is "yes", the report may be designated with a plus on the Conclusion Grading Worksheet depending on the work group's overall evaluation of the report.

MINUS (–)

Y N 1. Were the inclusion and exclusion criteria unclear or was there evidence of failure to adhere to defined criteria?

Y N 2. Were serious questions of bias introduced in the study (e.g., through the processes of subject selection, end point selection, and observation or data collection)?

Y N 3. Does the report show a statistically significant but clinically insignificant effect or, for a negative conclusion, lack power and sample size?

Y N 4. Are the results doubtfully generalizable to other populations?

Y N 5. Were other characteristics of a poorly-designed study clearly evident in the report (e.g., treatment and control groups different at baseline, low compliance with the intervention, important outcomes were not measured, inappropriate statistics for study design)?

If the answer to 2 or more of the above questions is "yes", the report may be designated with a minus symbol on the Conclusion Grading Worksheet depending on the work group's overall evaluation of the report.

NEUTRAL (Ø)

If the answers to the questions pertaining to the PLUS or MINUS criteria do not indicate that the report is exceptionally strong or exceptionally weak, the report should be designated with a neutral symbol on the Conclusion Grading Worksheet.

Table 2b. Systematic Review Quality Categories

PLUS (+)

Y N 1. Was the search for primary studies comprehensive (i.e., multiple sources) and current? Were clear criteria given for inclusion and exclusion of primary studies?

Y N 2. Was the quality of the articles included in the analysis assessed and reported?

Y N 3. If a meta-analysis was done, was homogeneity assessed? If no meta-analysis was done, did the authors state why not? Was there at least a narrative synthesis of the primary studies?

Y N 4. Are the primary studies included in the review applicable (i.e., generalizable) to the target population?

Y N 5. Are the conclusions valid (i.e., based on the primary evidence)?

If the answer to two or more of the above questions is "yes," the report may be designated with a plus on the Conclusion Grading Worksheet depending on the work group's overall evaluation of the report.

MINUS (-)

Y N 1. Was the search for primary studies incomplete or outdated? Were criteria for inclusion and exclusion of primary studies unclear or absent?

Y N 2. Was no attempt made to assess the quality of the primary studies included in the analysis?

Y N 3. If a meta-analysis was done, was the potential for homogeneity disregarded? If no meta-analysis was done, would such an analysis have been appropriate?

Y N 4. Are the primary studies included in the review doubtfully generalizable to the target population?

Y N 5. Are the conclusions doubtful based on the primary evidence?

If the answer to two or more of the above questions is "yes," the report may be designated with a minus on the Conclusion Grading Worksheet depending on the work group's overall evaluation of the report.

NEUTRAL (∅)

If the answers to the questions pertaining to PLUS or MINUS criteria do not indicate that the report is exceptionally strong or exceptionally weak, the report should be designated with a neutral symbol on the Conclusion Grading Worksheet.

Conclusion Grades

Conclusions and recommendations are graded either I, II, III, or Grade Not Assignable. Descriptions of the conclusion grades as well as examples of the types of evidence that would support a specific grade are presented in Table 3.

Table 3. Conclusion Grades

Grade I: The conclusion is supported by good evidence.

The evidence consists of results from studies of strong design for answering the question addressed. The results are both clinically important and consistent with minor exceptions at most. The results are free of any significant doubts about generalizability, bias, and flaws in research design. Studies with negative results have sufficiently large samples to have adequate statistical power.

Examples:
Supporting studies might consist of two or more randomized, controlled trials with consistent results or even a single well designed, well executed trial. The evidence might also come from a systematic review containing a meta-analysis of several trials with comparable methodologies and consistent results. For a question of the soundness of a diagnostic test, the evidence might be the results of a single well done comparison of the test against an established test for the same purpose, provided that there is no evidence to the contrary. For a question of the natural history of a disease, in the absence of evidence to the contrary, the evidence might be results from a single well done prospective cohort study.

Grade II: The conclusion is supported by fair evidence.

The evidence consists of results from studies of strong design for answering the question addressed, but there is some uncertainty attached to the conclusion because of inconsistencies among the results from the studies or because of minor doubts about generalizability, bias, research design flaws, or adequacy of sample size. Alternatively, the evidence consists solely of results from weaker designs for the question addressed, but the results have been confirmed in separate studies and are consistent with minor exceptions at most.

