

# Profiling Care Provided by Different Groups of Physicians: Effects of Patient Case-Mix (Bias) and Physician-Level Clustering on Quality Assessment Results

Sheldon Greenfield, MD; Sherrie H. Kaplan, PhD, MPH; Richard Kahn, PhD; John Ninomiya, MS; and John L. Griffith, PhD

**Background:** Patient characteristics (case-mix bias) and physician-level variation (clustering) are often overlooked in profiling the quality of care provided by different groups of physicians, such as specialties.

**Objective:** To examine the effect of case-mix bias and physician-level clustering on differences in quality of diabetes care between specialty groups participating in the American Diabetes Association's Provider Recognition Program.

**Design:** Retrospective record review of both process and outcome measures over 1 year and a cross-sectional patient survey. The sample included 29 solo and group practice sites in diverse regions of the United States. Of the 29 sites, 15 were endocrinology sites and 14 were primary care sites.

**Patients:** 1750 adults with diabetes.

**Measurements:** Process measures included frequency of hemoglobin A<sub>1c</sub>, lipid, and urine protein testing; blood pressure measurement; and foot and eye examinations. Outcome measures included A<sub>1c</sub> level, blood pressure, lipid levels, and patient satisfaction. Patient case-mix variables included age, sex,

health status, level of education, ethnic minority status, and duration of diabetes.

**Results:** Unadjusted differences between endocrinologists and generalists were statistically significant for most process and outcome measures. Inclusion of patient case-mix variables reduced the statistical significance of specialty differences for some quality measures. After accounting for the substantial physician-level clustering, observed differences between specialties were no longer statistically significant for any of the quality measures except patient satisfaction.

**Conclusions:** The findings underscore the importance of designing physician profiling studies with sufficient power to account for physician-level variation (clustering) as well as patient case-mix. Studies that are not designed with both sufficient numbers of physicians and patients per physician may distort differences in quality of care between physician groups.

*Ann Intern Med.* 2002;136:111-121.

[www.annals.org](http://www.annals.org)

For author affiliations, contributions, and current addresses, see end of text.

For a glossary of terms, see end of text.

See editorial comment on pp 153-154.

Identification of physician characteristics associated with optimal management of chronic disease could serve as the basis for designing quality improvement initiatives. However, assessment of the quality of care provided by physicians with different characteristics (for example, generalists vs. specialists, male vs. female, or those in health maintenance organizations vs. those with fee-for-service arrangements) presents a unique set of methodologic problems that stem from the nature of outpatient medicine. First, if older or sicker patients with multiple diseases have different needs for health care services and different health outcomes independent of the quality of care they receive, physicians who see such patients may appear to provide lower quality of care than do those who see younger patients with less comorbid disease. Accounting for these patient characteristics (that is, avoiding case-mix bias) is therefore an essential feature of fair and accurate comparisons of physicians' quality of care.

Second, patients with specific characteristics (such as age, sex, or health problems) choose and remain with physicians who have specific characteristics (such as age, sex, specialty training, proximity, or practice style). Patients in a physician's practice (or even those seen by a small group of physicians practicing together) might therefore "cluster," that is, be more like each other and differ from patients who are drawn to another physician's practice. Thus, patients are not independent observations, as assumed by standard statistical techniques. While also inherent in any unit or "cluster" being compared (such as a health plan, a hospital, or a state), the smaller and more homogenous the unit (for example, the individual physician's office practice as opposed to a hospital), the greater the potential for a clustering problem (1). Accounting for physician-level clustering is therefore another key feature of scientifically sound profiling of physicians' quality of care.

Finally, to be reliable for profiling at the physician

level, quality-of-care measures must be chosen that register the physician's style, or "thumbprint." For example, measures of the process of medical care (such as test ordering or interpersonal behavior) are more directly under the physician's control and are therefore more likely than some outcome and utilization measures to register the physician's thumbprint (1, 2). To detect physician style reliably, a sufficient number of patients per physician must be sampled.

Although these methodologic issues have been recognized in the design and analysis of randomized trials (1, 3, 4), attempts to address both patient case-mix and physician-level clustering in observational studies of quality-of-care assessment have been limited (2, 5–12). As part of a continuing effort to draw attention to quality of care, the American Diabetes Association (ADA) developed the Provider Recognition Program to acknowledge physicians who provide high-quality diabetes care. To illustrate the effects of patient characteristics and physician-level clustering on quality assessment results among patients with diabetes, we used ADA Provider Recognition Program data to compare two groups of physicians, in this case generalists and specialists, whose results were expected to differ (7, 13).

## METHODS

### Definitions

"Groups of physicians" refers to physicians who share a characteristic, such as specialty; it does not designate a collection of physicians practicing together. "Site" or "practice" refers to the office setting, as in "solo or group practice" or "endocrinologist practices," not to the behavior of physicians (as in "practice patterns"). We use a broad definition of "case-mix" that includes such patient characteristics as age, sex, level of education, and health status. "Physician-level clustering" refers to the variation in any measurements (patient characteristics or quality-of-care measures) associated with the physician. In this study, it reflects the extent to which rates of performance of quality measures differ between physicians within a specialty. The effect of this clustering is summarized by the inflation factor (also called "design effect" [1]), which is the ratio of estimated variances of differences between specialty groups, with and without adjustment for clustering (see Glossary).

### Site Selection and Patient Sampling

A convenience sample of 15 managed care plans, 10 clinics, and 6 endocrinology practices responded to a nationwide ADA solicitation to participate in a diabetes-related quality-of-care study. Seven of the 15 managed care plans agreed to recruit physicians. All 10 clinics and all 6 of the endocrinology practices agreed to have some or all of their physicians participate. Thirty-one practice sites that included a total of 63 physicians agreed to participate in the study. Of these 31 sites, 2 dropped out because of the burden of data collection.

Each participating practice self-reported specialty of the practice. To avoid misclassification of practices, the investigators verified specialty. Physicians at each participating practice were asked to submit data only for those patients for whom they provided the principal diabetes care.

