Adherence to Prescribed Explicit Criteria During Utilization Review

An Analysis of Communications Between Attending and Reviewing Physicians

Lawrence C. Kleinman, MD, MPH; Elizabeth A. Boyd, MA; John C. Heritage, PhD

Context.—Utilization review (UR) seeks to improve quality and cost-efficiency of health care. However, how well the process works in practice has not been assessed.

Objective.—To describe the outcomes of a sample of physician reviews in terms of the explicit criteria that the UR was designed to implement.

Design.—Retrospective analysis of transcripts of precertification reviews.

Participants and Setting.—California physicians employed by a UR firm conducted 96 interviews from April 1990 to July 1991 with attending physicians who had proposed to insert tympanostomy tubes on a patient younger than 16 years and whose proposals had been found to be inappropriate on an initial screen.

Main Outcome Measures.—The appropriateness rating assigned to each case by the physician-reviewer and by the investigators using explicit criteria. Logistic regression identified factors associated with the reviewers' recommendations to perform surgery and with recommendations at variance from the criteria.

Results.—The reviewers recommended 78% of cases for surgery, of which only 29% were supported by the criteria or had extenuating circumstances. The criteria concurred with all 30 of the reviewers' recommendations against surgery. Two factors, female sex (odds ratio [OR], 8.2; 95% confidence interval [CI], 1.2-53.8) and previous tympanostomy tube insertion (OR, 30.9; 95% CI, 2.4-394.8) were associated with reviewer recommendations in favor of surgery that were at variance from the criteria, despite the lack of evidence for either as a mitigating circumstance.

Conclusion.—Physician reviewers were more lenient than the explicit criteria that the reviews were designed to implement. In no cases did the reviewers depart from the criteria's recommendations in favor of surgery.

JAMA. 1997;278:497-501

IMPROVING the quality of medical care represents a formidable challenge. Managed care organizations (MCOs), hospitals, and other health care organizations seek to hold clinicians accountable to guidelines, practice parameters, and explicit or implicit criteria in the name of improving quality or increasing the cost-efficiency of care. Physicians are understandably skeptical and concerned that the motivation for such interventions may be to limit medical costs regardless of the effect on quality. Consequently, implementation of guidelines and criteria remains problematic.[1-4]

Utilization review (UR) is one technique used to implement guidelines. One type of UR is the prospective precertification of proposed medical interventions.[5,6] A recent article suggests that the anticipation of being held accountable by prospective UR decreases the use of medical services in a nonspecific manner.[7] While physician accountability may be a worthy goal, the UR systems that affect clinician behavior must also be accountable. Currently, the actual process of individual URs and the factors that contribute to their outcomes remain obscure to most clinicians.[8] To explore this issue we followed up a recent study to evaluate the medical appropriateness of tympanostomy tubes for children with otitis media.[9,10] This allowed us to investigate an example of precertification UR as it occurred in practice and to describe the outcomes of reviews in terms of the explicit criteria on which they were based. We specifically wanted to answer the following questions:

1. Do physician-reviewers categorize the appropriateness of a proposed procedure according to explicit criteria? We hypothesized that, in general, they did.
2. Can we identify specific factors in the reviews that are associated with cases for which the physician-reviewer's recommendations vary from the criteria? We hypothesized that being male and having a history of tympanostomy tubes could be associated with reviewers' recommending surgery at variance from the criteria.

METHODS

Derivation of Explicit Criteria

First, a series of mutually exclusive, clinically detailed scenarios, termed indications, was developed. These *indications* incorporated such factors as the child's age, as well as clinical factors, such as findings on physical examination or duration of symptoms. The appropriateness of each potential indication was

From the Departments of Pediatrics (Dr Kleinman) and Sociology (Ms Boyd and Dr Heritage), University of California, Los Angeles; and the Department of Pediatrics, Harvard Medical School, Boston, Mass (Dr Kleinman).

Dr Kleinman served as a physician-reviewer for Value Health Sciences, Inc, Santa Monica, Calif, prior to the study. He served as vice president from December 1994 to June 1996 and owns stock in the company.

The views and opinions contained herein are the authors' and not necessarily those of the Ambulatory Pediatric Association or the Robert Wood Johnson Foundation.

