Rigorous Development does not Ensure that Guidelines are Acceptable to a Panel of Knowledgeable Providers

Teryl K. Nuckols, MD, MSHS[1,2], Yee-Wei Lim, MD, PhD[1], Barbara O. Wynn, MA[1], Soeren Mattke, MD, DSc[1], Catherine H. MacLean, MD, PhD[3], Philip Harber, MD, MPH[4], Robert H. Brook, MD, ScD[1,2], Peggy Wallace, RN, MA, MS[5], Rena H. Garland[1], and Steven Asch, MD, MPH[1,2,6]

[1]RAND Corporation, Santa Monica, CA, USA; [2]Division of General Internal Medicine and Health Services Research, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [3]WellPoint Incorporated, Thousand Oaks, CA, USA; [4]Division of Occupational and Environmental Medicine, Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [5]ACS Healthcare Solutions Incorporated, Playa del Rey, CA, USA; [6]VA Greater Los Angeles Healthcare System, Los Angles, CA, USA.

BACKGROUND: Rigorous guideline development methods are designed to produce recommendations that are relevant to common clinical situations and consistent with evidence and expert understanding, thereby promoting guidelines' acceptability to providers. No studies have examined whether this technical quality consistently leads to acceptability.

OBJECTIVE: To examine the clinical acceptability of guidelines having excellent technical quality.

DESIGN AND MEASUREMENTS: We selected guidelines covering several musculoskeletal disorders and meeting 5 basic technical quality criteria, then used the widely accepted AGREE Instrument to evaluate technical quality. Adapting an established modified Delphi method, we assembled a multidisciplinary panel of providers recommended by their specialty societies as leaders in the field. Panelists rated acceptability, including "perceived comprehensiveness" (perceived relevance to common clinical situations) and "perceived validity" (consistency with their understanding of existing evidence and opinions), for ten common condition/therapy pairs pertaining to Surgery, physical therapy, and chiropractic manipulation for lumbar spine, shoulder, and carpal tunnel disorders.

RESULTS: Five guidelines met selection criteria. Their AGREE scores were generally high indicating excellent technical quality. However, panelists found 4 guidelines to be only moderately comprehensive and valid, and a fifth guideline to be invalid overall. Of the topics covered by each guideline, panelists rated 50% to 69% as "comprehensive" and 6% to 50% as "valid".

CONCLUSION: Despite very rigorous development methods compared with guidelines assessed in prior studies, experts felt that these guidelines omitted common clinical situations and contained much content of uncertain validity. Guideline acceptability should be independently and formally evaluated before dissemination.

KEY WORDS: practice guidelines; quality assurance; health care; evidence-based medicine; attitude of health personnel; evaluation.
J Gen Intern Med 23(1):37–44
DOI: 10.1007/s11606-007-0440-9

INTRODUCTION

Physicians have been slow to incorporate clinical guidelines into their decision making because of 3 basic reasons: lack of knowledge about guidelines, barriers to guideline implementation, and unfavorable attitudes toward guidelines.[1–5] Among unfavorable attitudes, arguably the two most fundamental include physicians believing that a specific guideline does not apply to their patients and their questioning a guideline's interpretation of the evidence (i.e., its validity). If physicians question a guideline's relevance or validity, they may not even consider other issues. Other well-established unfavorable attitudes include believing that the recommendations do not appear cost-effective, lacking confidence in developers, and lacking self-efficacy or motivation.[6–8] Some guidelines are viewed unfavorably by very few physicians (1%), whereas others are viewed unfavorably by the majority (91%). Some physicians find no guidelines acceptable.[6]

In part to increase the likelihood that providers will view guidelines favorably,[9] experts have established formal development and evaluation standards, the most widely accepted of which is the "Appraisal of Guidelines Research and Evaluation" Instrument (AGREE).[10–14] National Guidelines Clearinghouse requires basic standards.[15] Nevertheless, empirical research has not examined whether rigorous development—i.e., good technical quality—consistently produces guidelines that providers find acceptable.

This document includes content previously published in a RAND report, available online at: http://www.rand.org/publications/MG/MG400/.
Received April 20, 2007
Revised September 27, 2007
Accepted October 8, 2007
Published online November 21, 2007

In 2004, legislation in California provided a compelling reason to consider this issue: it required the workers' compensation system to adopt a guideline for utilization management and mandated a systematic evaluation of guidelines.[16–18] But would a guideline selected on the basis of technical quality ensure a favorable reception among providers who treat occupational disorders?

One purpose of this study, sponsored by the California Department of Industrial Relations, was to answer this question. We identified promising guidelines for musculoskeletal disorders and appraised them using AGREE. Our objective was to assess the clinical acceptability of each guideline—i.e., providers' views about how useful each guideline would be in clinical practice. We examined the 2 fundamental dimensions mentioned above: whether providers experienced in caring for musculoskeletal disorders believe that the guidelines applied to common clinical situations (*perceived comprehensiveness*) and whether they agree with the guidelines' interpretations of the evidence (*perceived validity*) (defined below).

