Are Nonspecific Practice Guidelines Potentially Harmful? A Randomized Comparison of the Effect of Nonspecific Versus Specific Guidelines on Physician Decision Making

Paul G. Shekelle, Richard L. Kravitz, Jennifer Beart, Michael Marger, Mingming Wang, and Martin Lee

Objective. To test the ability of two different clinical practice guideline formats to influence physician ordering of electrodiagnostic tests in low back pain.

Data Sources/Study Design. Randomized controlled trial of the effect of practice guidelines on self-reported physician test ordering behavior in response to a series of 12 clinical vignettes. Data came from a national random sample of 900 U.S. neurologists, physical medicine physicians, and general internists.

Intervention. Two different versions of a practice guideline for the use of electrodiagnostic tests (EDT) were developed by the U.S. Agency for Health Care Policy and Research Low Back Problems Panel. The two guidelines were similar in content but varied in the specificity of their recommendations.

Data Collection. The proportion of clinical vignettes for which EDTs were ordered for appropriate and inappropriate clinical indications in each of three physician groups were randomly assigned to receive vignettes alone, vignettes plus the nonspecific version of the guideline, or vignettes plus the specific version of the guideline.

Principal Findings. The response rate to the survey was 71 percent. The proportion of appropriate vignettes for which EDTs were ordered averaged 77 percent for the no guideline group, 71 percent for the nonspecific guideline group, and 79 percent for the specific guideline group ($p = .002$). The corresponding values for the number of EDTs ordered for inappropriate vignettes were 32 percent, 32 percent, and 26 percent, respectively ($p = .08$). Pairwise comparisons showed that physicians receiving the nonspecific guidelines ordered fewer EDTs for appropriate clinical vignettes than did physicians receiving no guidelines ($p = .02$). Furthermore, compared to physicians receiving nonspecific guidelines, physicians receiving specific guidelines ordered significantly more EDTs for appropriate vignettes ($p = .0007$) and significantly fewer EDTs for inappropriate vignettes ($p = .04$).

Conclusions. The clarity and clinical applicability of a guideline may be important attributes that contribute to the effects of practice guidelines.

Key Words. Practice guidelines: nonspecific, specific; physician decision making

Practice guidelines are being developed and disseminated with the goal of improving healthcare by helping physicians to make better clinical decisions. The Institute of Medicine has identified eight important attributes of guidelines: validity, reproducibility/reliability, clinical applicability, clinical flexibility, clarity, multidisciplinary process, scheduled review, and documentation (Institute of Medicine 1992). Most of the work on guideline development has focused on the validity of the guideline (Audet, Greenfield, and Field 1990; Eddy 1990; Woolf 1992; Hayward, Wilson, Tunis, et al. 1995; Grimshaw, Eccles, and Russell 1995; Grimshaw and Russell 1993a; Eccles, Clapp, Grimshaw, et al. 1996). While validity is necessary for a practice guideline to achieve its intended effects, it may not be sufficient. Clinicians need to be able both to understand the guideline and apply it to individual patients. There has been no experimental work on measuring the effects of these other attributes of guidelines. We took advantage of two alternative forms of guidelines for the use of electrodiagnostic tests, both developed by the U.S. Agency for Health Care Policy and Research (AHCPR) Low Back Problems Clinical Practice Guideline Panel, to test the potential effects of the clinical applicability and clarity of the guideline on a random sample of the guideline's intended target users.

METHODS

This study evaluated the effect of two different versions of a practice guideline on physician decision making in low back pain. The practice guideline

---

Supported by the Center for the Study of Provider Behavior, a VA Health Services Research Program. Cover letters supporting the survey were provided by the American College of Physicians, the American Academy of Neurology, and the American Academy of Physical Medicine and Rehabilitation. The conclusions presented herein are those of the authors and do not necessarily represent the position of the Department of Veterans Affairs, the American College of Physicians, the American Academy of Neurology, or the American Academy of Physical Medicine and Rehabilitation.

Address correspondence to Paul G. Shekelle M.D., Ph.D., Senior Research Associate, Veterans Affairs Health Services Research & Development Service, West Los Angeles Veterans Affairs Medical Center (111G), 11301 Wilshire Blvd., Los Angeles CA 90073. Dr. Shekelle, Michael Marger, M.P.H., Mingming Wang, M.P.H., and Martin Lee, Ph.D. are from the Center for the Study of Provider Behavior, West LA and Sepulveda VAMC. Jennifer Beart, M.A. is from the University of California, Los Angeles School of Medicine, and Richard L. Kravitz, M.D., M.S.P.H. is from the Division of General Medicine and Center for Health Services Research in Primary Care, University of California, Davis. This article, submitted to *Health Services Research* on July 22, 1998, was revised and accepted for publication on March 1, 1999.

concerned the use of electrodiagnostic tests (electromyography and nerve conduction velocity tests, abbreviated EDTs) in patients with acute and subacute low back pain syndromes. Physician decision making was assessed using 12 structured clinical vignettes. General internists, neurologists, and physical medicine specialists identified from the American Medical Association Masterfile were randomly assigned to receive the vignettes alone (the control group), the vignettes plus a nonspecific version of the guideline (Guideline group A), or the vignettes plus a specific version of the guideline (Guideline group B).

