Grading the Evidence for Diabetes Performance Measures

CONTEXT. Grading scientific evidence is a critical step in developing practice guidelines and quality performance measures.

GENERAL QUESTION. What is the most useful way to grade evidence?

SPECIFIC QUESTION. How should we grade the recommended clinical practices for patients with diabetes?

STANDARD APPROACH. Hierarchical grading systems (e.g., grades I, II, and III), such as that used by the U.S. Preventive Services Task Force, have traditionally been used to rank the research designs of studies that support a particular clinical practice.

POTENTIAL DIFFICULTIES. Many studies that support the clinical practices of diabetes care do not clearly conform to the categories traditionally used in hierarchical grading systems. As a result, there is a tendency to inaccurately characterize the level of evidence, leading to the phenomenon of evidence inflation or evidence deflation. To avoid exaggerating the evidence, important sources of information may be excluded, resulting in an understatement of the available supporting evidence.

ALTERNATE APPROACH. This paper offers a more descriptive typologic system that uses the study design and an explanatory modifier to grade the evidence supporting the clinical practices of diabetes care. The study grades are randomized, controlled trial (RCT); RCT–embedded component; RCT–treatment only; RCT–different population; observational study–risk factor; and expert opinion. Using this grading system, the authors were able to more accurately describe the best available evidence supporting the clinical practices of diabetes care.

There is increasing interest nationwide in using evidence-based medicine, practice guidelines, and quality performance measures derived from guidelines to make the the practice of medicine more effective. As a result, several methods and typologic systems for grading the evidence supporting various aspects of clinical practice have evolved.[1–4] In this paper, we use our experience in grading evidence for the clinical practices of diabetes care to delineate the limitations of common hierarchical methods for structuring evidence. We then suggest and demonstrate an alternate descriptive typologic system.

Background

Standard Approach to Grading Evidence

The standard approaches to grading evidence usually use some form of hierarchical classification system to describe and rank the research designs of studies that support a particular clinical practice. In 1979, the Canadian Task Force on the Periodic Health Examination[1] began the first comprehensive effort to classify evidence. The goal of the task force was to address uncertainties about the value of clinical services

The abstract of this paper is available at ecp.acponline.org.

UNANSWERED QUESTIONS: RESEARCH UNDER CONSTRUCTION

SHEILA H. ROMAN, MD, MPH
Department of Medicine
Albert Einstein College of Medicine
New York, NY

STACEY B. SILBERZWEIG, MS, RD
Department of Health Policy

ALBERT L. SIU, MD, MSPH
Department of Medicine and Health Policy

Mount Sinai School of Medicine
New York, NY

Eff Clin Pract. 2000;3:85-91.

See related editorial on pages 94–95.

TABLE 1

Standard Hierarchical System for Grading Evidence*

GRADE	EXPLANATION
I	Evidence obtained from at least one properly designed, randomized, controlled trial
II-1	Evidence obtained from well-designed controlled trials without randomization
II-2	Evidence obtained from well-designed cohort or case–control analytic studies, preferably from more than one center or research group
II-3	Evidence obtained from multiple time series with or without the intervention; dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence
III	Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

*Adapted from the U.S. Preventive Services Task Force.[2]

offered during routine examination of asymptomatic persons. Other organizations, including the U.S. Preventive Services Task Force, have since adopted variations on this initial classification system.[2–4] **Table 1** outlines the standard hierarchical ranking system.

Our Experience Using the Standard Approach in Diabetes

An interdisciplinary steering committee was convened and used modified Delphi techniques to select clinical practices of care, which were included in a set of guidelines and performance measures for patients with diabetes in the ambulatory care setting. MEDLINE searches were performed by using combinations of the terms *diabetes, diabetes outcome*, and *outcome* and the relevant clinical practice and outcome of interest. To find additional articles, we reviewed the reference sections of both previously identified sources and the articles obtained by searching MEDLINE. Studies were assessed for methodologic thresholds (number of patients, characterization of the patient population, definition of interventions, assessments of outcomes, and methodologic rigor), and evidence was graded by using the traditional hierarchical classification scheme. **Table 2** shows the results of this process for common performance measures used to evaluate the quality of diabetes care.