We chose an expert panel method because it enabled us to compare the acceptability of individual guidelines, ensuring that the providers had detailed knowledge of the guidelines and generally favorable attitudes toward guidelines (so that any low ratings would reflect perceived flaws in individual guidelines rather than generally unfavorable attitudes). Also, using a multidisciplinary expert panel and well-established panel methods might yield a more reliable,[19] informative, and nuanced evaluation than alternative methods, such as focus groups or a survey of community providers. Because we sought to assess clinical acceptability rather than to establish objective validity or determine the reliability of guideline development methods, panelists assessed guidelines' recommendations and embedded literature summaries as community providers do, basing judgments on their own knowledge and experience rather than on an independent literature review.

METHODS

We searched for guidelines covering musculoskeletal disorders, selected guidelines meeting inclusion criteria, and assessed technical quality. We then developed and applied a method for assessing clinical acceptability. A detailed report is available online at http://www.rand.org/publications/MG/MG400/.[20]

Identifying and Selecting Guidelines

We searched MEDLINE and National Guidelines Clearinghouse using keywords covering musculoskeletal disorders, surveyed specialty society web sites, interviewed workers' compensation experts, and contacted state governments. Next, we applied inclusion criteria addressing technical quality, scope, and cost (Table 1).

Of 73 documents identified,[20] some were not guidelines and many did not cover multiple tests and therapies, were out of date, or lacked an updating plan. Two specialty society guidelines and 3 proprietary utilization management guidelines met all inclusion criteria (Table 2).

Table 1. Inclusion Criteria for Guidelines Receiving Further Evaluation

Criterion	Definition	Source
A guideline	"Systematically developed statements that assist practitioner and patient decisions about appropriate health care for specific clinical circumstances."	Based on literature[9]
Evidence-based, peer-reviewed	Interpreted to mean based, at a minimum, on a systematic review of literature published in medical journals included in MEDLINE	Specified in legislation[16–18]
Nationally recognized	Interpreted to mean accepted by the National Guidelines Clearinghouse; published in a peer-reviewed U.S. medical journal; developed, endorsed, or disseminated by an organization based in two or more U.S. states; currently used by one or more U.S. state governments; or in wide use in two or more U.S. states	Specified in legislation[16–18]
Cover common and costly tests and therapies for disorders of spine, arm, and leg	Guidelines that covered several of the most common and costly tests and therapies in the California workers' compensation system: magnetic resonance imaging of the spine, spinal injections, spinal surgery, physical therapy, chiropractic manipulation, surgery for nerve compression syndromes, shoulder surgery, and knee surgery[19]	Selected with policymakers[20]
Reviewed or updated at least every 3 yrs	Reviewed or updated between June 2001 and June 2004, and we confirmed a plan for updating at 3-yr intervals or less	Based on literature,[43] selected with policymakers[20]
Developed by a multidisciplinary clinical team	A clinical team including at least 3 major types of providers that care for injured workers	Based on literature,[9,11] selected with policymakers[20]
Cost less than $500 per individual user in California		Developed with policymakers[20]

Table 2. Guidelines Meeting All Inclusion Criteria

Guideline name	Abbreviated name	Developer	Last review	Reference
Clinical Guidelines by the American Academy of Orthopedic Surgeons	AAOS	American Academy of Orthopedic Surgeons*	2001–2002	[44]
American College of Occupational and Environmental Medicine Occupational Medicine Practice Guidelines	ACOEM	American College of Occupational and Environmental Medicine*	2003	[45]
Optimal Treatment Guidelines, part of Intracorp Clinical Guidelines Tool(R)	Intracorp	Intracorp†	2004	[46]
McKesson/Interqual Care Management Criteria and Clinical Evidence Summary	McK	McKesson Corporation†	2004	[47]
Official Disability Guidelines (ODG)—Treatment in Workers Compensation	ODG	Work Loss Data Institute†	2004	[48]

*Medical specialty society.
†Proprietary guideline developer.

Appraising Technical Quality

For these 5 guidelines, we evaluated technical quality using AGREE, which assesses the resources and processes used to develop guidelines via 6 domains (Table 3) having 2 to 7 questions each.[11]

Internationally, AGREE is the preferred tool for assessing guidelines and its reliability has been demonstrated.[11–13,21] However, its validity remains unproved,[12,22] and it does not assess guideline content or the underlying evidence.[12,13,23] A question within AGREE's rigor-of-development domain stipulates that independent, external reviews by both clinical and methodological experts should be performed, but it does not specify what such reviews should entail.[11]

AGREE is supported by detailed instructions; training is not required.[24] At least 2 raters but preferably 4 raters are needed. We had a physician with a Ph.D. in health services research and an author of a paper on guideline quality rate AGREE questions using guideline content, developers' answers, and corroborating evidence.[30] The AGREE method creates standardized domain scores from individual ratings, incorporating any variability into overall scores.[11]