Creation of Guidelines

The process by which the two different guidelines were created has previously been described in detail (Shekelle and Schriger 1996). In brief, the AHCPR Low Back Problems Clinical Practice Guideline Panel created two different versions of guidelines for the use of EDTs. The guideline panel had 23 members, who represented a wide variety of back pain specialists including orthopedic and neurosurgical specialists, primary care physicians, neurologists, rehabilitation physicians, chiropractors, as well as others, including a patient. The first set of guidelines was created from a systematic review of the literature and a roundtable informal consensus method, and resulted in the guideline presented in Figure 1. In this study we have designated these guidelines as "nonspecific." Several months later, the same group using the same systematic literature review applied the appropriateness method expert judgment process to create an alternative set of guidelines for the use of EDTs, shown in Figure 2. As opposed to the informal roundtable consensus method, the appropriateness method is comprehensive and specific in that it creates appropriateness criteria for a large number of clinically detailed

Figure 1: Nonspecific Guidelines as Printed in the Physician Survey

PRACTICE GUIDELINE

1. Needle EMG and H-reflex may be useful in assessing questionable nerve root dysfunction in patients with leg symptoms lasting longer than four weeks (regardless of whether patients also have back pain).
2. If the diagnosis of radiculopathy is obvious and specific on clinical examination, electrophysiologic testing is not recommended.
3. SEP may be useful in assessing suspected spinal stenosis and spinal cord myelopathy.

Figure 2: Specific Guidelines as Printed in the Physician Survey

PRACTICE GUIDELINE

I. In patients with low back problems of greater than four weeks duration, electrophysiologic tests are in general most useful in assessing patients in the following clinical situations:

   a. Patients with definite lower limb neurologic deficits (with or without lower limb symptoms) and a nonconcordant single-level anatomic abnormality or anatomic abnormalities at multiple levels on CT or MRI.
   b. Patients with definite lower limb neurologic deficits (with or without lower limb symptoms) and who have not had a CT or MRI (or the CT or MRI is normal) and whose symptoms are getting worse on repeat examination.
   c. Patients with lower limb symptoms and equivocal lower limb neurologic deficits who have not had a CT or MRI (or the CT or MRI is normal) and whose symptoms are getting worse on repeat examination.
   d. Patients with lower limb symptoms and/or a positive straight-leg raise test and equivocal lower limb neurological deficits and a nonconcordant single-level anatomic abnormality or anatomic abnormalities at multiple levels on CT or MRI and whose symptoms are getting worse on repeat examination.
   e. Patients with lower limb symptoms and definite lower limb neurologic deficits and a potential peripheral neuropathic process whose symptoms are unchanged or getting worse on repeat examination.

II. Electrophysiologic tests are in general not useful in assessing patients with low back problems in the following clinical situations:

   a. Patients with symptoms of less than four weeks duration (including those with lower limb symptoms).
   b. Patients with no neurologic deficits, except those patients whose symptoms are getting worse on repeat examination and

continued

Figure 2: *Continued*

have either lower limb symptoms and no or a normal CT or MRI, or lower limb symptoms and/or a positive straight-leg raise and CT or MRI evidence of anatomic abnormalities at multiple levels.

c. Patients without definite lower limb neurologic deficits seen on the first visit.

III. There was panel disagreement on the usefulness of electrophysiologic tests in assessing patients with low back problems in the following clinical scenarios:

a. Patients with symptoms of greater than four weeks duration and equivocal or definite neurologic deficits and a concordant anatomic abnormality on CT or MRI.

b. Patients with symptoms of greater than four weeks duration and definite neurologic abnormalities seen on the first visit.

c. Patients with symptoms of less than four weeks duration with or without lower limb symptoms who have definite lower limb neurologic deficits and a potential peripheral neuropathic process.

patient presentations (termed indications); further, it does not force consensus. From 219 indications rated for appropriateness by the panel, staff created 11 guideline statements that were then reviewed by the panel. In this study we have designated these guidelines as "specific." Although the guideline panel judged in private voting, by a two-to-one margin, that incorporating all or some of the specific EDT guidelines into the final document was desirable, they ultimately chose instead to retain the nonspecific guidelines, primarily because the specific guideline statements had a style and substance dissimilar to the remainder of the low back guideline statements. A full explication of the appropriateness method has been published (Brook, Chassin, Fink, et al. 1986; Brook 1994) along with its use in this circumstance (Shekelle and Schriger 1996).