Much of the available evidence did not strictly support the performance of these clinical practices or clearly fit into the categories used in hierarchical grading systems. Using traditional systems, we classified these diabetes quality performance measures as grade III (which indicates consensus of the expert committee that developed them) to avoid overstating the evidence. However, this classification excluded important sources

TABLE 2

Grading of Selected Diabetes Performance Measures by Using the Standard Hierarchical System*

MEASURE	HIERARCHICAL GRADE	WHY THE EVIDENCE IS LIMITED
Hemoglobin A_{1c} testing	III	No RCT has demonstrated that hemoglobin A_{1c} testing in and of itself improves long-term clinical outcomes
Screening for proteinuria	III	No RCT has demonstrated the benefit of screening
Treating hyperlipidemia	III	Patients with diabetes were excluded or were a small subset of the patients enrolled in the lipid-lowering trials
Weight and exercise counseling	III	Modification of risk factors (e.g., obesity and sedentary lifestyle) has only demonstrated improvement in biological markers and not clinical outcomes

*RCT = randomized, controlled trial.

of information and actually understated the existing evidence. **Table 3** summarizes the evidence and the four grading dilemmas that arise when traditional grading systems are used to evaluate diabetes practices.

Four Grading Dilemmas

Dilemma 1: Inferences about One Component Based on Evidence from a Multicomponent Intervention

For various reasons, some randomized trials investigate clinical practices that have several components. If researchers are interested in only one of the components, summarizing the available evidence can be difficult. For example, regular monitoring of hemoglobin A_{1c} levels in patients with diabetes is supported by the results of the landmark Diabetes Control and Complications Trial (DCCT),[5] a large, well-constructed, randomized, controlled trial (RCT) that definitively established a cause-and-effect relation between glycemic control and microvascular outcomes in patients with type 1 diabetes mellitus. However, in that trial, hemoglobin A_{1c} testing was only one component of a multicomponent, multidisciplinary, intensive diabetes management program. This program decreased the development and progression of diabetes-related microvascular complications that lead to blindness, kidney failure, and neuropathic disease. Many of the recommended procedures included in that system of care—such as regular monitoring of hemoglobin A_{1c} levels; regular screening for microvascular complications; and patient interactions for self-management education, nutritional counseling, and behavioral counseling—are most strongly supported by evidence from the DCCT.

However, in the strictest of interpretations, DCCT evidence can be considered to support these care processes only when they are delivered within the context of the same multicomponent intensive management program that the DCCT itself used. We could therefore conclude that the DCCT does not provide evidence for regular hemoglobin A_{1c} testing in “real-world” clinical settings. Consequently, we could rely on less robust clinical trials or lower levels of evidence to support this practice. We could also use the DCCT and other studies with structured intensive management programs[6–8] to support regular monitoring of hemoglobin A_{1c} levels if we simultaneously acknowledge that the evidence may be overstated because it does not involve hemoglobin A_{1c} testing only.

Dilemma 2: Inferences about the Value of Early Diagnosis Based on Evidence for Early Treatment

As previously discussed by Mulrow and colleagues,[9] much evidence about health care is indirect and requires that diagnostic strategies and other clinical practices or interventions be linked with outcomes of care. This is

TABLE 3

Supplemental Evidence for Diabetes Performance Measures and Four Associated Grading Dilemmas*