Reprints: Lawrence C. Kleinman, MD, MPH, 306 Elliot St, Newton Upper Falls, MA 02164 (e-mail: Kleinman@HSPH.Harvard.edu).

then assessed by an expert panel using a 2-round modified Delphi method. The panel rated each indication as inappropriate, equivocal (including uncertain), or appropriate by using a 9-point ordinal scale (1-3 inappropriate, 4-6 equivocal, 7-9 appropriate). Appropriate tympanostomies were defined as those for which the expected health benefits of the surgery exceeded the expected health risks by a sufficient margin to make the procedure worth doing. In an attempt to increase the burden necessary to label an indication inappropriate, the rating of inappropriate was restricted to those indications for which both a majority of the panelists ranked the indication as inappropriate, and 8 of the 9 panelists agreed that the indication was not appropriate. Two or more panelists rating an indication appropriate in the face of 2 or more rating it inappropriate was considered to represent panelist disagreement and led to a rating of equivocal, regardless of the median rating. The criteria have been published previously.[9]

Application of Criteria by UR Firm

These criteria were adopted by a national health services firm, Value Health Sciences, Inc, Santa Monica, Calif (hereinafter called the "UR firm"), as part of a system for conducting prospective audits of the appropriateness of interventions that are proposed for patients. This system has been licensed to a number of clients—third-party payers and MCOs—that use it for prospective UR. As a rule, the UR firm recommends in favor of surgery for all cases found either equivocal or appropriate. Some clients use the recommendations to make decisions about reimbursement, while others do not link them to any financial sanction.

The UR firm implemented this method using a 2-step process that has been described previously.[9] At step 1, nurse-reviewers telephoned the attending physicians' offices to obtain clinical data. This interview was guided by a computer algorithm designed to obtain sufficient data to assess medical appropriateness according to the criteria. Once the computer determined that a case had at least equivocal appropriateness, the interview was terminated and the case recommended for surgery. Cases identified as potentially inappropriate at step 1 were then subjected to the step 2 physician reviews that are the subject of this article. In the step 2 review a physician reviewed the data from the step 1 review and then conducted a telephone interview of the attending physician, usually the otolaryngologist who recommended the surgery. All step 2 reviewers for the UR firm are board-certified physicians and licensed to practice in California. Both pediatricians and otolaryngologists reviewed pediatric otolaryngeal procedures preferentially. The purpose of the step 2 review was to determine if there was additional clinical information that would increase the level of appropriateness and/or extenuate in favor of surgery for the case under review. When the otolaryngologist indicated that his or her office had insufficient information, the reviewers also spoke to the attending primary care physician. The physician-reviewers were asked to recommend whether or not surgery was indicated on the basis of the clinical data and the explicit criteria.

Sample

Because of logistical considerations related to retrieving archived step 2 reviews, we limited this study to those cases reviewed between April 1, 1990, and July 31, 1991. There were 5214 cases reviewed during this period for 3 large clients, representing both fee-for-service and prepaid health plans and insuring a diverse population of over 5.6 million people in the United States at the time. All 3 applied financial sanctions (ie, denied reimbursement) against cases found inappropriate. A total of 1448 (28%) were screened as potentially inappropriate by the step 1 review. Of these, the UR firm conducted the step 2 physician review on 942 cases (65%), and these make up the sampling frame for the current study. The remainder of the step 2 reviews were conducted by the clients directly and were not preselected based on any criteria; the data from these client reviews were not available to the investigators.

We chose a probability sample of cases found inappropriate at step 1, oversampling low-volume reviewers and cases for which the reviewers recommended against surgery. We sought a total sample of 100 cases in the ratio of 2 cases recommended for surgery to each case recommended against, sufficient that a finding of no cases incorrectly labeled inappropriate would imply with more than 95% confidence that the actual rate was less than 10%. We used the sampling ratios to construct weights that allow us to present data representative of the sampling frame. Four cases had to be removed from the analysis because of incomplete data, leaving a final sample of 96 step 2 reviews. This sample included 66 cases where the reviewer overturned the step 1 finding of inappropriate (recommended in favor of surgery), and 30 cases where the step 1 finding was sustained. This sample is sufficient so that a zero result would imply an actual rate of less than 10% with 93% confidence.

