The Unreliability of Individual Physician "Report Cards" for Assessing the Costs and Quality of Care of a Chronic Disease

Timothy P. Hofer, MD, MS
Rodney A. Hayward, MD
Sheldon Greenfield, MD
Edward H. Wagner, MD
Sherrie H. Kaplan, PhD
Willard G. Manning, PhD

Context Physician profiling is widely used by many health care systems, but little is known about the reliability of commonly used profiling systems.

Objectives To determine the reliability of a set of physician performance measures for diabetes care, one of the most common conditions in medical practice, and to examine whether physicians could substantially improve their profiles by preferential patient selection.

Design and Setting Cohort study performed from 1990 to 1993 at 3 geographically and organizationally diverse sites, including a large staff-model health maintenance organization, an urban university teaching clinic, and a group of private-practice physicians in an urban area.

Participants A total of 3642 patients with type 2 diabetes cared for by 232 different physicians.

Main Outcome Measures Physician profiles for their patients' hospitalization and clinic visit rates, total laboratory resource utilization rate and level of glycemic control by average hemoglobin A_{1c} level with and without detailed case-mix adjustment.

Results For profiles based on hospitalization rates, visit rates, laboratory utilization rates, and glycemic control, 4% or less of the overall variance was attributable to differences in physician practice and the reliability of the median physician's case-mix–adjusted profile was never better than 0.40. At this low level of physician effect, a physician would need to have more than 100 patients with diabetes in a panel for profiles to have a reliability of 0.80 or better (while more than 90% of all primary care physicians at the health maintenance organization had fewer than 60 patients with diabetes). For profiles of glycemic control, high outlier physicians could dramatically improve their physician profile simply by pruning from their panel the 1 to 3 patients with the highest hemoglobin A_{1c} levels during the prior year. This advantage from gaming could not be prevented by even detailed case-mix adjustment.

Conclusions Physician "report cards" for diabetes, one of the highest-prevalence conditions in medical practice, were unable to detect reliably true practice differences within the 3 sites studied. Use of individual physician profiles may foster an environment in which physicians can most easily avoid being penalized by avoiding or deselecting patients with high prior cost, poor adherence, or response to treatments.

JAMA. 1999;281:2098-2105 www.jama.com

PROVIDER PROFILING IS NOW widely practiced or attempted in many health care systems.[1-4] Those who are paying for or "managing" health care are seeking ways to make health care providers more accountable for both the cost and quality of the care that they supply. Reports comparing hospital mortality rates across hospitals and the Health Plan Employer Data and Information Set performance measures across health plans have been widely implemented.[3,5] The goal of the more recent attempts at physician profiling is to hold a single individual (the physician) accountable for what happens to a specific group of patients.[3,6-8] Developing and disclosing profiles to consumers is a crucial element of the proposed Consumer Bill of Rights and Responsibilities[9] and is considered an important part of creating efficient health care markets.[9] Surveys have shown that up to 80% of group practices with capitated patients profile the resource utilization of their physicians.[1] Furthermore, for many hospitals and managed care organizations

For editorial comment see p 2142.

Author Affiliations: Veterans Affairs Center for Practice Management and Outcomes Research, and Department of Internal Medicine, University of Michigan, Ann Arbor (Drs Hofer and Hayward); Primary Care Outcomes Research Institute, New England Medical Center, Boston, Mass (Drs Greenfield and Kaplan); Center for Health Studies, Group Health Cooperative, Seattle, Wash (Dr Wagner); and Department of Health Studies, University of Chicago, Chicago, Ill (Dr Manning).
Corresponding Author and Reprints: Timothy P. Hofer, MD, MS, Ann Arbor VA HSR&D (US Mail), 3rd Floor, Lobby L, PO Box 130170, Ann Arbor, MI 48113 (e-mail: thofer@umich.edu).

individual physician profiling has become an integral part of medical staff appointment and issuing clinical privileges.[8]

Although increasing professional accountability is a laudable goal, profiling ventures can be quite expensive (adding as much as $0.59 to $2.17 per member per month),[9] may have very serious consequences for physicians and could potentially be harmful to patients through distortion of physician incentives.[4,6,8] In their most benign use, profiles are used simply to educate physicians about their practices relative to their peers. However, the most severe consequence of poor performance on a profile is the potential loss of a managed care contract or hospital admitting privileges.[8]