Development of Clinical Vignettes

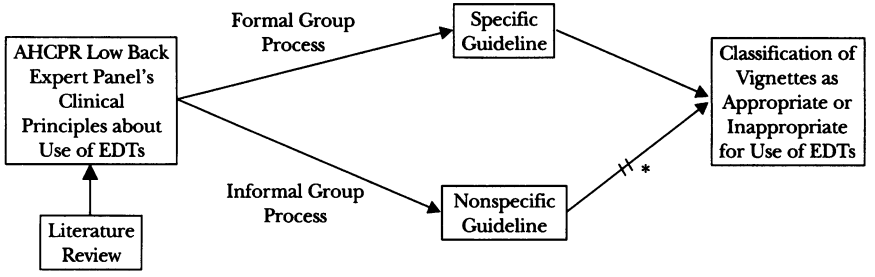
We created 12 clinical vignettes that described in words and pictures patients with back pain syndromes who might present to a physician and be

considered for a possible evaluation with EDTs. In theory we could have used either version of the EDT guideline in order to classify these scenarios, because both versions represented the same clinical principles derived from the experts' assessment of the literature and their clinical judgment. However, in practice we found it impossible to classify the vignettes unambiguously using the nonspecific version. Consequently, we were able to use only the specific guidelines to classify these vignettes as appropriate or inappropriate for the use of EDTs using the specific guidelines. The relationship among the clinical principles governing the use of EDTs, the two guideline types, and classification of the vignettes is depicted in Figure 3. Five vignettes described patient syndromes that the panel judged to be inappropriate for EDT use; an additional five vignettes described patient syndromes that the panel judged to be appropriate for EDT use. Two vignettes described patient syndromes that were judged as neither clearly appropriate nor inappropriate for EDT use. We pilot tested these vignettes to assess their understandability and clinical relevance on a convenience sample of general internists, neurologists, and physical medicine physicians at our respective institutions, and revised the vignettes accordingly. Figure 4 presents one of the final vignettes in words and pictures as it appeared in the survey, and the appendix presents all of the final vignettes in words only.

Survey Instrument

In addition to the vignettes, all survey participants were asked nine questions about practice setting, the numbers of back pain patients typically seen, and

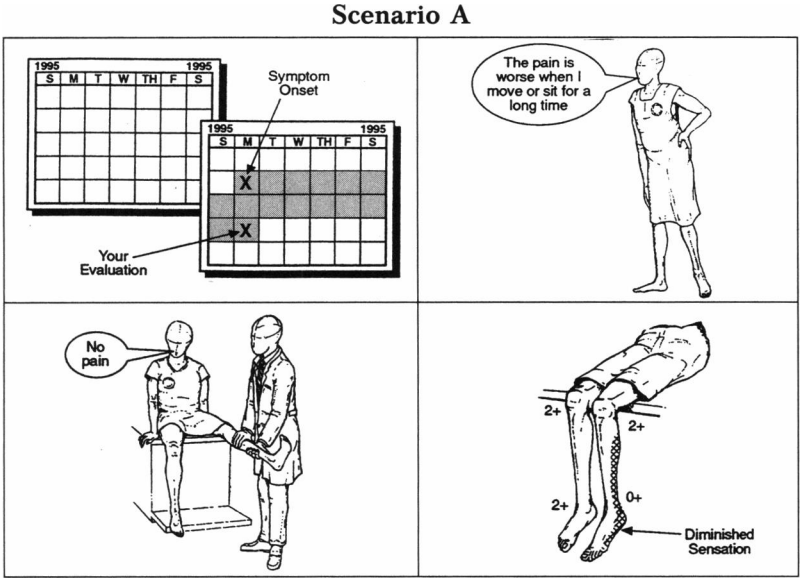
Figure 3: Relationship Between Clinical Principles for Use of Electrodiagnostic Tests, Two Types of Guidelines, and Classification of Clinical Vignettes



* Too nonspecific to allow reliable classification

EDT=electrodiagnostic tests

Figure 4: A Sample Clinical Vignette as it Appeared in the Physician Survey



A 42-year-old man seeks care for the first time because of two weeks of moderate low back pain that is worse with movement and sitting. The pain does not radiate to either leg. Straight leg raising is negative, but there is definite sensory loss along the posterior left leg and a diminished left ankle jerk.

Would you order, perform or refer this patient for electrodiagnostic studies (*electromyography, somatosensory evoked potentials, or nerve conduction velocity*) at this time? (Circle one.)

YES NO

---

the number of EDTs typically ordered. Guideline group A also received the nonspecific guidelines depicted in Figure 1, and Guideline group B received the specific guidelines depicted in Figure 2. Groups A and B were instructed:

> "Before responding to the vignettes, please consider the following draft guideline about the use of electrodiagnostic studies for patients with low back pain syndromes. It was developed by a national multi-disciplinary

panel of back experts assembled by the Agency for Health Care Policy and Research, a branch of the United States Public Health Service."

Groups A and B were further queried: "To what extent do you agree that this guideline describes optimal clinical practice with respect to the use of electrodiagnostics in patients with low back pain?" and the response items ranged from strongly agree to strongly disagree on a 5-point scale.