Data Collection

With the knowledge of both parties, the step 2 reviews were routinely audiotaped by the UR firm. We transcribed these tapes verbatim, and one of us (E.A.B.) abstracted all clinical information from the step 1 review, the transcript, and the reviewer's summary of the case. The abstraction coded the level of documentation of each clinical finding and noted contradictions within the data. As parental reports are not always reliable, the criteria explicitly required episodes of otitis media or the presence of an effusion to be documented,[11] either by firsthand knowledge of the proposing physician or secondhand knowledge from another clinician. Unless verified by a clinician, a patient's (or parent's) statement was considered undocumented information, and this was consistent with the training that the reviewers received. We calculated the κ statistic to assess the validity of the abstraction method compared with a detailed content analysis. The κ was 0.93, indicating excellent agreement.[12]

These abstractions were then reviewed by the first author who was blind to the outcome of the specific review and unaware of the exact proportion of reviews in the sample that were ultimately overturned. This investigator assigned each case to an indication based only on the documented clinical data. We assigned appropriateness ratings to each indication based on the criteria.[9]

We considered the criteria to recommend in favor of surgery if the appropriateness rating was equivocal (median rating 4, 5, or 6, or panelist disagreement) or appropriate (median 7, 8, or 9), or if extenuating clinical circumstances were present (in which case we assigned an appropriateness rating of 9). Extenuating circumstances were clinical situations that the literature suggests make a child more likely to benefit from tympanostomy tube surgery. Examples include craniofacial abnormalities (such as an unrepaired cleft palate), cholesteatoma, Down syndrome, and certain ethnicities, such as some Eskimo children. We also considered children with repaired cleft palates and with cleft lips to have extenuating circumstances, even though this is not well supported by the literature.

For the purposes of the current study we abstracted the following data in addition to that collected from the first abstraction: previous history of tympanostomy tube surgery; comments suggesting a positive outcome from the previous surgery; and the difference (in days) between the documented duration of effusion and the longest duration of

Table 1.—Review Outcome by Appropriateness*

	Criteria					
	Inappropriate		Equivocal		Appropriate	
Reviewer	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
Not recommended	30	21	0	0	0	0
Recommended	47	54	13	14	6†	7†

*The category of appropriateness is based on the findings of an expert panel. Cases with extenuating circumstances were considered appropriate. Findings were adjusted on the basis of sampling weights to be representative of the sampling frame.

†Of these, 4 (raw) and 5 (adjusted) had extenuating clinical circumstances (see text).

effusion asserted during the review. We took the decision of the expert panel not to consider a history of previous tympanostomy tubes to be a specific decision that such a history was not an extenuating circumstance. The literature regarding the natural history of otitis media supports this point of view.[13,14] Nonetheless, we conducted an additional analysis that considered either documentation or assertion by the reviewed physician of a positive outcome from previous tympanostomy tubes to constitute an extenuating circumstance.

Data Analysis

We assessed agreement between the recommendations of the reviewers and the criteria through the use of contingency tables, calculating the κ statistic to quantify the level of agreement beyond chance. We used McNemar χ to assess whether the reviewers and the criteria were similar in their level of discrimination when identifying care as inappropriate. A large value for χ suggests that the 2 processes classify in a disparate manner.[12] We used the SAS PROC T TEST to compare the length of documented effusion and the mean appropriateness rating between those cases that were recommended for surgery and those that were not.[15] We also compared the length of effusion between cases characterized by the criteria as appropriate, equivocal, and inappropriate using SAS PROC GLM.[15]

To simplify the results of the regression analysis we desired a single clinical variable that would integrate the severity of acute otitis media and otitis media with effusion. However, no commonly accepted measure exists. For this purpose only, we defined a clinical severity variable as 0.5 times the duration of effusion (in months) if the child had no acute otitis media, the duration of effusion if the child had occasional episodes of acute otitis media, and 1.5 times the duration of effusion if the child had frequent episodes of acute otitis media ($\geq$3 episodes in 6 months, or $\geq$ 4 episodes in 12 months). We ran models using this clinical variable as well as using separate variables for the frequency of recurrent acute otitis media and duration of effusion, individually and jointly. All produced similar results, and we present our findings using the single variable.