MEASURE	SUPPLEMENTAL EVIDENCE	GRADING DILEMMA
Hemoglobin A_{1c} testing	The DCCT, UKPDS, and other RCTs show that multicomponent intensive therapy protocols (which included hemoglobin A_{1c} testing) were associated with lower risk for microvascular complications of diabetes	Inference about one component based on evidence from a multicomponent intervention
Screening for proteinuria	RCTs document the benefit of treatment of proteinuria with ACE inhibitors to delay progression from microproteinuria to macroproteinuria and renal failure	Inference about the value of early diagnosis based on evidence for early treatment
Treating hyperlipidemia	Multiple RCTs have documented that treating hyperlipidemia decreases cardiac end points and mortality rates in nondiabetic patients	Inference based on evidence from a different population
Weight and exercise counseling	Multiple risk factors have been defined for type 2 diabetes mellitus (e.g., obesity and sedentary lifestyle)	Inference based on risk factor identification

**ACE = angiotensin-converting enzyme inhibitors; DCCT = Diabetes Control and Complications Trial; RCT = randomized, controlled trial; UKPDS = United Kingdom Prospective Diabetes Study.*

TABLE 4
Proposed Descriptive Evidence Grading System*

GRADE	EXPLANATION
RCT	Evidence from randomized, controlled trial
RCT–embedded component	One component of a multicomponent intervention
RCT–treatment only	Indirectly supported by evidence from a treatment intervention
RCT–different population	Evidence largely derived from a different patient population
Observational study–risk factor	Evidence for modification based on the association between risk factor and bad outcome
Expert opinion	Evidence based on expert opinion, consensus panel, or clinical experience

**RCT = randomized, controlled trial.*

certainly the case for many of the clinical practices recommended for diabetes, which are almost always separated from the ultimate clinical end points of interest by a chain of clinical events and intermediate outcomes. Furthermore, the level of evidence along the chain of clinical events may vary or may be nonexistent at certain points. Clinical practices that precede the time frame of a given intervention in a chain of clinical events may not be considered part of the randomized trial evaluating the intervention. However, these clinical practices may be of great interest to persons who are developing practice guidelines or quality performance measures.

For example, consider the recommendations for nephropathy screening. Randomized trials of patients with diabetes mellitus and proteinuria have shown that treatment with angiotensin-converting enzyme inhibitors effectively decreases progression to renal failure.[10–12] However, no randomized trials have been done to determine when or at what intervals patients with diabetes should be tested for proteinuria or microalbuminuria. Clearly, in this example, the scientific evidence supporting specific interval screening for proteinuria is not as strong as evidence from randomized trials. However, stating that the evidence for routine testing is nonexistent ignores the randomized trial evidence that supports treatment on the basis of test results. Ignoring this evidence would result in undervaluing the screening procedure and understating its potential to improve clinical care and outcomes.

Dilemma 3: Inferences Based on Evidence from a Different Population

Existing grading systems do not address the ways in which evidence should be used when the only or best available evidence is from a similar patient population, not from the population of interest. Most frequently, the population of interest is a subgroup in a randomized trial or participants who have been excluded from randomized trials or other studies conducted on a broader group of diagnoses or patients. If information from the larger trial is not available or is inadequate to allow evaluation of the efficacy of the therapy in the subgroup of patients, how should the evidence supporting treatment for the subgroup be rated?

An example of this phenomenon is the literature on cholesterol-lowering agents and prevention of cardiovascular end points. Patients with diabetes were initially excluded from these studies and represented only a small part of a much larger study sample in subsequent studies. It was not until publication of the Scandinavian Simvastatin Survival Study (4S)[13] that efficacy in diabetes treatment could be examined in a subgroup analysis. However, many experts consider subgroup analyses suspect and recommend that they be avoided.[14] Again, although it is inaccurate to generalize the rating from the strongest evidence (in this case the much larger RCT from which the subgroup was derived), stating that the evidence is nonexistent ignores the randomized trial, the subgroup analysis, and other evidence. For patients with diabetes, in whom atherosclerotic cardiovascular disease is associated with high morbidity and mortality, it would be counterintuitive to ignore the findings regarding lipid lowering in nondiabetic populations or subgroup analyses.