The usefulness of physician profiles depends on their reliability and accuracy. Even so, little published evidence on the reliability of any of the commonly used profiling systems exists.[10] In part, this lack of attention to reliability comes from a lack of data. Few profilers have data with adequate risk adjustment measures, enough observations per provider, and enough providers to try to estimate the true reliability of their measures. The Diabetes Patient Outcomes Research Team study, which examined the type of health care provided to patients with type 2 diabetes living in 3 geographically different locations and enrolled organizationally diverse sites, allows for the opportunity to evaluate the reliability of provider profiling for one of the most common chronic medical conditions. We examined the utility of several potential profile measures, including hospitalization rates, visit rates, total laboratory resource use, and glycemic control. We evaluated (1) the reliability of physician profiles, (2) the importance of detailed case-mix adjustment, (3) the clinical and economic magnitude of variations in physicians' practices, and (4) the potential for physicians to improve their profile measures by gaming the system (ie, through patient selection rather than changing their practice).

METHODS

Sites and Patient Sample

The study sample included 3642 patients with diabetes cared for at 3 geographically and organizationally distinct sites: (1) a staff-model health maintenance organization (HMO) on the West Coast, (2) an urban university teaching clinic in the Midwest, and (3) a group of private practice physicians in a New England urban area.[11] Patients were eligible if they were older than 30 years and were prescribed either insulin or sulfonylureas, or if they met 1 of the following laboratory criteria: (1) a fasting plasma glucose level greater than 7.8 mmol/L (>140 mg/dL), (2) a random plasma glucose level greater than 11.1 mmol/L (>200 mg/dL), or (3) a glycosylated hemoglobin (HbA_{1c}) level greater than 3 SDs from the mean.

At 2 sites (HMO and private practice), physicians were randomly selected (within age and sex strata), but a universal sample of primary care clinic physicians were selected at the third site. Physicians reviewed a list of eligible patients with diabetes under their care who were identified by means of pharmacy and laboratory databases (at the HMO and teaching clinic sites) or from their records (private practice site). The physicians excluded patients who did not have type 2 diabetes, were too ill for follow-up, or who either did not speak English or did not have a family interpreter (9% of the patients were excluded). Of the patients who were contacted by telephone and requested to participate in the study, 18% declined. An additional 15% failed to return their baseline surveys. The final cohort consisted of 3642 patients, 1730 from the HMO site, 787 from the university teaching clinic site, and 1125 from the private practice site, who met the eligibility criteria described above. The 232 physicians across all 3 sites had an average of 16 patients who responded to the survey, with physicians averaging 21 patients at the HMO site, 9 patients at the urban teaching hospital, and 18 patients at the private practice site. Less than 5% of the physicians were endocrinologists. Patients were asked in the baseline survey to confirm that we had correctly identified the physician primarily responsible for their diabetes care.

At the HMO site, extensive information was available from medical information systems including hospital discharge and visit records and clinical laboratory and pharmacy systems.[12] Analyses that used medical information systems–based measures such as total laboratory utilization and the results of HbA_{1c} tests are thus limited to this site.

Variables

Dependent Variables. Resource utilization measures were collected from a self-administered survey and included total hospitalizations and total number of office visits in the previous 6-month period. Laboratory utilization measures were constructed at the HMO site where we had access to all of the laboratory records. Each test was mapped to a relative value unit constructed to reflect actual laboratory costs (in 1992 dollars) for performing each test.[12,13] Diabetes control in 1991 and 1992 was based on the average HbA_{1c} level each year at the HMO site.[12]

Independent Variables. The variables used to predict utilization and diabetes control included demographic variables (patient sex and age), physician and site, socioeconomic status (including income, education, and employment status), duration of diabetes, and health status measures. We measured generic health status using the 36-item Medical Outcomes Study Short-Form Health Survey and comorbidity using a previously described and validated diabetes-specific instrument, the Total Illness Burden Index.[12,14] This measure uses what patients have reported about all of their diseases and symptom intensity to characterize the total disease burden, including the presence and severity of diabetic complications.

Analysis

Our overall analytic strategy was to construct case-mix–adjusted regression models for the resource utilization and glucose control–dependent variables.