Physician Sample

A list of the names and addresses of randomly sampled physicians in three specialties was obtained from Buckley-Dement Direct Marketing Services (Chicago). Buckley-Dement maintains an up-to-date listing of medical and osteopathic physicians in the United States, using information provided by the American Medical Association (AMA) and the American Osteopathic Association. The AMA database has been found to be an excellent source of information on all medical doctors regardless of AMA membership. The three specialties chosen were selected because they represent a substantial fraction of the target audience for AHCPR guidelines: general internists, because they are primary care physicians; and neurologists and physical medicine physicians, because a prior study identified these two specialties as most likely to order or perform electrodiagnostic tests (Cherkin et al. 1994). Samples of 300 names for each specialty were received.

Survey Process

Cover letters encouraging participation were obtained from leaders of the respective specialty societies (i.e., the American College of Physicians, the American Academy of Neurology, and the American Academy of Physical Medicine and Rehabilitation) and were included in the appropriate survey packets. Each of the three types of survey (vignettes alone, vignettes plus nonspecific guidelines, and vignettes plus specific guidelines) was randomly assigned to each of the three specialty groups (i.e., 100 members of each specialty were assigned to each of the three guideline groups). All surveys were mailed out in November 1995, accompanied by a $10 cash incentive. Nonrespondents to the first-round mailing were sent a second mailing, and nonrespondents to the second-round mailing were sent a third-round mailing. We attempted to get telephone numbers from directory assistance for a random sample of 70 of the nonrespondents to the third-round mailing and found that 24 percent could not be contacted because they had moved or left practice.

Statistical Analysis

The percentages of correct EDT ordering practice were determined on the basis of all vignettes and separately for those vignettes rated as appropriate and those rated as inappropriate. An average percentage was determined for each respondent in each of these categories and was then averaged across respondents. To compare these results across the three different categories of respondents (control, nonspecific guideline, and specific guideline) the nonparametric Kruskal-Wallis test was used (analogous to a one-way analysis of variance) (Glantz 1992). Pairwise comparison of the groups employed the nonparametric Mann-Whitney test (analogous to the unpaired *t*-test). To formulate confidence intervals comparing these proportions, we employed the standard large-sample interval for the difference of two means. To analyze the results for individual vignettes across guideline groups, the standard chi-square test was used. Finally, in comparing the practice characteristics across the guideline groups, we used the Kruskal-Wallis test for continuous data and the chi-square test for categorical data.

RESULTS

Response Rate

Of 900 physicians in the random sample, 70 had moved, leaving no forwarding address, or had died, and 545 returned surveys (crude response rate: 66 percent; response rate corrected for noncontactable nonrespondents: 71 percent). The response rate differed significantly by specialty (general internists were less likely to respond than were neurologists or physical medicine physicians), but it did not vary by type of survey (vignettes alone, vignettes plus nonspecific guideline, or vignettes plus specific guideline).

Practice Setting of Respondents

Table 1 presents data about the practice setting of the respondents, according to the type of survey received. No statistically significant differences were found among groups on the mean amount of time devoted to patient care or other activities, the mean total number of patients seen or the mean number of patients with sciatica seen, the mean number of electrodiagnostic studies ordered or performed per year, or the types of practice arrangements and reimbursement received.

Table 1: Comparison of Means of Practice Characteristics Among Three Survey Groups

Practice Characteristics	*No Guidelines (n=184)*	*Nonspecific Guidelines (n=192)*	*Specific Guidelines (n=169)*
Proportion of time spent in			
Direct patient care	71%	71%	78%
Teaching	8%	7%	7%
Research	5%	7%	3%
Administration	10%	8%	6%
Number of outpatients seen in a typical week	54	48	52
Number of sciatica patients seen in a typical month	41	38	37
Number of electrodiagnostic studies ordered or performed in the past year	86	64	91
Proportion of respondents who personally perform			
Electromyography	58%	52%	48%
Nerve conduction studies	57%	52%	48%
Proportion of respondents in multispecialty groups	29%	33%	30%
Proportion of patients with the following forms of health insurance			
Prepaid (capitated)	16%	18%	19%
Fee-for-service	29%	31%	29%
No insurance	9%	9%	9%
Government pay (Medicare and Medicaid)	45%	43%	44%

Note: There were no statistically significant differences among the means.

Responses to Individual Clinical Vignettes

Table 2 presents for each vignette the mean proportion of respondents in each survey group who ordered an EDT. Overall, EDTs were ordered in 75 percent of the vignettes judged appropriate (range 64 percent to 92 percent), 31 percent of the vignettes judged inappropriate (range 10 percent to 47 percent), and 53 percent of the vignettes judged uncertain (range 34 percent to 70 percent).