To maximize the reliability of physician-level quality measures, approximately 35 patients per site were to be sampled. All but 3 practice sites recruited 27 or more patients; the average number of patients sampled per site was 67. All patients received most of their diabetes care at the practice site. Practice sites recruited consecutive patients starting at an index date.

### Data Collection

Clinical data were abstracted from medical records for all visits to participating physicians in the 12 months immediately preceding the index date. The care assessed was delivered between October 1994 and July 1996. Following detailed ADA specifications, personnel at each practice site abstracted study measures from patients' medical records. Abstraction by sites was determined to agree with expert ADA personnel by performing duplicate abstraction at two participating practices.

A survey to collect patient case-mix variables and ratings of interpersonal care was also administered at the time of recruitment. A majority of patients (69%) returned the survey by mail; the remainder (31%) completed the survey in the physician's office. The method of survey administration did not affect the variables analyzed. Completed patient surveys were obtained from 1258 (73.5%) patients. Compared with responders, nonresponders had a higher mean hemoglobin A<sub>1c</sub> level (0.091 [9.1%] vs. 0.087 [8.7%]), had had diabetes for

less time (10.8 vs. 14.3 years), and fewer received most of their care from an endocrinologist (68% vs. 73%).

Since clinical measures not found in the medical record were scored as not performed rather than as missing, none of these measures was imputed. Missing data from the survey exceeded 5% for only four variables. Values were imputed for single-item measures by using Hot-Deck (14); multi-item scales were imputed by using each respondent's average of completed items for those completing at least 50% of scale items. Although multiple imputation methods are available, a single imputation procedure for missing data was chosen for this project in order to focus on the issues of clustering and case-mix, apart from additional sources of variability.

### Measures

A panel of experts recruited by the ADA identified process and outcome measures that together defined a comprehensive evaluation of quality of care for diabetes. These measures were lipid profile (frequency of measurement and levels of total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and triglycerides), dilated eye examination (a note, report, or letter from a physician or optometrist documenting the date of the examination, or a dated photograph of the retina), foot examination (reference to a visual inspection or sensory examination of the feet), blood pressure, urinalysis or any test for urinary protein (such as 24-hour collection or dipstick protein testing), and self-monitoring of blood glucose (any mention of patient-collected blood glucose values and any mention of patient self-adjustment of diet or treatment on the basis of blood glucose results).

Glycated hemoglobin was recorded as hemoglobin A<sub>1c</sub> or total glycated hemoglobin. Total glycated hemoglobin values were converted to hemoglobin A<sub>1c</sub> values. For items that appeared more than once during the study year, the most recent value recorded was used.

Satisfaction with care was measured by using four well-tested survey items, rated on a five-point scale from "excellent" to "poor" (15): personal manner of the health care provider, overall diabetes care, explanation of laboratory results, and how well questions were answered. The proportion of "excellent" scores was used in the analysis. Self-reported health status was measured by using a single five-point rating scale from "excellent" to "poor." This item has been shown to reflect the sum-

mary of diabetes complications and comorbid conditions (16).

### Statistical Analysis

All data were analyzed by using SAS software (17). We used univariate analysis to examine frequency distributions and variability in study measures. Derived variables from the patient survey were tested for internal consistency reliability by using the Cronbach  $\alpha$  value; reliability coefficients exceeded 0.70 for all study measures. Characteristics of respondents and nonrespondents were compared by using *t*-tests for continuous variables and chi-square tests for proportional variables.

For each process and outcome measure, regression models were used to calculate estimated odds ratios and 95% CIs comparing the two specialties. Odds ratios were estimated from regression models with no adjustments, inclusion of patient case-mix only, inclusion of physician-level clustering only, and inclusion of both case-mix and physician-level clustering. Although the odds ratio may not be the optimal metric when performance rates are extremely high or low (18, 19), we used odds ratios only to illustrate the effects of patient case-mix bias and physician-level clustering on specialty comparisons. Models were adjusted for patient age, level of education, ethnic minority status, sex, duration of diabetes, treatment with insulin, and health status. Health status and education were entered as continuous variables. Crude and covariate-only adjusted estimates were obtained from logistic regression models.

Generalized estimation equations (20) were used to investigate the additional effect (beyond case-mix bias) of physician-level clustering on significance testing. This effect, summarized by the inflation factor, is derived from the covariate-adjusted model and the full model with both covariate and cluster adjustments. It reflects the proportional decrease (or increase) in sample size required to see statistically significant differences between the groups being compared given the extent of clustering of patients by physician. It can be written as  $IF = 1 + (n - 1) \times p$ , where *IF* is the inflation factor, *p* is the intraclass correlation coefficient, and *n* is the number of patients seen per physician. The intraclass correlation coefficient measures how similar patients are within practices and how different they are from patients in other practices (21). In this study, after adjust-

**Table 1. Proportion of Patients Conforming with Process and Outcome Measures, by Specialty and Unadjusted for Patient Characteristics**

Quality Measure	Patients of Endocrinologists (n = 1250)	Patients of Generalists (n = 500)	Difference (95% CI)	P Value
	%		percentage points	
<b>Process measure*</b>				
Hemoglobin A <sub>1c</sub> measurement	90	79	11 (6 to 14)	<0.001
Lipid measurement	45	51	-6 (-11 to 0)	0.042
Urine protein measurement	58	42	16 (12 to 22)	<0.001
Blood pressure	91	92	-1 (-2 to 3)	>0.2
Eye examination	41	34	7 (2 to 12)	0.008
Foot examination	75	54	21 (16 to 26)	<0.001
Self-measurement of blood glucose	83	73	10 (6 to 15)	<0.001
<b>Outcome</b>				
Hemoglobin A <sub>1c</sub> value < 0.10 (<10.0%)†	79	76	3 (1 to 8)	0.153
Hemoglobin A <sub>1c</sub> value < 0.08 (<8.0%)	38	32	6 (1 to 12)	0.029
Total cholesterol level < 5.18 mmol/L (<200 mg/dL)	46	43	3 (-4 to 9)	>0.2
Low-density lipoprotein cholesterol level < 3.37 mmol/L (<130 mg/dL)	51	58	-7 (-14 to 0)	0.063
High-density lipoprotein cholesterol level‡	58	51	7 (0 to 14)	0.046
Triglyceride level < 2.26 mmol/L (<200 mg/dL)	68	57	11 (5 to 17)	<0.001
Blood pressure ≤ 140/90 mm Hg	58	52	6 (0 to 11)	0.037
Satisfaction with interpersonal care rated as "excellent"	38	22	16 (11 to 22)	<0.001

\* Blood pressure was measured twice annually; all other process measures were obtained once annually.