Dilemma 4: Inferences Based on Risk Factor Identification

Available methods used to grade evidence allow for consideration of case–control and nonrandomized cohort study designs. These designs have occasionally been used to evaluate the effect of clinical practices but are more often used to identify risk factors (with or without modification). It is not easy to summarize the latter type of evidence with existing systems for grading evidence. Some would argue that identification of a risk factor

TABLE 5

Grading of Selected Diabetes Performance Measures by Using the Proposed Descriptive System*

MEASURE	PROPOSED GRADE	EXPLANATION
Hemoglobin A_{1c} testing	RCT–embedded component	One component of a multicomponent intervention
Screening for proteinuria	RCT–treatment only	Indirectly supported by evidence from a treatment intervention
Treating hyperlipidemia	RCT–different population	Evidence largely derived from patients without diabetes
Weight and exercise counseling	Observational study–risk factor	Evidence largely derived from observations about the associations between obesity and sedentary lifestyles and diabetes

**RCT = randomized, controlled trial.*

provides little evidence on the ability of clinical practices to modify risk and that such evidence should perhaps not be graded as acceptable.

Multiple risk factors have been defined for type 2 diabetes mellitus and for the development and progression of associated microvascular and macrovascular outcomes. Modification of some risk factors (e.g., glycemic control) has been shown to improve clinical outcomes (e.g., microvascular). For other risk factors (e.g., obesity and sedentary lifestyle), however, the effect of modification (e.g., weight loss and exercise) on clinical outcomes remains heterogeneous and less certain. However, it is probably unreasonable to completely discount such evidence.

Alternate Approach: A Descriptive System for Evidence Grading

Table 4 shows our proposal to address these four grading dilemmas. Instead of the traditional hierarchical grading system, we propose using a descriptive typologic system that consists of the study design and an explanatory modifier.

RCT–Embedded Component

For clinical practices that are one component in a previously studied program of care, we propose adding the modifier "embedded component" to the study design that describes the larger work. This modifier identifies the clinical practice of interest as a component of a larger intervention and allows more accurate description of the evidence. The example of regular monitoring of hemoglobin A_{1c} levels to reduce the risk for microvascular complications would therefore be graded as "RCT–embedded component." This grade indicates that this clinical practice was one component of a multicomponent, structured study intervention, such as the DCCT[5] or the United Kingdom Prospective Diabetes Study.[6] It also indicates that the available evidence supports the monitoring of hemoglobin A_{1c} levels in the context of the patient education and intensive medical management provided in the larger study. We believe that this grade more accurately describes the quality of the evidence that supports monitoring hemoglobin A_{1c} levels.

RCT–Treatment Only

We propose assigning a descriptive modifier of "treatment only" to clinical practices preceding a treatment intervention that has been the subject of clinical studies. For example, we would assign a grade of "RCT–treatment only" to indicate that testing for proteinuria in patients with diabetes is indirectly supported by randomized trials evaluating the efficacy of angiotensin-converting enzyme inhibitors in patients with diabetic proteinuria.[10–12]

RCT–Different Population

To identify evidence that is generalized from studies in a related patient population, we propose using the modifier "different population." Most of the evidence for the efficacy of controlling lipid levels is from patients without diabetes; therefore, it would be characterized as "RCT–different population" for patients with diabetes. This modifier allows more accurate representation of the specificity of the evidence.

Observational Study–Risk Factor

If the best available evidence identifies modifiable risk factors and the effect of risk factor modification is not established, we propose using the modifier "risk factor" to indicate evidence that is based on studies of the risk factor. Clinical practices relating to weight control or exercise in patients with diabetes might be graded as "observational study–risk factor."