The Effect of Different Practice Guidelines on Physician Test Ordering

No differences were found among the three groups in the mean proportion of EDTs ordered in response to the 12 vignettes (Table 3). However, for the

Table 2: Proportion of Respondents Ordering Electrodiagnostic Tests for Clinical Vignettes

	Proportion Ordering Electrodiagnostic Test			
Clinical Vignettes	*No Guideline Group (n=184)*	*Nonspecific Guideline Group (n=192)*	*Specific Guideline Group (n=169)*	p-*Value**
Judged appropriate				
C	77%	68%	79%	.053
D	66%	67%	79%	.010
E	85%	90%	92%	.136
I	66%	64%	67%	.884
K	84%	65%	80%	.001
Judged inappropriate				
A	39%	29%	24%	.014
B	47%	46%	45%	.859
G	12%	10%	12%	.825
J	27%	36%	22%	.013
L	37%	38%	34%	.740
Judged uncertain				
F	69%	62%	70%	.227
H	46%	34%	39%	.065

*p-Value for differences in mean proportion among groups using the Kruskal-Wallis test.

five appropriate vignettes, physicians assigned to the specific guidelines group ordered more EDTs than did physicians assigned to the nonspecific guidelines group (79% vs. 71%; difference 8%, 95% CI for the difference, 4%–13%). For the five inappropriate vignettes, physicians in the specific guideline group ordered fewer EDTs than did those in the nonspecific group (27% vs. 32%; difference −5%; 95% CI for the difference, −10%–0%). Physicians assigned to the nonspecific guidelines group ordered fewer EDTs for appropriate vignettes than did physicians assigned to the control group who did not receive any guidelines (71% vs. 77%; difference −6%; 95% CI for the difference, −11%– −1%).

We defined an "optimal test-ordering score" as the mean proportion of EDTs ordered "correctly" (meaning the sum of the appropriate vignettes for which EDTs were ordered and the inappropriate vignettes for which EDTs were not ordered). Physicians assigned to the specific guidelines group had significantly higher scores than did those in both the nonspecific guidelines group and the control (no guidelines) group (Table 3).

Two-way Kruskal-Wallis analysis of variance showed, in addition to a significant effect of the type of guideline, a significant effect of the provider

Table 3: Effect of Guidelines on Electrodiagnostic Test Ordering

	Mean Proportion of Electrodiagnostic Tests Ordered				*Difference Between Groups (95% Confidence Interval)*		
Clinical Vignettes Aggregated	*No Guideline Group*	*Nonspecific Guideline Group*	*Specific Guideline Group*	*Difference Among Groups**	*Nonspecific Versus No Guideline*	*Specific Versus No Guideline*	*Specific Versus Nonspecific*
All 12 vignettes	54%	50%	53%	$p = .16$	−4 (−9, .4)	−1 (−6, 3)	−3 (−7, 1)
Five appropriate vignettes	77%	71%	79%	$p = .002$	−6 (−11, −1)	2 (−2, 7)	8 (4, 13)
Five inappropriate vignettes	32%	32%	27%	$p = .08$	0 (−5, 5)	−5 (−10, 0)	−5 (−10, 0)
Optimal test ordering†	71%	68%	75%	$p = .0002$	−3 (−6, .4)	4 (1, 7)	7 (3, 10)

*Exact p-values calculated from the Kruskal-Wallis test.

†Combination of correct responses for both appropriate and inappropriate vignettes.

type, with general internists ordering fewer tests than neurologists or physical medicine physicians (means of 44%, 60%, and 54% of vignettes, respectively, $p = .0001$). A significant interaction effect between the type of guideline and type of physician was looked for; none was found.

No difference was found in the degree to which physicians who received guidelines agreed that the guideline represented optimal practice. The mean response on a 1–5 scale from strongly agree to strongly disagree was 2.58 for both guideline groups, indicating that, on average, both physician groups "somewhat agreed" that the guidelines represented optimal practice.

DISCUSSION

In this randomized controlled trial we tested the potential effect on physician decision making of two alternative versions of a guideline for the use of electrodiagnostic studies for patients with low back pain syndromes. We used random sampling from the AMA Masterfile, achieved a good response rate to our survey, and found no differences among study groups on relevant practice setting variables. Therefore, we believe our results to be both internally valid (as the only differences among groups was exposure to alternative forms of the guideline) and generalizable to our intended target population: primary care physicians and the specialists who perform electrodiagnostic tests, who comprise the intended audience for this AHCPR guideline.

A principal finding of this study was that the effect of both guidelines was small. We expected as much, based on the literature about changes in provider behavior that show passive dissemination to have small effects, usually (Grimshaw and Russell 1993b). We also had hypothesized that the appropriateness method guideline would have a beneficial effect on physician decision making while the informal consensus method guideline would have no effect. We found that the appropriateness method guideline was associated with a significant improvement in optimal test ordering (a combination of a rise in appropriate test ordering and a decrease in inappropriate test ordering). In contrast, the only potential effect of the informal consensus method guideline was a significant decrease in appropriate test ordering, with no effect on inappropriate test ordering. In all cases, the test-ordering behavior of physicians who received the appropriateness method guideline was significantly better than that of physicians who received the informal consensus method guideline. We also found differential test ordering among physician specialties but no interaction between guideline type and physician specialty.