† Unadjusted mean values were 0.087 (8.7%) for endocrinologists and 0.09 (9.0%) for generalists ( $P < 0.05$ ).

‡ Specified levels were >0.91 mmol/L (>35 mg/dL) for men and >1.17 mmol/L (>45 mg/dL) for women.

ment for case-mix, the intraclass correlation coefficient indicates the consistency of the physician's behavior across patients in his or her practice. A high intraclass correlation occurs when the quality measures are similar for patients within a physician's practice and differ from those for patients in another physician's practice. An inflation factor of 1 implies no physician-level clustering (standard analytic approaches assuming independence of observations might therefore be appropriate), and those greater than 1 represent data with positive intraclass correlation. These methods allowed us to examine differences between specialists and generalists after accounting first for patient case-mix, then for physician-level clustering, and finally for both.

#### Role of the Funding Source

This research was supported by the American Diabetes Association, which developed and maintains the Provider Recognition Program. Data from the Provider Recognition Program were analyzed by the authors with support from the ADA. The ADA had no role in the design, conduct, and reporting of the study.

## RESULTS

### Patient and Practice Sample Characteristics

The 29 participating sites were located in 13 states; 14 were in urban locations, 10 in were in suburban locations, and 5 were in rural communities. There were 13 individual physician sites (solo practices) and 16 multiple physician (group practices) sites. Of the solo practitioners sampled, 6 were endocrinologists and 7 were internal medicine or family practice physicians (generalists). Of the multiple physician practices, 9 were endocrinology sites and 7 were generalist sites.

A total of 1750 adult patients participated in the study. The patients ranged in age from 18 to 92 years (mean age [ $\pm$ SD], 59  $\pm$  15 years), 48% were male, 81% were white, and 32% had attended college. We did not distinguish between patients with type 1 and type 2 diabetes because the measures used to assess care were the same for both groups and diabetes type was not uniformly reported.

### Relationship of Quality Measures to Physician Specialty

We first examined the relationship between physician groups (specialties) and quality measures. In a

patient-level analysis unadjusted for patient characteristics, process and outcomes measures differed significantly between specialties (Table 1). Most process measures were performed significantly more often by endocrinologists than generalists. In addition, more patients of endocrinologists achieved specified outcome levels than did patients of generalists; however, only triglyceride level less than 2.26 mmol/L (<200 mg/dL) and patient satisfaction were statistically significant ( $P < 0.01$ ). When the specified level of hemoglobin A<sub>1c</sub> was defined as less than 0.08 (<8.0%), 38% of patients seen by endocrinologists compared with 32% of patients seen by generalists achieved this level ( $P = 0.029$ ).

#### Variation in Patient Characteristics by Specialty

We next compared the characteristics of patients of endocrinologists with those of patients of generalists. Patients of endocrinologists differed significantly from patients of generalists for four of the seven patient characteristics studied (Table 2). Compared with patients of generalists, patients of endocrinologists were significantly younger and better educated, more had had diabetes for 15 years or longer, and more were taking insulin.

#### Relationship of Specialty to Quality Measures after Accounting for Patient Case-Mix and Physician-Level Clustering

Tables 3 and 4 show results of a three-stage examination of the effects of specialty on each of the quality measures in unadjusted analysis, analysis accounting for patient case-mix only, and analysis accounting for both patient case-mix and for physician-level clustering. We examined the statistical significance of odds ratios and

the 95% CIs comparing the two specialties for each adjustment procedure. Odds ratios between the columns “Case-Mix Adjustment Only” and “Full Model Adjustment” in Tables 3 and 4 are therefore similar, but the CIs in the latter column are wider. The inflation factor in Tables 3 and 4 indicates the relative effect of physician-level clustering compared with that of case-mix adjustment only. Larger inflation factors indicate greater effects of physician-level clustering. An inflation factor of approximately 4.0 or greater, for example, indicates an approximate doubling of the CI (the square root of 4).

Although the unadjusted comparisons of specialties for each process measure were similar to the results shown in Table 1, the differences between specialties in the performance of annual lipid tests, eye examinations, and self-monitoring of blood glucose were no longer statistically significant after adjustment for patient characteristics (Table 3).

Accounting for physician-level clustering increased the variance associated with the specialty variable, such that no difference between specialties in process measures remained statistically significant (Table 3). The inflation factor exceeded 3.0 in all cases, suggesting that adjustment for clustering at the physician level substantially reduced the effective sample size and therefore reduced the study’s statistical power to detect specialty differences in quality measures. The observed intraclass correlations for process measures, a reflection of the physician-level clustering, were moderately high compared with those in the literature (2, 22). The intraclass correlation coefficient for the hemoglobin A<sub>1c</sub> process measure, for example, was 0.18; adjustment for patient case-mix did not change this value.