Synthesis

In **Table 5**, the evidence for selected diabetes performance measures is evaluated by using the descriptive evidence grading system. When **Tables 2** and **5** are compared, it is evident that use of a descriptive rather than hierarchical typologic system obviates the grading dilemmas we have discussed, allows better representation of the available evidence, and more accurately characterizes the level of evidence.

Discussion

Three factors have produced a surge of interest in measuring the processes and outcomes of diabetes care: 1) studies documenting the heavy financial burden of diabetes and its related complications on the U.S. health care system[5]; 2) the increasing incidence and prevalence of type 2 diabetes mellitus in the United States (6% of the general population and up to 20% of those 65 years of age or older), particularly in minority populations[16]; and 3) the documentation of underuse of diabetes-related processes of care in varied clinical settings (e.g., fee for service and managed care, academic and nonacademic) across the United States.[17–23]

As a result of these factors, a plethora of guidelines and measures of performance for diabetes care have been developed at both local and national levels. Inherent in all such efforts is an attempt to grade the evidence supporting the recommended practices. As we have demonstrated, the traditional hierarchical approach to grading does not adequately characterize the available evidence. We believe that a descriptive typologic system is more accurate and more flexible. Such a system is less likely to understate or overstate existing evidence and may help identify areas that require increased attention in the development and construction of valid and reliable performance measures.

One limitation of our grading system is that it does not directly address the quality of the actual studies under evaluation; for example, a guideline may be based on a randomized trial, but the trial may be of poor quality. Of course, this is also a limitation of the traditional hierarchical grading system. In the future, researchers could use discrete thresholds to assess inclusion and exclusion criteria, adequacy of numbers of participants, characterization of the patient population, definition of the interventions, rigorousness of the study design, and outcomes.

Although we have used diabetes quality performance measures as an example, we believe that the grading dilemmas we encountered apply to many clinical practices. A recent structured review of guidelines published in the peer-reviewed medical literature found a lack of adherence to established methodologic standards, especially in the identification, evaluation, and synthesis of scientific evidence.[24] The grading dilemmas we have described may have contributed to this lack of synthesis. Finally, because initiatives that have embraced existing hierarchical grading methods have sometimes focused on narrow, well-circumscribed problems or interventions based on RCTs, it could be argued that hierarchical grading systems have limited the influence of evidence-based medicine.

We believe that a descriptive typologic system for grading evidence can 1) allow specific clinical practices to be supported by more scientific evidence; 2) allow more flexibility in assigning value to supportive evidence from related clinical topics or patient populations; and 3) recognize the role of established modifiable risk factors in informing clinical practice even when evidence on the effectiveness of their modification has yet to be established. We believe that the use of a descriptive typologic system for grading evidence can evaluate a study's contributions beyond its basic design and will better inform clinicians about the evidence supporting clinical practices. We also think that this type of grading system will offer a superior platform from which to develop guidelines and derive performance measures. Finally, we feel that the use of descriptive modifiers in grading evidence will increase clinicians' ability to interpret best evidence and will broaden the influence of evidence-based medicine.

Take-Home Points

- **Grading scientific evidence is a critical step in developing practice guidelines and quality performance measures.**
- **The standard approach—classifying studies as grade I, II, or III—focuses on study design and excludes important sources of information.**
- **Many diabetes practices have not been directly studied in randomized, controlled trials.**
- **The evidence underlying diabetes guidelines often comes from randomized trials but is indirect (e.g., patients randomly assigned to receive a multicomponent intervention instead of hemoglobin A_{1c} monitoring only).**
- **To accommodate these subtleties, we offer a more descriptive typologic system for grading evidence.**