Examples:
Supporting studies might consist of three or four randomized, controlled trials with differing results although overall the results support the conclusion. The evidence might also be the results of a single randomized, controlled trial with a clinically significant conclusion but doubtful generalizability. Alternatively, the evidence might come from a systematic review containing a meta-analysis of randomized trials with similar methodologies but differing results. For a question of causation, the evidence might consist of two independent case-control studies with similar conclusions. The evidence might also consist of several careful case series reports with similar conclusions from investigators working separately.

Table 3. Conclusion Grades (continued)

Grade III: The conclusion is supported by limited evidence.

The evidence consists of results from studies of strong design for answering the question addressed, but there is substantial uncertainty attached to the conclusion because of inconsistencies among the results from different studies or because of serious doubts about generalizability, bias, research design flaws, or adequacy of sample size. Alternatively, the evidence consists solely of results from a limited number of studies of weak design for answering the question addressed.

Examples:
For a question of efficacy of medical treatment, the evidence might consist of three or four randomized trials with contradictory results or serious methodological flaws; or the evidence might be a systematic review of several trials with contradictory results or serious methodological flaws. The evidence might also consist of a single trial that used historical controls. Alternatively, for a question of efficacy, the evidence might consist of one case series report. For a question of causation, the evidence might consist of results from a single case-control study, unconfirmed by other studies.

Grade Not Assignable: There is no evidence available that directly supports or refutes the conclusion.

There is no evidence that directly pertains to the conclusion because either the studies have not been done or the only relevant information is in the form of medical opinion papers.

Examples:
The literature cited might consist of a review article citing only single case reports. The literature cited might also be an editorial, a consensus report, or a position statement from a national body without citations of the results of research studies. (In both cases, if research studies are cited, they should govern the assignment of the grade to the conclusion.) Alternatively, the literature cited may be of strong design but the outcome measures do not have direct bearing on the question being addressed in the conclusion.

Summary of Process

The process for reaching a conclusion grade, specifically for a guideline in development, is summarized in Figure 2. For guidelines undergoing revision and for technology assessment reports, a similar process is followed.