The Figure shows the variability in the proportion

Table 2. Comparisons of Patient Characteristics (Case-Mix) by Specialty Group

Characteristic	Patients of Endocrinologists (n = 1250)	Patients of Generalists (n = 500)	Difference (95% CI)*	P Value†
Mean age, y	57	62	5 (4 to 6)	<0.001
Men, %	48	48	0 (–5 to 6)	>0.2
College or greater education, %	36	20	16 (11 to 21)	<0.001
Nonwhite ethnicity, %	23	19	4 (0 to 10)	0.070
Duration of diabetes ≥ 15 y, %	39	26	13 (8 to 19)	<0.001
Use of insulin, %	68	45	23 (18 to 29)	<0.001
Mean health status score‡	53	51	2 (0 to 6)	0.080

\* Differences between percentages are expressed as percentage points.

† Comparisons of mean differences in proportions of patients with selected characteristics by practice specialty.

‡ Self-reported health status was rated on a 5-point scale from “excellent” to “poor.” Scores have been transformed to range from 0 to 100. Higher scores reflect better health.

**Table 3. Odds Ratios for Process Measures, by Method of Adjustment**

Process Measure*	Odds Ratio (95% CI)†			Inflation Factor‡
	Unadjusted	Case-Mix Adjustment Only§	Full Model Adjustment	
Hemoglobin A <sub>1c</sub>	2.24 (1.69–2.97)	1.94 (1.44–2.62)	1.94 (0.81–1.24)	8.73
P value	<0.001	<0.001	0.14	
Lipids	0.81 (0.66–0.99)	0.86 (0.69–1.07)	0.86 (0.50–1.49)	6.40
P value	0.04	0.18	>0.2	
Urine protein	1.96 (1.59–2.42)	1.81 (1.45–2.26)	1.81 (0.83–3.96)	12.70
P value	<0.001	<0.001	0.14	
Blood pressure	0.95 (0.65–1.38)	0.79 (0.53–1.17)	0.79 (0.30–2.04)	5.74
P value	>0.2	>0.2	>0.2	
Eye examination	1.34 (1.08–1.66)	1.23 (0.98–1.54)	1.23 (0.62–2.43)	8.98
P value	<0.01	0.08	>0.2	
Foot examination	2.52 (2.03–3.13)	2.35 (1.86–2.98)	2.35 (0.86–6.47)	18.53
P value	0.07	<0.001	0.1	
Self-measured blood glucose	1.79 (1.40–2.28)	1.16 (0.88–1.53)	1.16 (0.35–3.90)	19.43
P value	0.001	>0.2	>0.2	

\* Blood pressure was measured twice annually; all other process measures were obtained once annually.

† Odds ratios > 1.0 favor endocrinologists.

‡ Inflation factor (or design effect) is the ratio of the estimated SEs of the regression coefficients for full model versus case-mix model only.

§ Adjusted for patient characteristics (case-mix bias), including age, sex, ethnicity, educational level, duration of diabetes, and health status.

|| Adjusted for both patient characteristics (case-mix bias) and cluster effect.

of patients with unadjusted HbA<sub>1c</sub> values less than 0.10 (10%) for each practice site within specialty. Considerable variation was observed between physicians within each specialty.

Similar results were obtained for each of the outcome measures (Table 4). However, although more patients of endocrinologists had hemoglobin A<sub>1c</sub> values less than 0.08 (8%), more had blood pressure values less than 140/90 mm Hg, and more rated their interpersonal care as “excellent”; adjustment for patient characteristics did not change the statistical significance of differences

between specialties in hemoglobin A<sub>1c</sub> values or satisfaction results. Borderline specialty differences in lipid values in favor of generalists were statistically significant after case-mix adjustment only. Results for other lipid measures were similar (data not shown). After case-mix adjustment, differences in patients’ blood pressure that initially favored endocrinologists became statistically insignificant. Accounting for physician-level clustering increased the variance associated with specialty differences for all outcome measures except satisfaction with care, such that differences between specialties were no

**Table 4. Odds Ratios for Outcome Measures, by Method of Adjustment**

Outcome Measure	Odds Ratio (95% CI)*			Inflation Factor†
	Unadjusted	Case-Mix Adjustment Only‡	Full Model Adjustment§	
Hemoglobin A <sub>1c</sub> value < 0.10 (<10.0%)	1.22 (0.93–1.60)	1.32 (0.99–1.76)	1.32 (0.67–2.59)	5.51
P value	0.15	0.06	>0.2	
Hemoglobin A <sub>1c</sub> value < 0.08 (<8.0%)	1.31 (1.03–1.67)	1.54 (1.19–1.99)	1.54 (0.92–2.57)	3.90
P value	<0.03	<0.03	0.10	
Low-density lipoprotein cholesterol level < 3.37 mmol/L (<130 mg/dL)	0.77 (0.58–1.02)	0.69 (0.51–0.93)	0.69 (0.46–1.04)	1.34
P value	0.07	0.02	0.12	
Blood pressure ≤ 140/90 mm Hg	1.25 (1.01–1.54)	1.06 (0.85–1.33)	1.06 (0.71–1.59)	3.23
P value	<0.04	>0.2	>0.2	
Satisfaction with care	2.20 (1.64–2.96)	1.95 (1.43–2.66)	1.95 (1.14–3.33)	2.98
P value	<0.001	<0.001	0.02	

\* Odds ratios > 1.0 favor endocrinologists.

† The inflation factor (or design effect) is the ratio of the estimated SEs of the regression coefficients for full model versus case-mix model only.

‡ Adjusted for patient characteristics (case-mix bias), including age, sex, ethnicity, educational level, duration of diabetes, and health status.

§ Adjusted for both patient characteristics (case-mix bias) and cluster effect.

longer statistically significant (Table 4). As with process measures, outcome measures had moderately high intra-class correlations (for example, 0.12 for hemoglobin A<sub>1c</sub> value < 0.10 [ $<10\%$ ]) that were not affected by case-mix adjustment.

Reduction in the specification of the hemoglobin A<sub>1c</sub> outcome measure from less than 0.10 ( $<10\%$ ) to less than 0.08 ( $<8\%$ ) reduced the inflation factor, decreased the contribution of physician clustering, and increased the contribution of patient case-mix to differences between specialties (Table 4).