We used STATA to perform Huber regressions.[16] Huber regressions were used because they offer the advantage of logistic regression on a binary outcome variable and are able to adjust the standard errors for the lack of independence of cases performed by the same reviewer. For 1 analysis the outcome variable was the reviewer's decision to recommend for or against surgery. For the other analysis the outcome variable was the concordance or discordance of the decision with the criteria. We evaluated 6 variables in the final logistic models: the clinical variable defined above; an indicator variable regarding a history of previous tympanostomy tube insertion; the numerical appropriateness rating; the number of days that the asserted length of effusion exceeded the documented duration of effusion (as a marker for the impact of undocumented information on the reviews); age; and sex.

RESULTS

Reviewer and Sample Characteristics

The sample included 13 reviewers: 3 pediatricians, 3 otolaryngologists, and 7 reviewers from other disciplines. Seventy-nine cases were reviewed by either a pediatrician or an otolaryngologist. The attending physicians were otolaryngologists in 89 cases and primary care physicians in 7. The mean age of the patients was 4.25 years (range, 3 months to 12.75 years), and 60% were boys. There was a history of frequent acute otitis media in 33%. Effusions were present a mean of 41 days (range, 0-305 days); 36% had a history of tympanostomy tube insertion.

Concordance Between Reviewers and Criteria

In 30 cases, the reviewers recommended against surgery; all were considered to be inappropriate by the criteria. In 66 cases the reviewers recommended in favor of surgery. Of these, the criteria termed 6 (9%) appropriate, 4 because of extenuating clinical circumstances that were not sought by the step 1 protocol (2 with Down syndrome, 1 each with a history of previous cholesteatoma and repaired cleft lip) and 2 because new information was identified by the physician review that was sought and not identified by the step 1 review. A further 13 (20%) were termed equivocal by the criteria, and the reviewer's recommendation for surgery was concordant with the UR firm's policy. The remaining 47 cases (71%) were termed inappropriate by the criteria, and thus their recommendation for surgery was at variance from the criteria. Table 1 shows the unadjusted and the weighted results.

The κ statistic that compares a criteria result of inappropriate with a recommendation against surgery by the reviewer is 0.28 (95% confidence interval [CI], 0.20-0.36). The McNemar χ is 6.9, which corresponds to $P<.001$ (2-sided). Thus, we safely can reject the null hypothesis and assert that the reviewers' assessments of appropriateness differed from those of the criteria. The reviewers were more lenient.

Factors and Findings

Cases the reviewers recommended for surgery had more days of documented effusion than those not recommended for surgery (mean [SD] days, 48 [53] vs 23 [24]; $P=.002$). However, the milder the disease, the greater the disparity between the recommendations of the reviewers and the criteria. For example, of 70 children with middle ear effusions documented for less than 60 days, the physician reviewers recommended surgery for 63%, the criteria or extenuating circumstances for only 11%. The criteria recommended against surgery for all 41 children in the least severe category of otitis media, defined as effusions for less than 60 days, who also had fewer than 3 episodes of acute ear infections in the 6 months prior to the review and fewer than 4 acute infections in the previous year, and who had no documented infections while on antibiotic prophylaxis and no extenuating circumstances. The reviewers recommended 56% of these children for surgery.

Although the attending physicians asserted a history of prior tympanostomy tube surgery in 36 cases (37.5%), in only 3 of these cases (8% of cases with prior tubes, 3% of cases overall) did the attending physician assert that the child had a positive outcome from the previous insertion of tubes. All 3 of those cases were recommended for surgery by the reviewers.

Logistic Regressions

The results of the Huber regressions are shown in Tables 2 and 3. In addition to the findings shown in the tables, the Huber models generate odds ratios (ORs) for each of the reviewers compared with the index reviewer (who was randomly

chosen and was the same for both models). In each model, 2 reviewers varied significantly from the index reviewer.