Appraising Clinical Acceptability

We evaluated the guidelines' acceptability to expert providers, considering 2 dimensions not included in AGREE: *perceived comprehensiveness* and *perceived validity*.[11] A multidisciplinary panel rated guidelines in a 2-round, modified Delphi process derived from the RAND/UCLA Appropriateness Method,[25] which has been shown to be moderately reliable and to have predictive validity.[26,27] Separate panels examining the same clinical topics have yielded similar conclusions (κ=0.51–0.83).[28] Panel judgments about the appropriateness of Surgery have predicted results of subsequent randomized trials,[29] and process-oriented quality measures developed using this method have been positively associated with patient outcomes.[30]

This panel included: 1 Internal Medicine physician, 2 occupational medicine physicians, 1 physiatrist, 1 physical therapist, 1 neurologist board-certified in pain management, 2 doctors of chiropractic, 2 orthopedic surgeons, and 1 neurosurgeon. We contacted relevant specialty societies seeking leaders with experience treating workers and practicing at least 20% of their time, then chose panelists from diverse locations and practice settings who self-reported having favorable attitudes toward guidelines and extensive experience in guideline development or implementation.

Next, we specified the content panelists would evaluate within the guidelines: *appropriateness* and *quantity* of care for common and costly therapies. Specifying *appropriateness* is part of the definition of a guideline;[9] *quantity* was required by the legislation.[16–18] *Appropriate care* meant potential benefits to individual patients exceeded risks irrespective of cost, *inappropriate care* meant risks to patients exceeded potential

Table 3. Technical Quality Appraisal: AGREE Instrument Standardized Domain Scores

Domain	Definition	AAOS	ACOEM	Intracorp	McK	ODG
Scope and purpose	The overall aim of the guideline, the specific clinical questions, and the target patient population	1.00	0.89	0.89	1.00	1.00
Stakeholder involvement	The extent to which the guideline represents the views of its intended users	0.54	0.79	0.79	0.88	0.79
Rigor-of-development	The process used to gather and synthesize the evidence, the methods to formulate the recommendations and to update them	0.81	0.88	0.83	0.88	0.81
Clarity and presentation	The language and format of the guideline	0.96	0.88	1.00	1.00	0.96
Applicability	The likely organizational, behavioral and cost implications of applying the guideline	0.17	0.33	0.33	0.61	0.72
Editorial independence	The independence of the recommendations and acknowledgement of possible conflict of interest from the guideline development group	1.00	1.00	0.75	1.00	0.92

AGREE standardized domain scores range from 0.00 (lowest) to 1.00 (highest).

benefits, and *care of uncertain appropriateness* was between the two. *Quantity* meant frequency, intensity, and duration of therapy.

To give the evaluation depth and breadth, we selected *10 condition/therapy pairs* that were common and costly in the California workers' compensation system (Table 4).[31] Considering appropriateness of care for each condition/therapy pair and, when relevant, quantity of care, yielded *16 topics* to rate. An occupational medicine physician and a registered nurse annotated guideline booklets to identify relevant content. Because we were unable to include all common and costly tests and therapies, we asked panelists to rate any residual content (i.e., content within each guideline other than the 16 topics) and each guideline overall.

For each of the 16 topics, any residual content, and each guideline overall, panelists rated *perceived comprehensiveness* and *perceived validity*. *Comprehensive* meant panelists believed that the guideline addressed appropriateness or quantity for most workers with the condition who might be considered candidates for the therapy. For the residual and overall ratings, *comprehensive* meant the guideline addressed most of the common and costly tests and therapies used in treating occupational disorders. *Valid* meant that recommendations regarding appropriateness or quantity were consistent with panelists' interpretations of existing evidence or, in the absence of higher quality evidence, their expert opinions.

Panelists rated guidelines independently during the first round then met in person to discuss areas of disagreement and rerate during the second round. For both rounds, panelists received instructions, planned interpretation methods, annotated guideline booklets, and unblinded electronic access to the guidelines. Using private, equally weighted ballots (our report contains samples),[20] panelists rated *perceived comprehensiveness* and *perceived validity* separately on 9-point scales with 9 as the highest.

Table 4. List of Topics Rated by Panel

Condition	Therapies performed for that condition	Topics rated by panel*
Lumbar spine disorders	Physical therapy	Appropriateness Quantity
	Chiropractic manipulation	Appropriateness Quantity
	Decompression surgery	Appropriateness†
	Fusion surgery	Appropriateness†
Carpal tunnel syndrome	Physical therapy	Appropriateness Quantity
	Chiropractic manipulation	Appropriateness Quantity
	Surgery	Appropriateness†
Shoulder disorders	Physical therapy	Appropriateness Quantity
	Chiropractic manipulation	Appropriateness Quantity
	Surgery	Appropriateness†
Totals	10 condition/therapy pairs	16 topics

**Panelists rated both perceived comprehensiveness and perceived validity of guideline content pertaining to each of the 16 topics.*
†Panelists did not rate the quantity of care for surgery because patients typically undergo a specific surgical procedure only once.