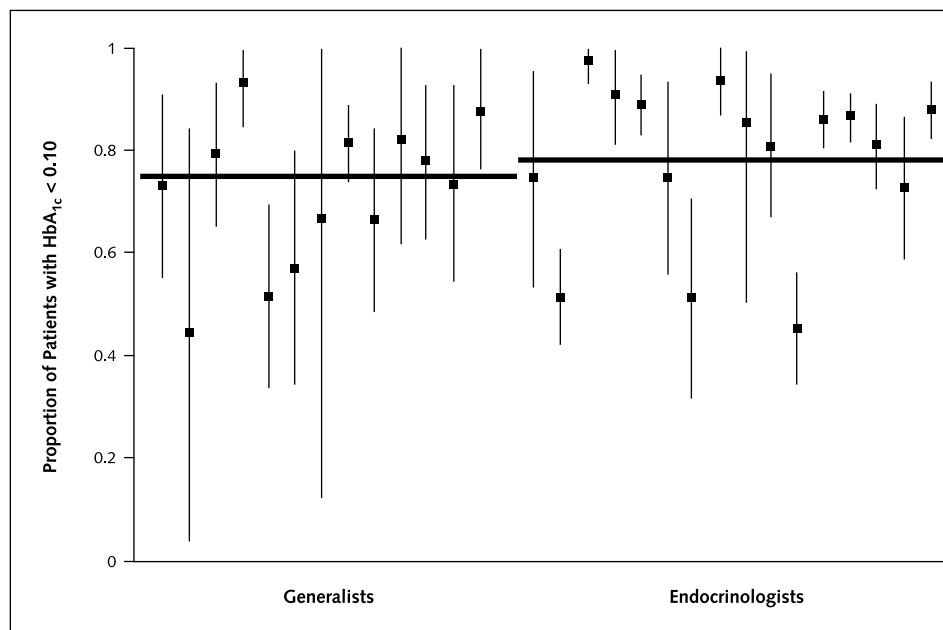
## DISCUSSION

Using data from the ADA Provider Recognition Program, we demonstrated that profiling and comparing the quality of care provided by different physician groups may be inaccurate if careful attention is not paid to patient case-mix and physician-level clustering. Although we compared generalists and specialists, our results would also apply to comparisons of quality of care provided by any group of physicians, such as those practicing in health maintenance organizations versus fee-

for-service arrangements, those in higher-volume versus lower-volume practices, or older versus younger physicians. This research has implications for four areas of the design and analysis of physician-level quality-of-care studies: case-mix measurement and adjustment; choice of quality measures; sampling, power, and physician-level reliability; and analysis of physician profile data.

Patient characteristics, including age, socioeconomic status, comorbid conditions, access to health resources, and health and lifestyle behaviors, have long been understood to have a direct and independent effect on health outcomes (16, 23–32). In our study, younger, better-educated patients who had diabetes of longer duration but slightly better overall health status were more prevalent in endocrinology practices than in generalist practices. Adjustment for these patient characteristics reduced the statistical significance of specialty differences for some of the process and outcome measures. These findings are consistent with those of other recent specialty comparisons where careful adjustment for case mix was performed (6, 33–37). Because case-mix measures are imperfect, however, there may still have been

**Figure.** Proportion of patients with unadjusted hemoglobin A<sub>1c</sub> values less than 0.10 (10%) for each practice site within specialty.



Squares represent the observed mean proportion of patients with hemoglobin A<sub>1c</sub> values less than 0.10 ( $<10\%$ ) seen at each site. Vertical lines represent 95% CIs. Bold horizontal lines represent the mean proportion of patients with hemoglobin A<sub>1c</sub> values less than 0.10 ( $<10\%$ ) for each specialty (76% of patients of generalists and 79% of patients of endocrinologists).

residual unmeasured case-mix differences between specialties. Improvement in these measures may improve quality-of-care comparisons between groups of physicians.

Simple adjustment for patient case-mix alone, however, could still lead to inaccurate inferences about differences in quality of care between groups of physicians. Quality measures must accurately reflect differences between the physician groups being compared and must be reliable at the physician or practice level. Our findings suggest that choosing only process or only outcome measures would have led to different conclusions about the quality of care provided by the two specialties. Specialty comparisons seldom use both process and outcome measures simultaneously (35, 38, 39), which, as our findings suggest, may be more appropriate. We also observed that choosing different thresholds for outcome measures may be important in measuring the effect of the physician versus the patient in quality-of-care assessments. As seen in **Table 4**, we observed a higher inflation factor (5.51) when the specified hemoglobin A<sub>1c</sub> level was set at less than 0.10 (<10%) than when it was set at less than 0.08 (<8%) (inflation factor, 3.90). In addition, patient characteristics were more significantly associated with achievement of lower versus higher hemoglobin A<sub>1c</sub> levels. Thus, higher versus lower thresholds may be a better reflection of the physician's contribution to quality of care. In future studies, comparisons of physicians' care could be improved by combining selected quality measures into multi-item scores (enhancing the reliability and validity of profile scores) and by choosing thresholds for measures that reflect physician (rather than patient) effects more accurately.

Although physicians' behavior may be highly consistent across patients in their practices, their behavior may or may not be similar to that of other physicians in their group (for example, specialty). This is the essence of physician-level clustering. We observed considerable variation between physicians within each of the specialty groups; in other words, physicians tended to behave more like individuals than like the specialty group to which they belonged. These large between-physician differences within the same specialty, coupled with the consistency of physicians' behavior across patients in their practices, increased the variance associated with specialty. Once this physician-level variation was taken into account, the specialty differences were no longer

statistically significant. To see statistically significant differences between specialties given the clustering that we observed, we would have needed to increase the size of the physician sample from 3- to 20-fold, depending on the quality measure being compared.