What can explain these results? One potential explanation—that physicians agreed more with the specific guideline than with the nonspecific guideline—is unlikely, because we found no difference between groups on acceptance of the guideline as representative of optimal practice. Similarly, the guidelines shared equally in the weight of authority invested in them because both were developed under the auspices of the AHCPR. A more likely explanation for our results, we believe, is that the informal consensus guideline is too nonspecific to allow physicians to understand for which patients the guideline is recommending either for or against the use of electrodiagnostic tests. For example, the decrease in appropriate test ordering in the nonspecific guideline is largely due to differences in test ordering on vignettes C and K. We believe that this is because the nonspecific guideline states that EDTs are not recommended if "the diagnosis of radiculopathy is obvious and specific on clinical examination," without further specifying the meaning of "obvious and specific." In the clinical indications rated for appropriateness of EDT use, it is clear that the panel did not consider the clinical situations described in vignettes C and K as "obvious and specific" for radiculopathy.

This difference in specificity between the two guidelines is directly related to the guideline development methods used. The two guidelines were each developed by the same AHCPR panel using the same literature review, but they varied in the method used to combine the literature with expert judgment. In areas of controversy, the informal consensus method tends to produce "lowest common denominator" statements that all panelists can agree on (Shekelle and Schriger 1996; Kosecoff, Kanouse, Rogers, et al. 1987). Unfortunately, such statements are sometimes too vague to allow physicians to act appropriately (McDonald and Overhage 1994). The finding that nonspecific guideline statements may actually decrease appropriate test-ordering behavior was unexpected. Thus, the overall effect of nonspecific guidelines may not be "no effect," but actually a deleterious effect. Because this finding was unexpected, more research is needed to see if it is replicable in other clinical situations and with other types of practice guidelines.

The major limitation to this study, which is general to all vignette-based studies, is that we are measuring physicians' reported intentions rather than their actual behavior. Physicians may answer surveys such as this in a way that they think is socially desirable, or "right," rather than the way in which they actually practice (Jones, Gerrity, and Earp 1990). While studies of the validity of vignette-based approaches to predicting individual physician behavior have reported mixed results, the available evidence supports the validity of vignette-based studies as an approach for measuring differences

between groups (Sandvik 1995; Braspenning and Sargent 1994; Wennberg, Dickens, Blener, et al. 1997; Carey and Garrett 1996). Most pertinently, Carey and colleagues reported a comparison between physicians' actual diagnostic radiology test–ordering behavior for patients with low back pain syndromes and the same physicians' responses to a series of clinical vignettes (Carey and Garrett 1996). The vignettes used by Carey and colleagues were very similar in form and content to the vignettes used in our study, and one of the vignettes was identical (vignette G). Although the absolute rate at which physicians order diagnostic tests for patients with low back pain may be overestimated by using clinical vignettes, the study concluded that vignettes accurately predicted aggregate comparisons of physician behavior among groups. Therefore, we believe that the available data support the idea that the relative differences among the groups we observed in our vignette study would be observed were we to measure actual physician behavior. Our experimental results with vignettes on the differences between specific and nonspecific guidelines are also supported by a recent observational study that did measure actual physician behavior. In that study, the second most important attribute in predicting lack of compliance with guidelines by Dutch general practitioners was that the guideline was vague and nonspecific (Grol, Dalhuijsen, Thomas, et al. 1998).

The differences we observed in this study were small in absolute magnitude. Whether our study underestimated the absolute magnitude of change in physician behavior (because of its weak implementation method) or overestimated this change (because of the potential upward bias on test ordering observed in some vignette studies) is unknown. However, we have shown that the likely direction of the effect with the specific guideline is positive, whereas the likely direction of the effect with the nonspecific guideline is negative.

Other possible mechanisms may have produced the observed effect. We have hypothesized that the effect stemmed from the differences in the clarity and clinical applicability that were reported when the physicians described for which patients the guideline was applicable, but it may instead have been attributable to differences in the terminology used by each guideline. The nonspecific guideline lists clinical situations where EDTs "may be useful," while the specific guideline lists clinical situations where EDTs are "in general most useful." This is another, but clearly different, manifestation of the clarity and clinical applicability of the guideline (i.e., the clarity of the utility of the procedure), and it is deserving of future study.

These results have important implications for those who seek to improve healthcare through the use of practice guidelines. Because both have been

developed by the AHCPR Low Back Problems panel, the nonspecific and the specific guidelines had equivalent face validity. Indeed, the physicians we surveyed endorsed their respective guidelines equivalently as representing optimal practice. Yet the two versions of the guidelines produced different results. This suggests that validity alone is not sufficient for guidelines to have their intended effects. The clinical applicability and clarity of the guideline is important too. In this case, specific guidelines may change physician decision making for the better, while nonspecific guidelines may change physician behavior for the worse. Testing for the acceptability and potential efficacy of newly created guidelines may allow the identification of ineffective or potentially harmful guidelines before they are disseminated.