During the 1-day second round in October 2004, panelists received AGREE results and the first-round ratings, including the distribution and panelists' own ratings relative to the distribution. An experienced moderator led the discussions of the 16 topics, any residual content, and each guideline overall. At the conclusion of each discussion, panelists rerated the guidelines. We used a modified Delphi method, rather than a consensus method that forces agreement, to allow different attitudes to be expressed and contend with one another and true agreement or disagreement to emerge. Panelists wrote comments on ballots and made closing statements.

We interpreted ratings as follows: a median rating of 7 to 9 without disagreement = *comprehensive* or *valid* (depending on which was being rated); a median rating of 1 to 3 without disagreement = *not comprehensive* or *not valid*; a median rating of 4 to 6 or any rating with disagreement = *intermediate comprehensiveness* or *uncertain validity*. We defined disagreement as follows: 4 or more panelists rated the guideline in the 1 to 3 range and 4 or more in the 7 to 9 range. In sensitivity analyses, we discarded the single highest and lowest ratings to determine whether outliers affected the results.[25]

For this paper, we summarized *perceived comprehensiveness* and *perceived validity* results across the 16 topics by determining the fraction and percent of topics in each interpretation category. We assumed that developers had not intended guidelines to cover topics that panelists judged to be *not comprehensive*.

RESULTS

Appraising Technical Quality

AGREE scores were generally high (Table 3). Regarding scope and purpose, each guideline had a well-articulated objective clinical questions and target population. RAND appraisers felt that recommendations were usually presented clearly. Rigor-of-development scores were high because developers clearly described the methods used to search for evidence and formulate recommendations.

Uninvolved individuals reviewed all 5 guidelines before publication. American Academy of Orthopedic Surgeons members reviewed AAOS. ACOEM developers described the review as extensive but provided no detail. A distinct unit within the parent company reviewed Intracorp. Various community and academic providers reviewed McKesson. Affiliates of specialty societies and typical end-users reviewed ODG. No developers reported whether clinical and methodological experts were included.

Relative weaknesses included the following. Stakeholder involvement scores were variable when developers did not clearly describe patient involvement or pilot testing. Three applicability scores were low because developers discussed few details about costs, implementation barriers, or criteria for monitoring adherence. Regarding editorial independence, 1 guideline (Intracorp) discussed potential development members' conflicts of interest in limited detail.

Appraising Clinical Acceptability

To summarize (Table 5), panelists felt that Intracorp was most *comprehensive* but *invalid*. Panelists judged the other 4 guidelines to be moderately *comprehensive* and *valid*.

Regarding *perceived comprehensiveness*, panelists frequently felt that the guidelines discussed therapies without defining appropriateness or quantity of care for most workers with the condition. Panelists judged all 5 guidelines to be *of intermediate comprehensiveness* overall. Of the 16 appropriateness and quantity topics that the panelists rated, Intracorp covered all 16 and was judged *comprehensive* for 69% of them. The other 4 guidelines covered fewer topics and were judged *comprehensive* for 50% to 60% of the ones they did cover.

Regarding *perceived validity*, panelists often found guidelines' statements to be only somewhat consistent with their understandings of the literature and opinions. Overall, AAOS and ACOEM were considered *valid*, McKesson and ODG were *of uncertain validity*, and Intracorp was *not valid*. Despite higher overall ratings for AAOS and ACOEM, these guidelines were *valid* for 40% and 50% of the topics they covered, respectively, whereas McKesson and ODG were *valid* for 43% and 36%, respectively. For Intracorp, 31% of covered topics were *not valid* and 6% were *valid*.

Tables 6 and 7 of the Appendix list *perceived comprehensiveness* and *perceived validity* by topic. Sensitivity testing did not affect results.

Two panelists' ballot comments expound upon the Intracorp ratings: "proscriptive", "poor documentation of validity", "no evidence provided", "inconsistent", "too restrictive", and "all conditions treated similarly". In the closing statements, 7 panelists commented that all 5 guidelines were barely satisfactory. As a group, the panelists were generally uncomfortable with the "proscriptive" nature of the proprietary guidelines compared with the specialty society guidelines.

DISCUSSION

We found that 5 major guidelines covering musculoskeletal disorders all performed extremely well on most technical quality standards—far better than many guidelines do.[10,30,32] Given the exceptional resources and processes used in their development, one would expect providers with especially favorable attitudes toward guidelines and extensive clinical knowledge to find these guidelines acceptable. However, our multidisciplinary expert panel found the best of them barely adequate and the worst invalid. Panelists thought that these guidelines neglected common clinical situations, and recommendations often diverged from panelists' interpretations of the literature and experience.