In quality assessment, comparison of two groups of physicians (such as specialists, those scoring over or under a benchmark value, or those in different care settings) requires both a sufficient number of patients per physician to ensure physician-level reliability and a sufficient number of physicians per comparison group to account for variation with the group. Paradoxically, to obtain a reliable profile of care at the level of an individual physician or a practice, a relatively large sample of patients from that practice might be required. However, having large samples of patients per practice does not ensure optimal comparisons of groups of physicians or practices and may actually substantially increase the number of physicians or practices needed to achieve adequate power for the comparison. To compare generalists and specialists, for example, when adequate numbers of patients are sampled to obtain reliable physician- or practice-level quality-of-care scores, considerably more physicians or practices may be needed to accurately compare the specialty groups. The greater the variability in care across physicians or practices, the larger the number of physicians needed to estimate differences between physician groups, such as specialties. The Appendix shows an example of this relationship. With 1750 patients, our study was designed with 80% power to detect significant differences between physician groups in the presence of the minimal intraclass correlation coefficients that we expected on the basis of previous research (2, 22). For example, for the hemoglobin A<sub>1c</sub> process measure, we had 80% power to detect an unadjusted difference of 6% between specialties. In the presence of the large inflation factor that we observed (8.7), the 11% difference between specialties in this measure was not statistically significant after adjustment for clustering. With the observed inflation factor for this measure, our study had 80% power for detecting only a larger difference (15%) between specialties. Therefore, our findings emphasize the importance of designing quality-of-care profiling studies with sufficient numbers of physicians in any comparison group to detect meaningful differences in the presence of at least moderate physician-level clustering.

Finally, our study underscores the importance of using appropriate statistical techniques to account for both patient case-mix and physician-level clustering when comparing groups of physicians. Because patients within physicians' practices are not independent observations, as assumed by many standard statistical techniques, and because physicians seem to behave more like individuals than like groups, failure to account for physician-level clustering by using an appropriate analytical method could lead to overestimation of the statistical significance of the groups being compared (1, 3, 5, 41–43). In addition to the generalized estimation equations that we used (20), other approaches also exist for analyzing cluster data (44–47).

It is assumed that some study patients might have been “co-managed” by generalists, endocrinologists, and other medical subspecialties and associate providers (for example, nurses, nutritionists, and podiatrists). We did not attempt to determine the level of co-management. We assume that such co-management would enhance patient care through the application of the experience, knowledge, and skills of multiple providers. However, to the extent that extensive co-management occurred, the unadjusted differences between specialties would have been less than those we observed.

Our study has at least two potential limitations. First, participating physicians volunteered, and patients were not randomly sampled within practices. These physicians may overrepresent those who were confident that they were providing high-quality care. Any bias introduced by sampling these practices might therefore result in higher performance on quality measures than would be expected in a more representative sample. However, even among these physicians, the variation between physicians in each specialty was unexpectedly large. Therefore, a broader sample of both physician groups would be unlikely to decrease the variation between physicians and therefore unlikely to yield findings different from those observed.

In addition, although medical record data are commonly assumed to be the “gold standard” for quality assessment, they may not be the optimal data source for this purpose. Physicians tend to record only positive findings and not document clinical care yielding negative findings. To affect our results, however, specialties would have had to differ in recording practices. A recent study suggests that specialties do not differ in this way (48).

In summary, most national quality-of-care initiatives fail to adjust adequately for patient characteristics and none adjust for variations between individual physicians or practices. The failure to take both of these important design and analytical features into account might lead to inaccurate conclusions in comparisons of profiles of quality of care provided by different groups of physicians.

## GLOSSARY

*Case-mix adjustment:* Statistical methods applied to physician comparisons to account for case-mix bias.

*Case-mix bias:* The nonrandom tendency of patients with certain characteristics (such as age, comorbid conditions, or severity of illness) that affect outcomes to choose physicians with certain characteristics (such as specialty). Ignoring this tendency in comparisons of physicians would introduce bias in those comparisons.

*Hierarchical modeling:* Statistical technique of accounting for nested data structures (for example, patients within physician). It does not assume independence of observations, accounts for multiple levels of sources of variability measured at different levels of the hierarchy, and allows estimation of random and fixed effects.

*Inflation factor:* The proportional increase (or decrease) in sample size, reflecting the variation around the group mean or rate for a measure due to differences between individuals in the groups, needed for observed differences between groups to reach statistical significance.

*Intraclass correlation:* With respect to physician profiling, a reflection of the consistency of the physician's behavior across patients (especially if adjusted for patient case-mix) and practices.

*Physician-level clustering:* The nonrandom clustering of patients with certain characteristics within a physician's practice; the variation between physicians in quality measures which, after case-mix adjustment, are attributable to differences in physicians' behavior.

*Physician-level reliability:* Reproducibility of a physician's quality measure (rate of a desired process or outcome) across multiple samples of patients in his or her practice.

*Validity:* The accuracy (or absence of bias) of measurement; the extent to which a measure actually measures what it is intended to measure. (Validity of the physician profile scores is not addressed in this paper.)

## APPENDIX

Following is an example of the relationship among design effects, patient and physician sample sizes, intraclass correlation coefficients, and effective sample sizes to detect differences between physician groups.

Suppose a fully randomized trial comparing two groups of

250 patients each had 80% power to detect a difference in the performance of a quality measure of 20% versus 31% (odds ratio, 1.8) for each of two physician groups, and patients were sampled with attention to the physician's practice from which they were drawn. If instead 20 physicians were sampled and then 25 patients of each physician were sampled (for the same total of 250 patients in each comparison group), the actual power of the study would depend on the intraclass correlation. Both the intraclass correlation and the number of observations in each cluster have implications for the inflation factor and the effective sample size. For example, with a very high intraclass correlation of 0.25, this nested design would have an inflation factor of 7.0 and have only 80% power to detect a much larger difference in the performance of the quality measure (20% vs. 52% [odds ratio, 4.2]) for the two physician groups. With the high intraclass correlation, each patient's response provides less independent information. Thus, adding more patients per physician only marginally increases the study's power. In the hypothetical extreme case of an intraclass correlation of 1.00, the effective sample size for this study would be 20, the number of participating physicians.