The findings in Table 2 confirm that higher appropriateness ratings are strongly associated with a recommendation for surgery. They also suggest that, controlling for other clinical factors, a history of prior tympanostomy tubes is associated with the reviewers' recommending surgery (OR, 30.94; 95% CI, 2.43-394.81), as is sex (OR, 8.15; 95% CI, 1.23-53.84). Undocumented information appeared to have a mild impact on the outcome of the review (OR for 30 days of undocumented effusion, 1.95; 95% CI, 1.06-3.58). The regression model, with discordance from the criteria as the outcome variable (Table 3), includes 3 predictor variables (age, clinical variable, and appropriateness rating) that have a defined relationship with the criteria and thus with discordance (because the reviewers never recommended against appropriate or equivocal reviews); they are included in the model to control for confounding while evaluating the other predictor variables. Two variables were found to be associated with reviewer recommendations at variance from the criteria: female sex (OR, 11.79; 95% CI, 2.35-59.11) and a history of prior tympanostomy tube surgery (OR, 9.28; 95% CI, 1.32-65.37). There was a suggestion that undocumented information also was related to recommendations that departed from the criteria, but it did not achieve statistical significance (OR for 30 days of undocumented effusion, 1.46; 95% CI, 0.94-2.25). The pseudo R^2 for the first model was 0.49, and for the second, 0.46.

Table 2.—Huber Model Predicting Reviewer Recommendation for Surgery

Variable	Odds Ratio (95% Confidence Interval)	P Value
Age, y	1.03 (0.78-1.36)	.83
Appropriateness rating* (range, 1-9)	4.55 (2.02-10.25)	<.001
Female sex	8.15 (1.23-53.84)	.03
Clinical severity†	1.02 (1.00-1.03)	.01
Excess duration‡	1.95 (1.06-3.58)	.03
History of prior tubes	30.94 (2.43-394.81)	.01

*The criteria held that ratings from 1-3 should not be recommended for surgery; extenuating circumstances were assigned a rating of 9.

†Clinical severity represents an interaction between the duration of effusion and the frequency of acute otitis media.

‡Excess duration represents, in 30-day intervals, the length of time that an effusion was asserted to have been present in excess of the documented duration of the effusion. This variable is included as a marker for the effect of undocumented clinical assertions on the outcome of the review.

Table 3.—Huber Model Predicting Reviewer Recommendations at Variance From the Criteria

Variable	Odds Ratio (95% Confidence Interval)	P Value
Age, y	0.92 (0.74-1.14)	.44*
Appropriateness rating† (range, 1-9)	0.26 (0.13-0.53)	<.001*
Female sex	11.79 (2.35-59.11)	.003
Clinical severity‡	1.02 (1.00-1.03)	.03*
Excess duration§	1.46 (0.94-2.25)	.09
History of prior tubes	9.28 (1.32-65.37)	.03

*These 3 variables have a defined relationship with the criteria and are included in this model only for the purposes of controlling for confounding while evaluating the other 3 variables.

†The criteria held that ratings from 1-3 should not be recommended for surgery; extenuating circumstances were assigned a rating of 9.

‡Clinical severity represents an interaction between the duration of effusion and the frequency of acute otitis media.

§Excess duration represents, in 30-day intervals, the length of time that an effusion was asserted to have been present in excess of the documented duration of the effusion. This variable is included as a marker for the effect of undocumented clinical assertions on the outcome of the review.

COMMENT

This study represents, to our knowledge, the first in-depth analysis of prospective UR.[17] We found that agreement between reviewers precertifying proposals for tympanostomy tube placement and the criteria that the reviews were designed to implement was poor. Although the reviewers never recommended against care supported by the criteria, they were less likely to consider care inappropriate than were the criteria. If we were to consider the goal of these reviews to be the strict implementation of the criteria (including the identification of extenuating circumstances), this would be analogous to a diagnostic test identifying cases that fail to meet explicit criteria. Following through on this analogy, the sensitivity of the step 2 review to identify care that the criteria recommend against would be 0.28 with a specificity of 1.0. The positive predictive value would be 1.0, and the negative predictive value, 0.28.