The case-mix–adjusted residuals from these models were then examined for how much they varied systematically by physician. This 2-step approach represents current profiling practice.[3,15] However, we confirmed our results using hierarchical regression for general linear models, an analytic technique less commonly used but more appropriate for this instance, for which patient observations are not independent but are clustered by physician.[16]

In the first set of analyses, case-mix adjustment included only age and sex, which are often the only patient characteristics that can be easily obtained from insurance company databases. Next we developed a full case-mix model, which included all of the independent variables described above. We used linear regression for the continuous dependent variables (laboratory relative value units and HbA_{1c} levels) and negative binomial regression for counts (hospitalizations and visits).[17] Variables for each site were included to remove site effects. In the 2-step approach, we looked for evidence of physician practice effects by analyzing whether residuals from these case-mix–adjusted models were associated with the physician identifier. We used random effects analysis of variance to quantify the variance between and within physician panels. In the hierarchical model approach, the physician level variance (along with its SD) is estimated as a separate parameter in the full case-mix model.

Quantifying the Amount of Physician Variation. The usual assessment of the amount of variation in physician practice patterns simply tabulates the average patient residual after case-mix adjustment by physician.[15] However, even with the most sophisticated and expensive case-mix adjustment, the above analysis will overstate (substantially in some cases) the true range of a physician practice effect. This is particularly true when the patient panel size is small or the physician effect on the profile is relatively small.[16,18-21] To account for a small signal (in this case, physician effect) to noise (variation due to unmeasured patient factors) ratio, a "shrinkage factor" must be applied to each physician's estimated profile. (A mathematical model describing this concept in the setting of constructing a physician profile measure may be obtained from the authors.)

Table 1. Demographics of Patients With Diabetes by Site*

Characteristics	HMO	Urban Teaching Hospital	Private Practice Physicians
Age, mean, y	63	61	62
Education			
<High school graduate	17	69	19
High school graduate	36	22	38
Some college	24	8	21
College or graduate	23	2	23
Employed	41	15	42
Sex, female	51	76	53
Income, $			
<15 000	18	94	28
15 000-29 000	36	5	27
≥30 000	46	1	45
Married	70	25	65
Ethnicity			
White	90	35	94
Black	4	64	4
Hispanic	1	<1	1
Asian	4	<1	1
Other	1	1	<1

*The sample size at the health maintenance organization (HMO) site was 1738, 790 at the urban teaching hospital, and 1132 at the private practice site. All data are percentages unless otherwise indicated.

The shrinkage factor can be thought of as adjusting the performance measure for the level of reliability. The lower the reliability, the more the physician's measure should be shrunk toward the mean of all the physicians. A statistic that describes the proportion of overall variation attributable to physician practice, after accounting for the shrinkage described above, is the intraclass correlation coefficient (ICC).[16,22] The ICC can be thought of as the maximum proportion of variance that can be explained by physician practice patterns[23] or as a reliability-adjusted R^2. It was calculated from the estimates of variance between and within physician panels.

Quantifying the Reliability of a Physician Measure. Reliability is the extent to which a measure gives the same result on repeated trials. Practically, reliability is often calculated as the correlation between 2 equivalent (but not identical) measures, such as items on a test scale.[23] For a report card measure of physician effect on glucose control, for example, each patient's level of glucose control, after case-mix adjustment, is considered an equivalent measure of the physician's effect on glucose control. The correlation between patient measurements within physician (the intra-class correlation coefficient or ICC) is thus the reliability of an estimate of the physician effect on glucose control based on a single patient measurement. The reliability of a physician profile, composed of a mean of n patient measurements is calculated as a function of n and the ICC using the Spearman-Brown prophecy formula.[23,24] A reliability of 0.80 suggests that 80% of the variation of an individual physician's profile is due to practice differences and 20% is due to chance variation and is often considered the minimum level necessary for making decisions about individual physicians based on a profile.[25] The reliability of a profile increases as the physician panel size increases and as the difference in practice patterns between physicians becomes larger.

Manipulating Profiles. Proponents of profiles argue that good case-mix ad-

justment can eliminate the advantage of caring for less sick panels of patients. For profiles of glycemic control at the staff-model HMO, we tested the effect of deliberate patient selection. We collected average HbA_{1c} levels of each patient for the years 1991 and 1992. After excluding physicians with fewer than 4 study patients, we calculated physician profiles for 1991 data and identified as outliers the 10% of physicians with the worst level of glycemic patient control after case-mix adjustment. (While there is no standard cutoff, those generating profiles frequently exclude profiles for providers with particularly small panel sizes.) Before estimating the 1992 profiles, we dropped from the panels of the physicians who were outliers in 1991 the few patients with the worst glycemic control in 1991 (above the 95th percentile overall) replacing them with patients who had average control in 1991. This simulates the effect of a physician eliminating the few (1-3) patients with the worst glycemic control as a strategy to improve their profile.