If sufficient development resources and processes do not ensure that guidelines are acceptable to providers, then a mechanism for evaluating this outcome of guideline development is needed. Drawing an analogy between assuring the quality of guidelines and assuring the quality of medical care, a framework for guideline evaluation can be borrowed from the Donabedian model.[33–36] "Structure" would include resources such as the developer's identity, participating providers' qualifications, and the strength of available literature. "Process" would represent the methods for communicating with stakeholders, searching the literature, formulating recommendations, and considering implementation issues. The ultimate outcome would be whether the target population's health improved after guideline implementation.[2] Adherence by providers would represent an intermediate outcome, and specific reasons for nonadherence would be more proximate intermediate outcomes.[6] The most proximate outcomes would be each recommendation's alignment with the best evidence available.[13,22]

Existing evaluation methods address the structure and process of guideline development and assume—erroneously—that favorable outcomes will result.[10,11,14] For example, having a well-defined scope and purpose should produce recommendations that providers find applicable to common clinical situations. Yet our panelists felt that none of the guidelines comprehensively covered one of the most common treatments

Table 5. Clinical Acceptability Appraisal: Summary of Panel Judgments Regarding Guideline Perceived Comprehensiveness and Perceived Validity for 16 Selected Topics, Any Residual Guideline Content Aside from Those 16 Topics, and Overall

	AAOS	ACOEM	Intracorp	McK	ODG
Perceived comprehensiveness of 16 topics, fraction (%)					
Not comprehensive					
All 16 topics	11/16 (69)	4/16 (25)	0/16 (0)	2/16 (13)	2/16 (13)
Intermediate					
All 16 topics	2/16 (25)	6/16 (38)	5/16 (31)	6/16 (38)	6/16 (38)
Topics covered*	2/5 (40)	6/12 (50)	5/16 (31)	6/14 (43)	6/14 (43)
Comprehensive					
All 16 topics	3/16 (19)	6/16 (38)	11/16 (69)	8/16 (50)	8/16 (50)
Topics covered*	3/5 (60)	6/12 (50)	11/16 (69)	8/14 (57)	8/14 (57)
Residual content	Not comprehensive	Comprehensive	Intermediate	Intermediate	Intermediate
Overall	Intermediate	Intermediate	Intermediate	Intermediate	Intermediate
Perceived validity of topics covered, fraction (%)					
Not valid					
Topics covered*	0/5 (0)	0/12 (0)	5/16 (31)	0/14 (0)	0/14 (0)
Uncertain					
Topics covered*	3/5 (60)	6/12 (50)	10/16 (63)	8/14 (57)	9/14 (64)
Valid					
Topics covered*	2/5 (40)	6/12 (50)	1/16 (6)	6/14 (43)	5/14 (36)
Residual content	N/A	Uncertain	Uncertain	Uncertain	Uncertain
Overall	Valid	Valid	Not Valid	Uncertain	Uncertain

**Excludes topics that were judged not comprehensive on the assumption that the guideline was not intended to cover those topics. For example, if 11 of 16 guidelines were judged "not comprehensive" then the guideline covered 5 of the 16 topics and only those 5 topics are included in this denominator.*

for low back problems, physical therapy. A multidisciplinary development group, rigorous methods for searching and synthesizing literature, and editorial independence should produce recommendations that are aligned with the best evidence available. However, another study found that guidelines with better methodological quality were not better aligned with a systematic literature review.[22] Even if a guideline's recommendations are perfectly concordant with the best available evidence, providers might still be dissatisfied if the recommendations are hard to apply or leave important clinical problems unaddressed because of gaps in the literature.

The opposite conjecture, that issues with development resources and processes can adversely affect the intermediate outcome of perceived validity, appears supported by our data. Panelists judged the 1 guideline with an imperfect editorial independence score, Intracorp, to be invalid, despite otherwise excellent technical quality. Others have noted that imperfect editorial independence can be devastating.[9] Also, panelists seemed to feel that developers of proprietary guidelines had less credibility than specialty societies. This may explain why ACOEM and McKesson had similar validity ratings for individual topics, but McKesson had lower overall validity ratings. Lastly, the variable AGREE scores for stakeholder involvement and applicability suggest that developers could have done more to increase acceptability to providers.

Several guideline evaluators, including AGREE authors, have argued that, in addition to being appraised for technical quality, all guidelines should be independently reviewed by clinical experts. However, given our findings, the mere performance of an external review appears insufficient to ensure clinical acceptability. Specific goals and methods for performing clinical reviews have not been established yet,[11,14,37] and they appear to be needed.