The finding that the reviewers were more lenient in their assessment of medical appropriateness than were the criteria is not unexpected. The reviewers are all licensed clinicians with substantial clinical experience. Their role in these reviews represents the cusp of a major change in medical practice and utilization management, and it is reasonable to expect them to give the benefit of the doubt to the proposing physicians, especially because reimbursement decisions are based on the review outcomes. Our qualitative analyses suggest that the reviewing physician and the attending physician actually "negotiate" facts together, much as parents have been found to negotiate clinical facts when interviewed about their child's medical history.[18] Moreover, the qualitative analyses demonstrate that conflict is less likely if the case is recommended for surgery. Indeed, it is both easier and quicker for the reviewer to complete a case that they recommend for surgery (E.A.B., unpublished data, December 1993). Thus, the path of least resistance may converge with traditional medical beliefs to promote more lenient recommendations than those of the criteria.

Two factors were associated with reviewers' recommendations that were at variance from the criteria: a history of tympanostomy tube insertion and female sex. Ongoing qualitative analyses of the step 2 interactions demonstrate that the reviewers orient to a history of prior tubes. Clinically, a history of previous tympanostomy tube surgery does not necessarily predict the subsequent need for tubes. In fact, neither the UR firm's explicit criteria nor the 1994 Agency for Health Care Policy and Research (AHCPR) Guideline on Otitis Media With Effusion considers a history of previous tubes as a mitigating factor in assessing the appropriateness of tubes.[14]

Boys are generally recognized as having a higher incidence of otitis media than girls. The finding that reviewers are more likely to recommend surgery for girls was unexpected and is not easily explained. Evidence exists that pain and suffering are treated more aggressively in girls than boys.[19] Ironically, the more aggressive provision of these surgeries to girls could reflect lower quality care, even as it implies better access. Further study is necessary. Health planners should be aware of the possibility that the implementation of other clinical guidelines may be similarly affected.

One interpretation of our findings could be that the criteria were too strict and that a positive impact of the physician reviews was to moderate the implementation of the criteria. To see if the criteria were too strict, we compared the recommendations of the criteria with those of the AHCPR Guideline for OME, developed in 1994 using another evidence-based method.[14] It classifies various interventions including tympanostomy tube insertion as recommended, optional, or not recommended in a variety of clinical circumstances. Its target population is children aged 1 through 3 years who have otitis media with effusion in the absence of other disease. Comparing the explicit criteria with the AHCPR guideline for the 64 children in our sample who fell within the target population, the AHCPR guideline recommended against tympanostomy tube surgery in all cases that the UR firm's criteria recommended against surgery. In fact, the UR firm's criteria supported tympanostomy more than twice as often as the AHCPR guideline. Thus, our explicit criteria do not appear to be overly restrictive.

It is important to recognize the limitations of the study. It represents an in-depth analysis of UR conducted by a single

UR firm for a single procedure. Thus, its findings should not be generalized broadly without confirmation in other populations. The children subjected to the step 1 reviews represent a large national population of children insured by 3 major insurance entities but are not a random sample of insured US children. Although the type of review we describe is all but routine today, at the time of the study it represented a more managed approach to care than average. There are only 96 cases, and this limits statistical power. Thus, although we demonstrate statistical significance for the key variables, some of the CIs in the logistic regressions are large.

CONCLUSION

This study demonstrates that the performance of criteria-based UR can be assessed reliably using a simple, straightforward method. In this study, reviewers never recommended against surgery that was supported by the criteria. Some practitioners may be reassured to learn that any systematic bias in the review process tends toward finding care more appropriate than would a strict application of the criteria; others may be disappointed at the amount of the reviewers' discordance from the criteria. These data may be sobering to those who anticipate a rapid reduction in health care costs through the implementation of criteria to eliminate inappropriate care.

Clinicians and health care policymakers have an opportunity and an obligation to hold those conducting UR accountable to explicitly state their performance standards and to demonstrate empirically how well they meet them. Doing so would help to reassure patients as it promotes a scientific basis for assessing and improving quality and cost-efficient care.

This work was funded in part by a grant from the Ambulatory Pediatric Association, McLean, Va, and by the Robert Wood Johnson Foundation, Princeton, NJ.