All of the above analyses were performed using the Stata Statistical Software Package[22,26] with the exception of the hierarchical models, which were estimated using MLwiN multilevel modeling software.[27]

RESULTS

Differences in Demographics and Profile Measures Between Sites

TABLE 1 shows the demographic characteristics of the patients with diabetes at each of the 3 sites. While age distributions were similar, there was a much larger percentage of black patients at the urban teaching clinic (64% vs 4%), as well as a lower percentage of married patients. Patients at the urban teaching clinic also had lower income, less education, and lower employment rates than those at the other 2 sites.

As shown in TABLE 2, the urban university teaching clinic site had almost twice the average number of hospitalizations per patient per year as the HMO site after full case-mix adjustment. The private practice site had about 3 patient visits per year more than the other 2 sites, a difference that remained significant after full case-mix adjustment (Table 2).

Impact of Reliability and Case Mix on Physician Visit and Hospitalization Rates

The profiling approach most commonly used by payers and administrators is to calculate simple age- and sex-adjusted measures that are averaged by physician to generate a physician profile. Using this approach, it would appear that 13% of the variation in outpatient visit rates and 8% of the variation for hospitalization rates are attributable to the physician (by an analysis of variance of the age- and sex-adjusted residuals, TABLE 3). However, accounting for reliability, particularly the low signal (physician practice effect) to noise (patient variability) ratio, and more extensive case-mix adjustment, a better estimate of the maximum possible amount of variation due to differences among physicians is closer to 4% for outpatient visits and 1% for hospitalizations (as represented by the ICC, Table 3). This implies that after adjusting for socioeconomic status, comorbidity, and health

Table 2. Variation in Hospitalization and Outpatient Visit Rates for Diabetic Patients Across the 3 Practice Sites*

	Mean Levels by Site†			Explained Variance (R^2)‡	
Variable	HMO	Urban Teaching Hospital	Private Practice Physician	Adjustment for Patient Characteristics§	Adding Site Available
Outpatient visits per year					
Unadjusted	9.4	9.4	11.3‖	. . .	0.00
Age- and sex-adjusted	9.4	9.2	11.4‖	0.00	0.00
Full case-mix	8.8	8.1	11.9‖	0.01	0.02
Hospitalizations, % with >1					
Unadjusted	11	21‖	12	. . .	0.01
Age- and sex-adjusted	11	21‖	12	0.00	0.02
Full case-mix adjusted	8	17‖	13‖	0.08	0.09

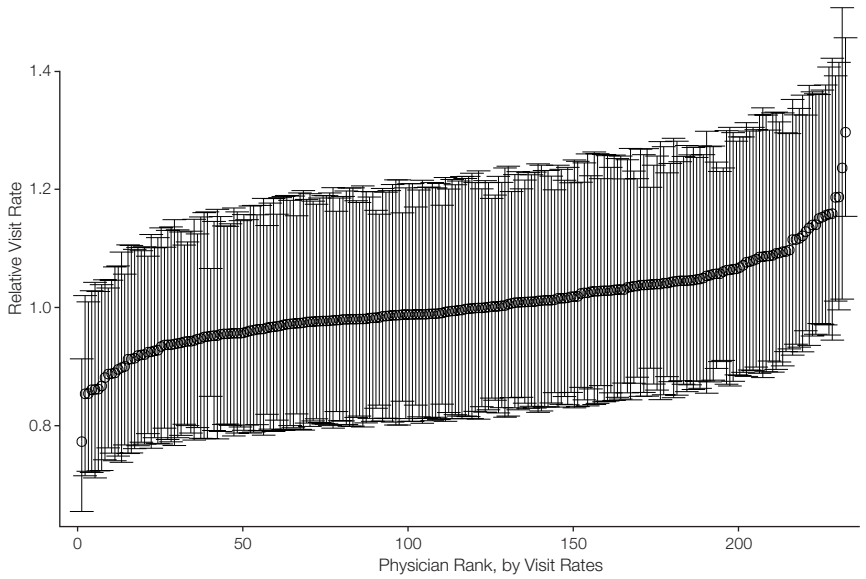
*Ellipses indicate not applicable.
†Adjusted levels obtained by setting covariates to mean value.
‡For these models pseudo-R^2s (which are on a 0-to-1 scale between a constant-only model and perfect prediction) are presented.
§The adjustment for patient characteristics, as shown in the row labels on the left, is either age- and sex-adjustment with only demographic and gender variables, or full case-mix adjustment with demographic, socioeconomic status, comorbidity, and health status covariates included. The pseudo-R^2 for each of these models is shown without site dummy variables (in this column) and with dummy variables for site (in the column at the far right). Thus, the difference in R^2 going down the column reflect the addition of more patient case-mix adjusters and the differences in R^2 going across represent the addition of dummy variables for site.
‖Differences from health maintenance organization (HMO) site significance at $P<.001$.