APPENDIX

Clinical Vignettes Used in Survey

Scenario A. A 42-year-old man seeks care for the first time because of two weeks of moderate low back pain that is worse with movement and sitting. The pain does not radiate to either leg. Straight leg raising is negative, but there is definite sensory loss along the posterior left leg and a diminished left ankle jerk.

Scenario B. A 47-year-old male patient is referred to you by a family practice colleague because of back and leg pain beginning six weeks ago. Examination by your colleague at that time revealed a positive straight leg raising test and a normal neurologic examination. Plain x-ray of the lumbar spine was normal and a CT scan showed a posterolateral herniated disc at L5-S1. Your colleague prescribed ibuprofen and reduced physical activity, but there has been no change in the patient's symptoms. Your examination today confirms the neurologic examination is normal. The straight leg raise is still positive.

Scenario C. A 35-year-old businessman has had low back pain radiating to the right leg for two months. Examination two weeks ago was consistent with an L5 radiculopathy, but a subsequent MRI scan of the spine showed a protuberant disc at the L3–4 level. During today's visit the patient reports that the pain is slightly better, but the neurologic examination remains abnormal (i.e., still consistent with L5 radiculopathy).

Scenario D. A 60-year-old office manager appears in follow-up for low back pain of seven weeks duration. During an initial exam four weeks ago, she

had a positive straight leg raising test on the left, diminished pin sensation in the web space of the great left toe, and weakened dorsiflexion of the left foot. Plain x-ray of the lumbar spine is normal. The patient's pain has become worse since the last visit.

Scenario E. A 52-year-old man seeks consultation for low back pain of eight weeks duration. Four weeks ago he saw a "back specialist" who performed an MRI; the patient hands you a report which reads "MRI normal." The pain radiates to the left leg. The patient notes that his left foot sometimes "feels numb." Neurologic exam reveals possible sensory loss in the left foot. Straight leg raising is positive. The patient's symptoms have been gradually getting worse.

Scenario F. A 51-year-old man returns for follow-up because of low back pain. Eight weeks ago he developed a "pulling pain" radiating into the right leg. When you saw him last month, examination showed questionable right lower extremity sensory loss, MRI showed protruding disc at L5-S1. Today the patient states that the pain is getting somewhat worse. Except for the equivocal sensory findings, neurological examination is normal.

Scenario G. A 35-year-old male auto mechanic presents with a four-day history of severe acute low back pain with radiation to the posterior calf and lateral foot. There is no known inciting event. Physical examination reveals some sensory deficits in this distribution and a diminished ankle reflex, but no motor weakness. Straight leg raising is limited to 45 degrees in the affected leg. Plain x-ray of the lumbar spine is normal.

Scenario H. A 42-year-old female radiology technician presents to your office complaining of a four-week history of back and left leg pain. The symptoms began shortly after she played soccer with her 10-year-old son. She was seen by a colleague on the second day who prescribed heat and nonsteroidal anti-inflammatory agents. Today, she continues to have pain radiating into the left calf. Physical examination reveals no motor deficits, and reflexes are normal, but there is reduced sensation over the calf and lateral aspect of the left foot. Straight leg raising is limited to 30 degrees. Plain x-ray of the lumbar spine is normal. A CT scan of the lumbar spine shows a posterolateral herniated nucleus pulposus at L5-S1.

Scenario I. A 30-year-old male truck driver with rheumatoid disease presents to your office with a six-week history of back and right leg pain. There is no known inciting event. Physical examination shows no sensory changes but

does reveal motor weakness and a diminished right ankle reflex. He has no limitation on straight leg raising. Plain x-ray of the lumbar spine is normal. No CT or MR has been performed. He has been under the care of a family practice colleague of yours, who has treated the patient with a short course of bedrest and ibuprofen. The patient's pain has not changed over the past three weeks.

Scenario J. You have been following a 38-year-old female teacher, who initially presented four weeks ago with low back pain radiating into the right thigh. There is no known inciting event. Physical examination, four weeks ago and at present, shows no motor or sensory changes in the affected leg, and reflexes are normal. Straight leg raising is limited to 45 degrees in the affected leg. A plain lumbosacral x-ray taken four weeks ago is normal. You have treated the patient with ibuprofen and moist heat. She returns in follow-up with her symptoms unchanged.

Scenario K. A 54-year-old male complains of low back pain radiating to the left leg and foot. The pain has waxed and waned over the past six weeks but is now constant. The patient notices that he has recently been stumbling over the affected foot. Straight leg raising is limited to 45 degrees in the affected leg. Neurologic exam shows weakness in the extension of the left foot. CT scan shows herniated discs as L2-3, L4-5, and L5-S1.

Scenario L. On his first visit to your office, a 23-year-old investment banker presents with low back pain radiating to the thigh. The patient has had this pain for 2-1/2 months. Examination reveals questionable diminished ankle reflexes. Straight leg raising is normal. No imaging studies have been performed.