Figure 2. Conclusion Grading Process for Guidelines in Development

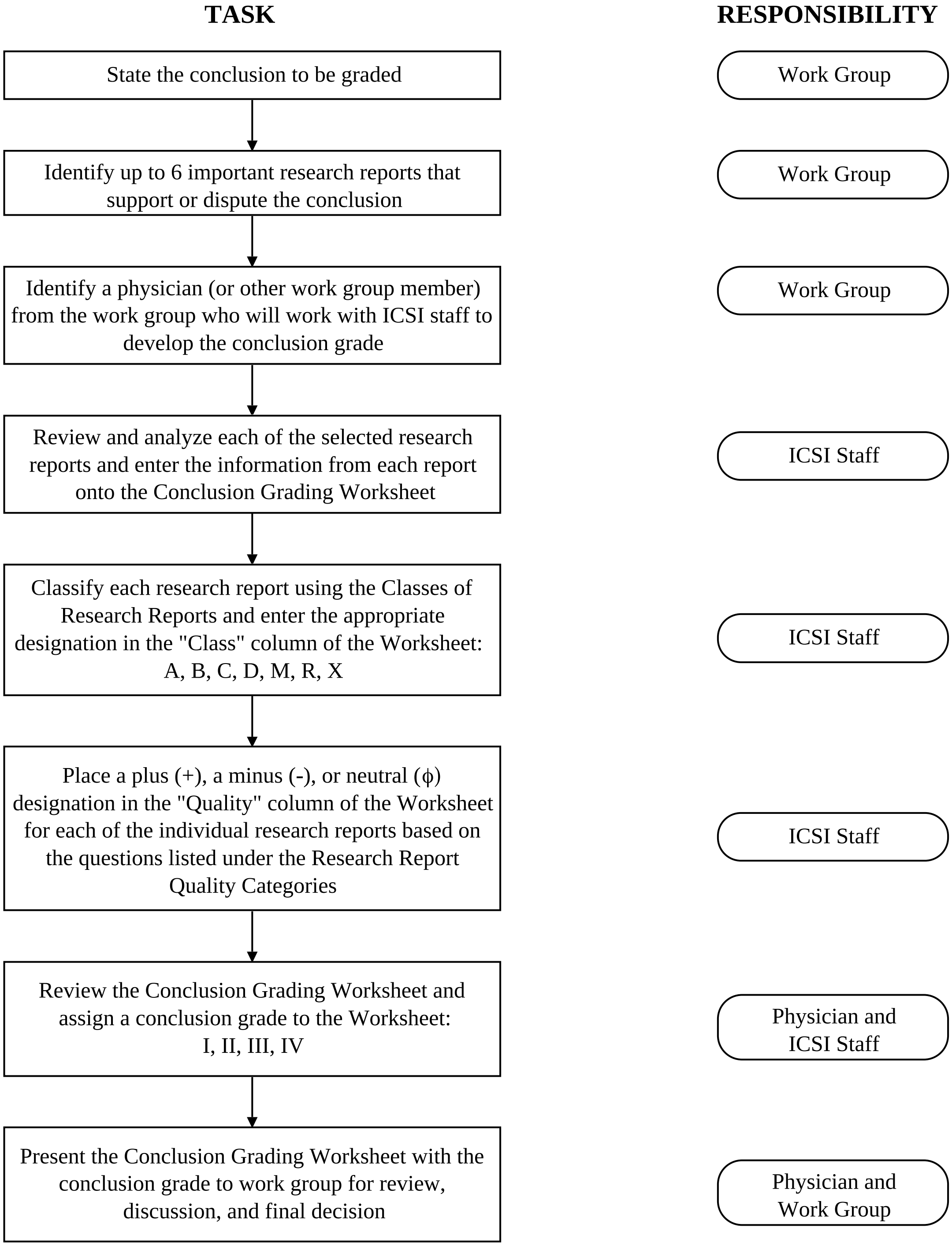
TASK
RESPONSIBILITY
State the conclusion to be graded
Work Group
Identify up to 6 important research reports that support or dispute the conclusion
Work Group
Identify a physician (or other work group member) from the work group who will work with ICSI staff to develop the conclusion grade
Work Group
Review and analyze each of the selected research reports and enter the information from each report onto the Conclusion Grading Worksheet
ICSI Staff
Classify each research report using the Classes of Research Reports and enter the appropriate designation in the "Class" column of the Worksheet: A, B, C, D, M, R, X
ICSI Staff
Place a plus (+), a minus (-), or neutral (ϕ) designation in the "Quality" column of the Worksheet for each of the individual research reports based on the questions listed under the Research Report Quality Categories
ICSI Staff
Review the Conclusion Grading Worksheet and assign a conclusion grade to the Worksheet: I, II, III, IV
Physician and ICSI Staff
Present the Conclusion Grading Worksheet with the conclusion grade to work group for review, discussion, and final decision
Physician and Work Group

Figure 2. This figure represents the process for reaching a conclusion grade. The left column states the task to be completed and the right column identifies who is responsible for completion of that task.

Descriptions of the "Classes of Research Reports" and the "Conclusion Grades" are included in each guideline and technology assessment report. The class of research report assigned to an individual article is presented at the end of the bibliographic citation for that article. The conclusion grades are incorporated into the text of the guideline or technology assessment report with a reference to the Appendix containing the conclusion grading worksheet. Therefore, the reader of the document is able to use the conclusion grading information in weighing the strength of the evidence supporting the conclusion statement. This knowledge should ultimately assist the physician in making decisions about patient care.

Guidelines and Technology Assessment reports both undergo a critical review process in which ICSI member medical groups have an opportunity to submit written critiques of the documents while still in draft form. It is expected that any critical evidence overlooked by the work group in their search of the literature would be identified during the review phase.

References

1. Braunwald E, Mark DB, Jones RH, et al. Unstable angina: diagnosis and management. Clinical Practice Guideline Number 10. Agency for HealthCare Policy and Research (AHCPR), March 1994.
2. U.S. Preventive Services Task Force. Guide to clinical preventive services (2nd Edition). Baltimore: Williams and Wilkins, 1996.
3. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 95(Suppl):2S-4S, 1989.
4. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 102(Suppl):305S-311S, 1992.
5. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ. Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* 104(Suppl):227S-230S, 1995.
6. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. IX. A method for grading health care recommendations. *JAMA* 274:1800-1804, 1995.
7. Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 49:749-754, 1996.
8. Riegelman RK, Hirsch RP. Studying a study and testing a test. Boston: Little, Brown & Co. 1996.
9. Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown & Co. 1987.
10. Friedman GD. Primer of epidemiology, 4th ed. St. Louis: McGraw-Hill, Inc. 1994.