Clinical appraisals of an individual guideline should have 2 goals: assessing providers' attitudes toward the guideline and identifying potential barriers to implementation.[6] Ideally, such appraisals should be structured enough to produce scientifically sound and reliable results, and detailed enough to identify problems with particular topics. The method we used could serve as an initial test of providers' attitudes toward guidelines. Studies should evaluate its reliability in this application, performance for other conditions and for individual guideline recommendations, and its ability to predict guideline acceptance by community providers. Our evaluation touched on 3 attitudes (perceived comprehensiveness, perceived validity, and confidence in developer); future appraisals should address others. Guideline developers should also assess providers' perceptions of implementation barriers because these perceptions are strongly associated with actual implementation.[38–41] An instrument for assessing implementation barriers has been published.[42]

In this study, several limitations could have influenced the relationship between technical quality and clinical acceptability. Including more guidelines would have permitted a quantitative assessment of this relationship, but we examined guidelines that exhibited little variability in technical quality, i.e., the guidelines that providers would be most likely to find acceptable. The fact that panelists judged one *not valid* is remarkable. AGREE could be insensitive to differences between well-developed guidelines, and it does not evaluate the underlying evidence or how developers conduct systematic literature reviews. Published evidence is less developed for musculoskeletal disorders than some conditions.

Our panelists assessed perceived validity rather than comparing guideline recommendations against a synthesis of evidence, as others have done,[22] but the latter would not reflect how providers examine guidelines in actual practice where the guidelines themselves must persuade readers of their validity. Attitudes toward individual guidelines vary greatly, and these findings might not apply to other conditions. Some panel physicians were unfamiliar with chiropractic manipulation of the extremities. Our panel method has not been validated for this application, and experts' opinions may differ from those of community providers.

Our inclusion criteria selected both clinical and utilization management guidelines. Although AGREE was not developed to assess the latter, its questions appear relevant. Panelists could have held biases against utilization management guidelines, but concealing developer identities would not have resolved this because the guidelines designed to limit utilization were obvious. Panelists rejected only one such guideline, highlighting its greater discrepancy between technical quality and clinical acceptability.

Nevertheless, our study indicates that rigorous development methods do not consistently produce guidelines that are acceptable to providers, which would likely limit the guidelines' usefulness and, ultimately, their ability to improve care. During guideline development, developers should do more to involve stakeholders and consider implementation issues. Before guideline dissemination, clinical acceptability should be independently and formally evaluated.

Acknowledgements: *Christine Baker, M.A., Executive Officer of the California Commission on Health and Safety and Workers' Compensation and Anne Searcy, M.D., Medical Director of the California Division of Workers' Compensation participated in the study conception and design (specifically, choosing the inclusion criteria for the guidelines). Paul G. Shekelle, M.D, Ph.D., The RAND Corporation, Santa Monica; Division of General Internal Medicine and Health Services Research, Department of Medicine David Geffen School of Medicine at the University of California, Los Angeles; and VA Greater Los Angeles Healthcare System, Los Angeles, California, provided comments on a draft of this manuscript. This study was sponsored by the California Department of Industrial Relations (the Division of Workers Compensation and the California Commission on Health and Safety and Workers' Compensation). The RAND Corporation provided support for the preparation of this manuscript. Although Christine Baker and Anne Searcy, Division of Workers' Compensation participated in the study conception and design, data acquisition, analysis, and preparation of this manuscript were completely independent of the funders.*

Conflicts of Interest: *None disclosed.*

Corresponding Author: *Teryl K. Nuckols, MD, MSHS; Rand Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138, USA (e-mail: teryl@rand.org).*

APPENDIX

Table 6. Panel Judgments Regarding Perceived Comprehensiveness for 16 Selected Topics, Any Residual Content Aside from Those 16 Topics, and Overall (C = Comprehensive, I = Intermediate, NC = Not Comprehensive)

Panel Judgments for Each of the 16 Topics	AAOS	ACOEM	Intracorp	McKesson	ODG
Lumbar Spine					
Physical Therapy					
Appropriateness	I	I	I	I	I
Quantity	NC	I	C	I	I
Chiropractic Manipulation					
Appropriateness	NC	I	I	I	C
Quantity	NC	NC	C	I	I
Decompression Surgery					
Appropriateness	C	C	C	C	I
Fusion Surgery					
Appropriateness	C	C	I	C	C
Carpal Tunnel					
Physical Therapy					
Appropriateness	NC	I	I	I	C
Quantity	NC	NC	C	I	I
Chiropractic Manipulation					
Appropriateness	NC	I	C	C	C
Quantity	NC	C	C	C	I
Surgery					
Appropriateness	C	C	C	C	C
Shoulder					
Physical Therapy					
Appropriateness	I	C	C	C	C
Quantity	NC	I	C	C	C
Chiropractic Manipulation					
Appropriateness	NC	NC	I	NC	NC
Quantity	NC	NC	C	NC	NC
Surgery					
Appropriateness	NC	C	C	C	C
Residual Content	NC	C	I	I	I
Overall	I	I	I	I	I

Table 7. Panel Judgments Regarding Perceived Validity for 16 Selected Topics, Any Residual Content Aside from Those 16 Topics, and Overall