Table 3. Amount of Variation in Hospitalization and Outpatient Visit Rates Attributable to a Physician Practice Style Effect

Variable	Age- and Sex-Adjusted	Case-mix Adjusted
Variation associated with physician		
Visits, %		
Unadjusted for reliability, R^2*	13	10
Reliability adjusted, ICC†	7	4
Reliability of physician visit rate‡	0.51	0.41
Hospitalizations, %		
Unadjusted for reliability, R^2‡	8	8
Reliability adjusted, ICC§	2	1
Reliability of physician visit rate‡	0.24	0.17

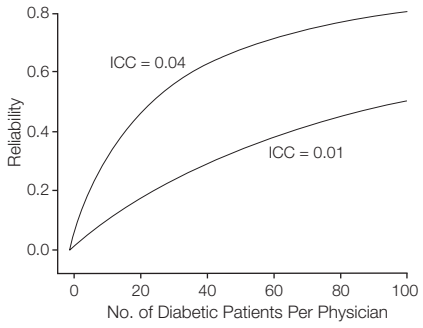
*R^2 from an analysis of variance with physician identifier as independent variables, the usual estimate of the amount of variation in the dependent variable explained by physician.
†Intraclass correlation coefficient (ICC) from a 1-way random effects analysis of variance with physician identifier as the independent variable. This gives an estimate of the reliability of a physician profile composed of a single patient and also represents an upper bound of the amount variation in overall resource use that can be explained by physician practice variation. It can thus be thought of as a reliability-adjusted R^2.
‡The reliability of a physician profile based on 16 patients with diabetes (the average panel size across all 3 sites). As the panel size goes up, the reliability of the profile will increase (see Figure 2). Data are based on the number of visits per year.

Figure 1. Comparison of Physicians' Visit Rate Profiles



Relative visit rate by physician (with 1.0 being an average profile) after adjustment for patient demographic and detailed case-mix measures. The error bars represent a 1.4 SE confidence interval (CI), so that overlapping CIs suggest that the difference between 2 physician visit rates is not statistically different ($P>.05$).[16] In this graph, although the overall physician effect on visits is statistically significant (see "Results" section), it is not possible to say that the physicians at the extremes are significantly different in their visit rates from any of the other physicians.

Figure 2. Reliability by Panel Size for 2 Levels of Physician Effect



This figure illustrates the reliability (on the y axis) of a physician profile measure constructed by averaging a patient utilization or laboratory measure across the number of patients with diabetes in a physician's panel (shown on the x axis). Two lines are shown representing 2 levels of the overall proportion of variance (represented by the intra-class correlation coefficient [ICC]) in a profiled measure accounted for by differences in physician practice found in the data. Thus for an ICC of 4% (the level of physician practice effect for patient visit rates), if a physician's panel includes 100 patients with diabetes, the reliability of the physician's visit rate score would be 0.80.

status, 99% of the variation in hospitalization rates and 96% of the variation in outpatient visit rates is due to unmeasured patient factors or chance. Eliminating all physician variation would have a relatively small effect on overall patient visit rates or hospitalization rates.

Using hierarchical analyses, we can test in a more appropriate statistical model whether this remaining physician effect is statistically significant. After accounting for both the low signal-to-noise ratio of the physician effect and full case-mix adjustment, the physician level variation in visit rates is still statistically significant (χ^2 6.6; $P = .01$). In contrast to the case of outpatient visits, the physician effect for hospitalization rates is not statistically significant ($P = .13$).

Although the overall physician practice effect for visit rates is statistically significant, the reliability of an individual physician visit rate (based on a panel with 16 patients) is only 0.40 (Table 3). Thus, more than 60% of the variation in the median physician's patient-visit–rate profile (1 – reliability) is due to error from chance variation. FIGURE 1 illustrates that this level of physician variation and panel sizes result in physician profile measures that are inadequately precise to distinguish physicians from one another. The vast majority of visit rates are within 10% (or 1 visit per year) of the mean. Even at the extremes, the confidence limits are wide enough that it would be difficult to say that any physician has a rate significantly different from any of the average physicians. At this level of physician effect (ICC = 4%), physicians would need to have more than 100 patients with diabetes for the profiles to have a reliability of 0.80 (FIGURE 2). At the HMO site, we could identify the total number of patients with diabetes per physician via medical information systems, even if the patients were not enrolled in the study. There, the median primary care physician had 29 total patients with diabetes, 90% had fewer than 60, and no physician (out of more than 250 at the HMO) had more than 85.