References

1. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979;121: 1193-254.
2. U.S. Preventive Services Task Force. Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force. 2d ed. Baltimore: Williams & Wilkins; 1996.
3. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1992;102:305S-11S.
4. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA. 1995;274:1800-4.
5. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. N Engl J Med. 1993;329: 977-86.
6. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet. 1998;325: 837-53.
7. Ohkubo Y, Kishikawa H, Araki E, et al. Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with non-insulin-dependent diabetes mellitus: a randomized prospective 6-year study. Diabetes Res Clin Pract. 1995;28:103-17.
8. Reichard P, Nilsson BY, Rosenqvist U. The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. N Engl J Med. 1993; 329:304-9.
9. Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. Ann Intern Med. 1997;127:989-95.
10. Lewis EJ, Hunsicker LG, Bain RP, Rohde RD. The effect of angiotensin-converting-enzyme inhibition on diabetes nephropathy. The Collaborative Study Group. N Engl J Med. 1993; 329:1456-62.
11. Mogensen CE, Christensen CK. Predicting diabetic nephropathy in insulin-dependent patients. N Engl J Med. 1984; 311:89-93.
12. Ravid M, Lang R, Rachmani R, Lishner M. Long-term renoprotective effect of angiotensin-converting enzyme inhibition in non-insulin-dependent diabetes mellitus. A 7-year follow-up study. Arch Intern Med. 1996;156:286-9.
13. Pyorala K, Pedersen TR, Kjekshus J, Faergeman O, Olsson AG, Thorgeirsson G. Cholesterol lowering with simvastatin improves prognosis of diabetic patients with coronary heart disease. A subgroup analysis of the Scandinavian Simvastatin Survival Study (4S). Diabetes Care. 1997;20:614-20.
14. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med. 1987;317:426-32.
15. Rubin RJ, Altman WM, Mendelson DN. Health care expenditures for people with diabetes mellitus, 1992. J Clin Endocrinol Metab. 1994;78:809A-809F.
16. National Diabetes Data Group. Diabetes in America. 2d ed. Bethesda, MD: National Institutes of Health; 1995;47-68;85-116.
17. Ho M, Marger M, Beart J, Yip I, Shekelle P. Is the quality of diabetes care better in a diabetes clinic or in a general medicine clinic? Diabetes Care. 1997;20:472-5.
18. Peters AL, Legorreta AP, Ossorio RC, Davidson MB. Quality of outpatient care provided to diabetic patients. A health maintenance organization experience. Diabetes Care. 1996;19:601-6.
19. Wisdom K, Fryzek JP, Havstad SL, Anderson RM, Dreiling MC, Tilley BC. Comparison of laboratory test frequency and test results between African-Americans and Caucasians with diabetes: opportunity for improvement. Findings from a large urban health maintenance organization. Diabetes Care. 1997;20:971-7.
20. Weiner JP, Parente ST, Garnick DW, Fowles J, Lawthers AG, Palmer RH. Variation in office-based quality. A claims-based profile of care provided to Medicare patients with diabetes. JAMA. 1995;273:1503-8.
21. Marshall CL, Bluestein M, Chapin C, et al. Outpatient management of diabetes mellitus in five Arizona Medicare managed care plans. Am J Med Qual. 1996;11:87-93.
22. Rubin RJ, Dietrich KA, Hawk AD. Clinical and economic impact of implementing a comprehensive diabetes management program in managed care. J Clin Endocrinol Metab. 1998;83: 2635-42.
23. Greenfield S, Rogers W, Mangotich M, Carney MF, Tarlov AR. Outcomes of patients with hypertension and non-insulin dependent diabetes mellitus treated by different systems and specialties. Results from the medical outcomes study. JAMA. 1995;274:1436-44.
24. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA. 1999;281:1900-5.

Acknowledgments

The authors thank Roslyn King and Althea Evans for their technical support in the preparation of this manuscript.

Grant Support

Supported in part by grants from the American Diabetes Association and the New York State Department of Health.

Correspondence

Sheila H. Roman, MD, MPH, Medical Director, Office of Outcomes Measurement and Research, Beth Israel Medical Center, First Avenue at 16th Street, New York, New York 10003; telephone: 212-420-3052; fax: 212-420-4708; e-mail: sroman@bethisraelny.org.