The authors wish to thank Marlene Nishimoto-Horowitz and Linda Schulman for administrative support, Terri Anderson, MA, for assistance in data retrieval and workup, and Jennifer Strand for transcribing many of the audiotapes. We are grateful to Value Health Sciences, Inc, Santa Monica, Calif, (VHS) for unrestricted access to the data. We also wish to thank Charlie Homer, MD, MPH, Children's Hospital, Boston, Mass; Jonathan Finkelstein, MD, MPH, Children's Hospital and Harvard Medical School and Harvard Pilgrim Health Care, Boston; Hal Morgenstern, PhD, University of California, Los Angeles; Nicholas Fiebach, MD, Yale University School of Medicine, New Haven, Conn; and Lewis First, MD, University of Vermont College of Medicine, Burlington, for various contributions of both substance and style. Ed Park, PhD, RAND provided assistance in creating the study design and the sampling scheme. Robert H. Brook, MD, ScD, RAND and University of California, Los Angeles, provided important conceptual and practical insights. We gratefully acknowledge the late Howard Freeman, PhD, University of California, Los Angeles.

References

1. Lomas J. Words without action? the production, dissemination, and impact of consensus recommendations. *Annu Rev Public Health.* 1991;12:41-65.
2. Wise CG, Billi JE. A model for practice guideline adoption and implementation. *J Qual Improv.* 1995; 21:465-476.
3. Hayward RA, McMahon LF, Bernard AM Jr. Evaluating the care of general medicine inpatients: how good is implicit review? *Ann Intern Med.* 1993; 118:550-556.
4. Rubin HR, Rogers WH, Kahn KL, Rubenstein LV, Brook RH. Watching the doctor watchers: how well do peer review organization methods detect hospital quality problems? *JAMA.* 1992:267:2349-2354.
5. Ellrodt AG, Conner L, Riedinger MS, Weingarten S. Implementing practice guidelines through a utilization management strategy: the potential and the challenges. *Qual Rev Bull.* 1992;18:456-460.
6. Weingarten SR, Riedinger MS, Conner L, et al. Practice guidelines and reminders to reduce duration of hospital stay for patients with chest pain: an interventional trial. *Ann Intern Med.* 1994;120:257-263.
7. Rosenberg SN, Allen DR, Handte JS, et al. Effect of utilization review in a fee-for-service health insurance plan. *N Engl J Med.* 1995;333:1326-1330.
8. Shapiro MF, Wenger NS. Rethinking utilization review. *N Engl J Med.* 1995;333:1353-1354.
9. Kleinman LC, Kosecoff J, Dubois RW, Brook RH. The medical appropriateness of tympanostomy tubes proposed for children under 16 years in the United States. *JAMA.* 1994;271:1250-1255.
10. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care.* 1986;2(1):53-63.
11. Paradise JL, Bluestone CD, Bachman RZ, et al. History of sore throat as an indication for tonsillectomy: predictive limitations of histories that are undocumented. *N Engl J Med.* 1978;298:409-413.
12. Fleiss JL. *Statistical Measures for Rates and Proportions.* 2nd ed. New York, NY: John Wiley & Sons Inc; 1986.
13. Rosenfeld RM. What to expect from the medical treatment of otitis media. *Pediatr Infect Dis J.* 1995; 14:731-738.
14. Stool SE, Berg AO, Berman S, et al. *Otitis Media With Effusion in Young Children: Clinical Practice Guideline Number 12.* Rockville, Md: Agency for Health Care Policy and Research, Public Health Service, US Dept of Health and Human Services; 1994. AHCPR publication No. 94-0622.
15. *SAS Procedures Guide.* 3rd ed. Cary, NC: SAS Institute; 1992.
16. *STATA Statistical Data Analysis.* Los Angeles, Calif: Computing Resource Center; 1987.
17. Kassirer JP. The quality of care and the quality of measuring it. *N Engl J Med.* 1993;329:1263-1265.
18. Tannen D, Wallat C. Interactive frames and knowledge schemas in interaction: examples from a medical examination/interview. *Soc Psychol Q.* 1987;50:205-216.
19. Maccoby EE, Jacklyn CN. *The Psychology of Sex Differences.* Stanford, Calif: Stanford University Press; 1974.

Printed and Published in the United States of America