Panel Judgments for Each of the 16 Topics	AAOS	ACOEM	Intracorp	McKesson	ODG
Lumbar Spine					
Physical Therapy					
Appropriateness	U	U	U	U	U
Quantity	N/A	U	U	U	U
Chiropractic Manipulation					
Appropriateness	N/A	U	NV	U	U
Quantity	N/A	N/A	NV	U	U
Decompression Surgery					
Appropriateness	V	V	U	V	U
Fusion surgery					
Appropriateness	V	U	NV	U	U
Carpal Tunnel					
Physical Therapy					
Appropriateness	N/A	U	U	U	V
Quantity	N/A	N/A	U	U	U
Chiropractic Manipulation					
Appropriateness	N/A	U	NV	V	V
Quantity	N/A	V	NV	V	U
Surgery					
Appropriateness	U	V	U	V	V
Shoulder					
Physical Therapy					
Appropriateness	U	V	U	V	U
Quantity	N/A	V	U	U	V
Chiropractic Manipulation					
Appropriateness	N/A	N/A	U	N/A	N/A
Quantity	N/A	N/A	U	N/A	N/A
Surgery					
Appropriateness	N/A	V	V	V	V
Residual Content	N/A	U	U	U	U
Overall	V	V	NV	U	U

V = valid, U = uncertain validity, NV = not valid, N/A = not applicable
Not Applicable: The content pertaining to these topics was judged not comprehensive and, therefore, on the assumption that the guideline was not intended to cover those topics, validity judgments are not applicable.

REFERENCES

1. **Mosca L, Linfante AH, Benjamin EJ, et al.** National study of physician awareness and adherence to cardiovascular disease prevention guidelines. Circulation. 2005;111(4):499–510.
2. **Bauer MS.** A review of quantitative studies of adherence to mental health clinical practice guidelines. Harv Rev Psychiatry. 2002;10(3):138–53.
3. **Cruz-Correa M, Gross CP, Canto MI, et al.** The impact of practice guidelines in the management of Barrett esophagus: a national prospective cohort study of physicians. Arch Intern Med. 2001;161(21):2588–95.
4. **Switzer GE, Halm EA, Chang CC, Mittman BS, Walsh MB, Fine MJ.** Physician awareness and self-reported use of local and national guidelines for community-acquired pneumonia. J Gen Intern Med. 2003;18(10):816–23.
5. **Christakis DA, Rivara FP.** Pediatricians' awareness of and attitudes about four clinical practice guidelines. Pediatrics. 1998;101(5):825–30.
6. **Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA.** Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282(15):1458–65.
7. **Ward MM, Vaughn TE, Uden-Holman T, Doebbeling BN, Clarke WR, Woolson RF.** Physician knowledge, attitudes and practices regarding a widely implemented guideline. J Eval Clin Pract. 2002;8(2):155–62.
8. **Cabana MD, Ebel BE, Cooper-Patrick L, Powe NR, Rubin HR, Rand CS.** Barriers pediatricians face when using asthma practice guidelines. Arch Pediatr Adolesc Med. 2000;154(7):685–93.
9. Committee to Advise the Public Health Service on Clinical Practice Guidelines, Institute of Medicine. In: **Field MJ, Lohr KN, eds.** Clinical Practice Guidelines: Directions for a New Program. Washington, DC: National Academies Press; 1990.
10. **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(20):1900–5.
11. The AGREE Collaboration. The Appraisal of Guidelines for Research & Evaluation (AGREE) Instrument. Available at: http://www.agreetrust.org/docs/AGREE_Instrument_English.pdf. Accessed March 21, 2007.
12. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care. 2003;12(1):18–23.
13. **Burgers JS.** Guideline quality and guideline content: are they related? Clin Chem. 2006;52(1):3–4.
14. **Shiffman RN, Shekelle P, Overhage JM, Slutsky J, Grimshaw J, Deshpande AM.** Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Ann Intern Med. 2003;139(6):493–8.
15. United States National Guideline Clearinghouse Inclusion Criteria. Available at: http://www.guideline.gov/about/inclusion.aspx. Accessed July 6, 2006.