Even low intraclass correlation can significantly affect the inflation factor and, thus, the study's power. For example, in the above design, an intraclass correlation of 0.05 would produce an inflation factor of 2.2, resulting in an effective sample size of 227. This study would then have 80% power to detect a difference in the performance of the quality measure of 20% versus 37% (odds ratio, 2.4) for the two physician groups. Therefore, studies must have a sufficient number of patients seen by each physician to estimate physician performance reliably, and to make appropriate case-mix adjustments, and they must have adequate numbers of physicians to have sufficient power to detect relevant differences between physician groups.

From New England Medical Center and Tufts University School of Medicine, Boston, Massachusetts; American Diabetes Association National Office, Alexandria, Virginia; and University of California, San Diego, California.

**Disclaimer:** The views expressed in this manuscript are those of the authors and do not necessarily reflect those of the American Diabetes Association.

**Grant Support:** By the American Diabetes Association, Alexandria, Virginia.

**Acknowledgments:** The authors thank the American Diabetes Association Provider Recognition Steering Committee for guidance and leadership in facilitating this study; the Health Outcomes Institute, particularly David Radosevich, PhD, and Michael Huber, for invaluable assistance and support; Grace A. Connolly, JD, for editorial input; and Elaine Ackerson for help with preparation of the manuscript.

**Requests for Single Reprints:** Sheldon Greenfield, MD, Primary Care Outcomes Research Institute, Tufts University School of Medicine, M&V 1, 136 Harrison Avenue, Boston, MA 02111; e-mail, sheldon.greenfield@tufts.edu.

**Current Author Addresses:** Drs. Greenfield and Kaplan: Primary Care Outcomes Research Institute, Tufts University School of Medicine, M&V 1, 136 Harrison Avenue, Boston, MA 02111.

Dr. Kahn: American Diabetes Association, 1701 North Beauregard Street, Alexandria, VA 22311.

Mr. Ninomiya: 351 Longden Lane, Solana Beach, CA 92075.

Dr. Griffith: Biostatistics Research Center, New England Medical Center, 750 Washington Street, NEMC 063, Boston, MA 02111.

**Author Contributions:** Conception and design: S. Greenfield, S.H. Kaplan, R. Kahn, J.L. Griffith.

Analysis and interpretation of the data: S. Greenfield, S.H. Kaplan, R. Kahn, J. Ninomiya, J.L. Griffith.

Drafting of the article: S. Greenfield, S.H. Kaplan, R. Kahn.

Critical revision of the article for important intellectual content: S. Greenfield, S.H. Kaplan, R. Kahn, J.L. Griffith.

Final approval of the article: S. Greenfield, S.H. Kaplan, R. Kahn.

Provision of study materials or patients: R. Kahn.

Statistical expertise: S.H. Kaplan, J. Ninomiya, J.L. Griffith.

Obtaining of funding: S. Greenfield, R. Kahn.

Administrative, technical, or logistic support: R. Kahn, J. Ninomiya.

Collection and assembly of data: R. Kahn, J. Ninomiya.

## References

1. Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). *J Health Serv Res Policy*. 2000;5:12-6. [PMID: 10787581]
2. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA*. 1999;281:2098-105. [PMID: 10367820]
3. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol*. 1981;114:906-14. [PMID: 7315838]
4. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics*. 1998;47:95-113.
5. Divine GW, Brown JT, Frazier LM. The unit of analysis error in studies about physicians' patient care behavior. *J Gen Intern Med*. 1992;7:623-9. [PMID: 1453246]
6. Greenfield S, Rogers W, Mangotich M, Carney MF, Tarlov AR. Outcomes of patients with hypertension and non-insulin dependent diabetes mellitus treated by different systems and specialties. Results from the medical outcomes study. *JAMA*. 1995;274:1436-44. [PMID: 7474189]
7. Harrold LR, Field TS, Gurwitz JH. Knowledge, patterns of care, and outcomes of care for generalists and specialists. *J Gen Intern Med*. 1999;14:499-511. [PMID: 10491236]
8. Streja DA, Rabkin SW. Factors associated with implementation of preventive care measures in patients with diabetes mellitus. *Arch Intern Med*. 1999;159:294-302. [PMID: 9989542]
9. Donohoe MT. Comparing generalist and specialty care: discrepancies, deficiencies, and excesses. *Arch Intern Med*. 1998;158:1596-608. [PMID: 9701093]

10. Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A*. 1999;159:385-444.
11. Normand SL, Glickman M, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association*. 1997;92:803-14.
12. McNeil BJ, Pedersen SH, Gatsonis C. Current issues in profiling quality of care. *Inquiry*. 1992;29:298-307. [PMID: 1398901]
13. Ho M, Marger M, Beart J, Yip I, Shekelle P. Is the quality of diabetes care better in a diabetes clinic or in a general medicine clinic? *Diabetes Care*. 1997;20:472-5. [PMID: 9096962]
14. Reilly M, Pepe M. The relationship between hot-deck multiple imputation and weighted likelihood. *Stat Med*. 1997;16:5-19. [PMID: 9004380]
15. Rubin HR, Gandek B, Rogers WH, Kosinski M, McHorney CA, Ware JE Jr. Patients' ratings of outpatient visits in different practice settings. Results from the Medical Outcomes Study. *JAMA*. 1993;270:835-40. [PMID: 8340982]
16. Greenfield S, Sullivan L, Dukes KA, Silliman R, D'Agostino R, Kaplan SH. Development and testing of a new measure of case mix for use in office practice. *Med Care*. 1995;33:AS47-55. [PMID: 7723461]
17. Statistical Analysis System. Version 8. Cary, NC: SAS Institute; 2000.
18. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common [Letter]. *BMJ*. 1998;317:1318. [PMID: 9804732]
19. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *N Engl J Med*. 1999;341:279-83. [PMID: 10413743]
20. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med*. 1998;17:1261-91. [PMID: 9670414]
21. Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addict Behav*. 1997;22:1-12. [PMID: 9022867]
22. Greenfield S, Kaplan SH, Silliman RA, Sullivan L, Manning W, D'Agostino R, et al. The uses of outcomes research for medical effectiveness, quality of care, and reimbursement in type II diabetes. *Diabetes Care*. 1994;17 Suppl 1:32-9. [PMID: 8088221]
23. Klein R, Klein BE, Moss SE, Cruickshanks KJ. Relationship of hyperglycemia to the long-term incidence and progression of diabetic retinopathy. *Arch Intern Med*. 1994;154:2169-78. [PMID: 7944837]
24. Nathan DM, McKittrick C, Larkin M, Schaffran R, Singer DE. Glycemic control in diabetes mellitus: have changes in therapy made a difference? *Am J Med*. 1996;100:157-63. [PMID: 8629649]
25. Wagner EH, Austin BT, Von Korff M. Improving outcomes in chronic illness. *Manag Care Q*. 1996;4:12-25. [PMID: 10157259]
26. Nathan DM, Singer DE, Godine JE, Harrington CH, Perlmutter LC. Retinopathy in older type II diabetics. Association with glucose control. *Diabetes*. 1986;35:797-801. [PMID: 3721064]
27. Lloyd CE, Wing RR, Orchard TJ, Becker DJ. Psychosocial correlates of glycemic control: the Pittsburgh Epidemiology of Diabetes Complications (EDC) Study. *Diabetes Res Clin Pract*. 1993;21:187-95. [PMID: 8269821]
28. Pringle M, Stewart-Evans C, Coupland C, Williams I, Allison S, Sterland J. Influences on control in diabetes mellitus: patient, doctor, practice, or delivery of care? *BMJ*. 1993;306:630-4. [PMID: 8461816]
29. Kaplan GA, Keil JE. Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation*. 1993;88:1973-98. [PMID: 8403348]
30. Winkleby MA, Fortmann SP, Barrett DC. Social class disparities in risk factors for disease: eight-year prevalence patterns by level of education. *Prev Med*. 1990;19:1-12. [PMID: 2320553]
31. Harris MI, Eastman RC, Cowie CC, Flegal KM, Eberhardt MS. Racial and ethnic differences in glycemic control of adults with type 2 diabetes. *Diabetes Care*. 1999;22:403-8. [PMID: 10097918]
32. Zaslavsky AM, Hochheimer JN, Schneider EC, Cleary PD, Seidman JJ, McGlynn EA, et al. Impact of sociodemographic case mix on the HEDIS measures of health plan quality. *Med Care*. 2000;38:981-92. [PMID: 11021671]
33. Chen J, Radford MJ, Wang Y, Krumholz HM. Care and outcomes of elderly patients with acute myocardial infarction by physician specialty: the effects of comorbidity and functional limitations. *Am J Med*. 2000;108:460-9. [PMID: 10781778]
34. Frances CD, Go AS, Dauterman KW, Deosarasingh K, Jung DL, Gettner S, et al. Outcome following acute myocardial infarction: are differences among physician specialties the result of quality of care or case mix? *Arch Intern Med*. 1999;159:1429-36. [PMID: 10399894]
35. Ayanian JZ, Guadagnoli E, McNeil BJ, Cleary PD. Treatment and outcomes of acute myocardial infarction among patients of cardiologists and generalist physicians. *Arch Intern Med*. 1997;157:2570-6. [PMID: 9531225]
36. Carey TS, Garrett J, Jackman A, McLaughlin C, Fryer J, Smucker DR. The outcomes and costs of care for acute low back pain among patients seen by primary care practitioners, chiropractors, and orthopedic surgeons. The North Carolina Back Pain Project. *N Engl J Med*. 1995;333:913-7. [PMID: 7666878]
37. Poses RM, McClish DK, Smith WR, Huber EC, Clemo FL, Schmitt BP, et al. Results of report cards for patients with congestive heart failure depend on the method used to adjust for severity. *Ann Intern Med*. 2000;133:10-20. [PMID: 10877735]
38. Mitchell JB, Ballard DJ, Whisnant JP, Ammering CJ, Samsa GP, Matchar DB. What role do neurologists play in determining the costs and outcomes of stroke patients? *Stroke*. 1996;27:1937-43. [PMID: 8898795]
39. Kaplan SH, Greenfield S, Gandek B, Rogers WH, Ware JE Jr. Characteristics of physicians with participatory decision-making styles. *Ann Intern Med*. 1996;124:497-504. [PMID: 8602709]
40. Jollis JG, DeLong ER, Peterson ED, Muhlbaier LH, Fortin DF, Califf RM, et al. Outcome of acute myocardial infarction according to the specialty of the admitting physician. *N Engl J Med*. 1996;335:1880-7. [PMID: 8948564]
41. Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomization trials. *J Clin Epidemiol*. 1993;46:123-31. [PMID: 8437028]
42. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am J Epidemiol*. 1994;140:279-89. [PMID: 8030631]
43. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health*. 1995;85:1378-83. [PMID: 7573621]
44. Carr GJ, Portier CJ. An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicology data. *Biometrics*. 1993;49:779-91. [PMID: 8241373]
45. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics*. 1984;40:961-71. [PMID: 6534418]
46. Gatsonis CA, Epstein AM, Newhouse JP, Normand SL, McNeil BJ. Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction. An analysis using hierarchical logistic regression. *Med Care*. 1995;33:625-42. [PMID: 7760578]
47. Binder D. On the variances of asymptotically normal estimators from complex surveys. *International Statistics Review*. 1983; 51:279-92.
48. Solomon DH, Schaffer JL, Katz JN, Horsky J, Burdick E, Nadler E, et al. Can history and physical examination be used as markers of quality? An analysis of the initial visit note in musculoskeletal care. *Med Care*. 2000;38:383-91. [PMID: 10752970]