REFERENCES

Audet, A. M., S. Greenfield, and M. Field. 1990. "Medical Practice Guidelines: Current Activities and Future Directions." *Annals of Internal Medicine* 113 (9): 709–714.

Braspenning, J., and J. Sargent. 1994. "General Practitioners' Decision Making for Mental Health Problems: Outcomes and Ecological Validity." *Journal of Clinical Epidemiology* 47 (12): 1365–72.

Brook, R. H. 1994. "The RAND/UCLA Appropriateness Method." In *Clinical Practice Guideline Development: Methodology Perspectives,* edited by K. A. McCormick, S. R. Moore, and R. A. Siegel, pp. 59–70. Agency for Health Care Policy and Research, Pub. No. 95-0009. Rockville, MD: Public Health Service.

Brook, R. H., M. R. Chassin, A. Fink, D. H. Solomon, J. Kosecoff, and R. E. Park. 1986. "A Method for the Detailed Assessment of the Appropriateness of Medical Technologies." *International Journal of Technology Assessment in Health Care* 2 (1): 53–63.

Carey, T. S., and J. Garrett. 1996. "Patterns of Ordering Diagnostic Tests for Patients with Acute Low Back Pain." *Annals of Internal Medicine* 125 (10): 807–14.

Cherkin, D. C., R. A. Deyo, K. Wheeler, and M. A. Ciol. 1994. "Physician Variation in Diagnostic Testing for Low Back Pain: Who You See Is What You Get." *Arthritis and Rheumatism* 37 (1): 15–22.

Eccles, M., Z. Clapp , J. M. Grimshaw, et al. 1996. "Developing Valid Guidelines: Methodological and Procedural Issues from the North of England Evidence-Based Guideline Development Project." *Quality in Health Care* 5 (1): 44–50.

Eddy, D. M. 1990. "Practice Policies: Guidelines for Methods." *Journal of the American Medical Association* 263 (9): 1839–41.

Glantz, P. A. 1992. *Primer of Biostatistics, 3rd edition,* pp. 324–48. New York: McGraw-Hill, Inc.

Grimshaw, J. M., M.P. Eccles, and I. T. Russell. 1995. "Developing Clinically Valid Practice Guidelines." *Journal of Evaluation in Clinical Practice* 1: 37–48.

Grimshaw, J. M., and I. T. Russell. 1993a. "Achieving Health Gain Through Clinical Guidelines: I. Developing Scientifically Valid Guidelines." *Quality in Health Care* 2 (4): 243–48.

———. 1993b. "Effect of Clinical Guidelines on Medical Practice: A Systematic Review of Rigorous Evaluations." *The Lancet* 342 (8883): 1317–322.

Grol, R., J. Dalhuijsen, S. Thomas, C. in't Veld, G. Rutten, and H. Mokkink. 1998. "Attributes of Clinical Practice Guidelines that Influence the Use of Guidelines in General Practice: Observational Study." *British Medical Journal* 317 (7162): 858–61.

Hayward, R. S. A., M. C. Wilson, S. R. Tunis, E. B. Bass, and G. Guyatt. 1995. "User's Guide to the Medical Literature, VIII. How to Use Clinical Practice Guidelines: A. Are the Recommendations Valid?" *Journal of the American Medical Association* 274 (7): 570–74.

Institute of Medicine. 1992. *Guidelines for Clinical Practice: From Development to Use,* edited by M. J. Field and K. N. Lohr. Washington, DC: National Academy Press.

Jones, T. V., M. S. Gerrity, and J. Earp. 1990. "Written Case Simulations: Do They Predict Physicians Behavior?" *Journal of Clinical Epidemiology* 43 (8): 805–15.

Kosecoff, J., D. E. Kanouse, W. H. Rogers, L. McCloskey, C. M. Winslow, and R. H. Brook. 1987. "Effects of the National Institutes of Health Consensus Development Program on Physician Practices." *Journal of the American Medical Association* 258 (19): 2708–13.

McDonald, C. J., and J. M. Overhage. 1994. "Guidelines You Can Follow and Trust: An Ideal and an Example." *Journal of the American Medical Association* 271 (11): 872–73.

Sandvik, H. 1995. "Criterion Validity of Responses to Patient Vignettes: An Analysis

Based on Management of Female Urinary Incontinence." *Family Medicine* 27 (6): 388–92.

Shekelle, P. G., and D. L. Schriger. 1996. "Using the Appropriateness Method in the Agency for Health Care Policy and Research Clinical Practice Guideline Development Process." *Health Services Research* 31 (4): 453–68.

Wennberg, D. E., J. D. Dickens, L. Blener, F. J. Fowler, D. N. Soule, and R. B. Keller. 1997. "Do Physicians Do What They Say? The Inclination to Test and Its Association with Coronary Angiography Rates." *Journal of General Internal Medicine* 12 (3): 172–76.

Woolf, S. H. 1992. "Practice Guidelines, A New Reality in Medicine: Methods of Guideline Development." *Archives of Internal Medicine* 152 (5): 946–52.