16. State of California. Assembly Bill 749 (Calderon), Workers' Compensation: Administration and Benefits; 2002.
17. State of California. Senate Bill 228 (Alarcón), Workers' Compensation; 2003.
18. State of California. Senate Bill 899 (Poochigan), Workers' Compensation; 2004.
19. **Hutchings A, Raine R, Sanderson C, Black N.** An experimental study of determinants of the extent of disagreement within clinical guideline development groups. Qual Saf Health Care. 2005;14(4):240–5.
20. **Nuckols TK, Wynn BO, Lim Y, et al.** Evaluating Medical Treatment Guideline Sets for Injured Workers in California. Santa Monica, CA: RAND Corporation; 2005. MG-400-ICJ.
21. **Grol R, Cluzeau FA, Burgers JS.** Clinical practice guidelines: towards better quality guidelines and increased international collaboration. Br J Cancer. 2003;89(Suppl):S4–8.
22. **Watine J, Friedberg B, Nagy E, et al.** Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. Clin Chem. 2006;52(1):65–72.
23. **Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D.** A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care. 2005;17(3):235–42.
24. The AGREE Collaboration. The Appraisal of Guidelines for Research & Evaluation (AGREE) Instrument Training Manual. Available at: http://www.agreetrust.org/docs/AGREE_INSTR_MANUAL.pdf. Accessed March 22, 2007.
25. **Fitch K, Bernstein SJ, Aguilar MD, et al.** The RAND/UCLA Appropriateness Method User's Manual. Santa Monica, CA: RAND Corporation; 2001. MR-1269-DG-XII/RE. Available at: http://www.rand.org/publications/MR/MR1269/.
26. **Shekelle PG.** The appropriateness method. Med Decis Making. 2004;24(2):228–31.
27. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
28. **Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE.** The reproducibility of a method to identify the overuse and underuse of medical procedures. N Engl J Med. 1998;338(26):1888–95.
29. **Shekelle PG, Chassin MR, Park RE.** Assessing the predictive validity of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. Int J Technol Assess Health Care. 1998;14(4):707–27.
30. **Higashi T, Shekelle PG, Adams JL, Kamberg CJ, Roth CP, Solomon DH, Reuben DB, Chiang L, MacLean CH, Chang JT, Young RT, Saliba DM, Wenger NS.** Quality of care is associated with survival in vulnerable older patients. Ann Intern Med. 2005 Aug 16;143(4):274–81.
31. **Wynn BO, Bergamo G, Shaw RN, Mattke S, Dembe A.** Medical Care Provided California's Injured Workers: An Overview of the Issues. Santa Monica, CA: RAND Corporation; 2006. WR-394-ICJ.
32. **Hasenfeld R, Shekelle PG.** Is the methodological quality of guidelines declining in the U.S.? Comparison of the quality of U.S. agency for health care policy and research (AHCPR) guidelines with those published subsequently. Qual Saf Health Care. 2003;12(6):428–34.
33. **Brook RH, McGlynn EA, Cleary PD.** Quality of health care. Part 2: measuring quality of care. N Engl J Med. 1996;335(13):966–70.
34. **Donabedian A.** Explorations in Quality Assessment and Monitoring, Vol. 1. The Definition of Quality and Approaches to its Assessment. Ann Arbor, MI: Health Administration Press; 1980.
35. **Donabedian A.** Explorations in Quality Assessment and Monitoring, Vol. 2. The Criteria and Standards of Quality. Ann Arbor, MI: Health Administration Press; 1982.
36. **Donabedian A.** Explorations in Quality Assessment and Monitoring, Vol. 3. The Methods and Findings of Quality Assessment and Monitoring: An Illustrated Analysis. Ann Arbor, MI: Health Administration Press; 1985.
37. **Shekelle PG, Woolf SH, Eccles M, Grimshaw J.** Clinical guidelines: developing guidelines. BMJ. 1999;318(7183):593–6.
38. **Maue SK, Segal R, Kimberlin CL, Lipowski EE.** Predicting physician guideline compliance: an assessment of motivators and perceived barriers. Am J Manag Care. 2004;10(6):383–91.
39. **Mottur-Pilson C, Snow V, Bartlett K.** Physician explanations for failing to comply with "best practices". Eff Clin Pract. 2001;4(5):207–13.
40. **Grol R, Dalhuijsen J, Thomas S, Veld C, Rutten G, Mokkink H.** Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ. 1998;317(7162):858–61.
41. **Burgers JS, Grol RP, Zaat JO, Spies TH, van der Bij AK, Mokkink HG.** Characteristics of effective clinical guidelines for general practice. Br J Gen Pract. 2003;53(486):15–9.
42. **Shiffman RN, Dixon J, Brandt C, et al.** The GuideLine Implementability Appraisal (GLIA): development of an instrument to identify obstacles to guideline implementation. BMC Med Inform Decis Mak. 2005;5:23–34.
43. **Shekelle PG, Ortiz E, Rhodes S, et al.** Validity of the Agency for Healthcare Research and Quality Clinical Practice Guidelines: how quickly do guidelines become outdated? JAMA. 2001;286(12):1461–7.
44. American Academy of Orthopaedic Surgeons. Guidelines and support documents for hip pain, knee injury, knee osteoarthritis, low back pain/sciatica, shoulder pain and wrist pain. 2001–2003.
45. American College of Occupational and Environmental Medicine. Occupational Medicine Practice Guidelines: Evaluation and Management of Common Health Problems and Functional Recovery in Workers, Second Edition. Beverly Farms, Massachusetts; 2004.
46. Intracorp. Intracorp Optimal Treatment Guidelines. Philadelphia, Pennsylvania; 2003. Available at: http://www.intracorp.com.
47. McKesson. InterQual Care Management Criteria for Workers' Comp & Clinical Evidence Summaries. (Formerly QualityFirst® Workers' Compensation/Disability Guidelines). Newton, Massachusetts. 2004.
48. Work Loss Data Institute. Official Disability Guides (ODG), Treatment in Workers' Compensation. San Diego, CA; 2004. Available at: http://www.worklossdata.com.