Profiling Laboratory Resource Use and Glycemic Control

Physician visit and hospitalization rates may be influenced by many factors and perhaps are less determined by physician practice than profiles based on measures more closely tied to specific physician interventions. Thus, we next examined whether profiles of total laboratory utilization and the average level of glycemic control, constructed from computerized laboratory data available only at the HMO site, would be more reliable measures of physician practice differences.

For age- and sex- adjusted total laboratory costs (laboratory relative value units), physician practice style appears to account for 7% of the variation in laboratory costs. However, after full case-mix and reliability adjustment, the ICC was only 2.6%, in the middle of the 1% to 4% ICC range found for hospitalizations and visits, respectively. The reliability of the median physician's profile (21 study patients per physician at the HMO) was 0.38 after full case-mix adjustment.

The percentage of variation accounted for by physician practice differences or ICC can be difficult to interpret at a practical or intuitive level. A more direct il-

lustration of the magnitude of the physician effect on total laboratory cost and how the usual physician profiles shrink dramatically after adjusting for reliability is shown in FIGURE 3. The usual physician profiles (shown on the left side of Figure 3) would suggest that if the physicians in the 90th percentile of utilization were able to change their practice habits to be similar to the practice of physicians in the 10th percentile, the HMO could realize an annual savings of $113 for each of its patients with diabetes. However, after adjusting the estimates for panel size and reliability, the potential annual cost savings resulting from changing the highest to the lowest decile is closer to $40 per patient.

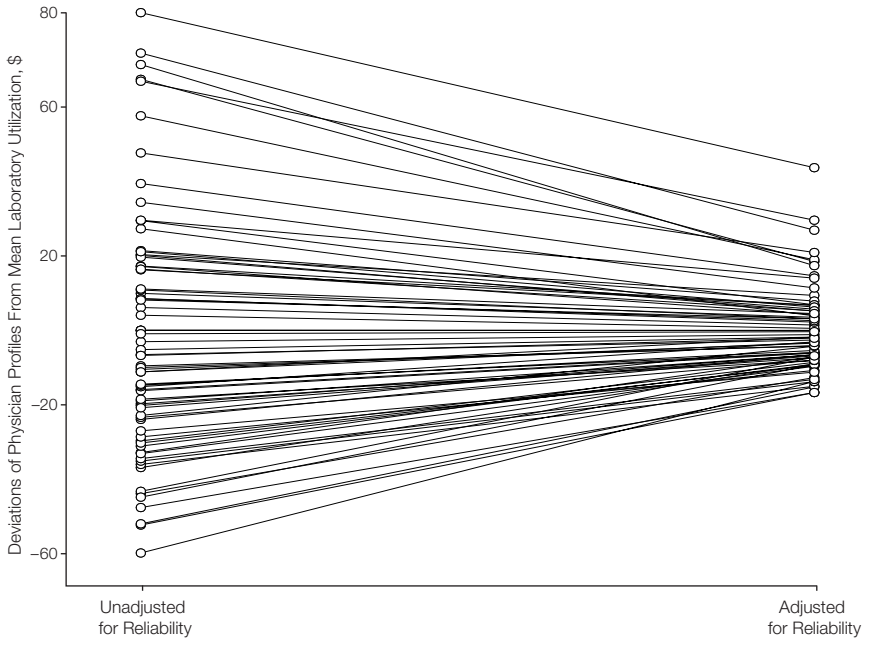
Similarly, for profiles that measure the level of patients' glycemic control, when applying a simple age- and sex-adjusted model, the physician identifier appears to account for 8% of the total variance in level of control (or a range of 2 percentage points between the average HbA_{1c} level achieved by patients of physicians in the 10th percentile vs those in the 90th percentile). However, the reliability and case-mix adjusted physician effect is only 3.3%. Thus, the better estimate of the true range of physician practice between the 10th and the 90th percentiles is 0.7 percentage points and the reliability of an individual physician's HbA_{1c} profile is 0.38 (for physicians with 21 patients).

Does Full Case-Mix Adjustment Prevent Gaming the System? Ideally, full case-mix models would eliminate or reduce the perverse incentive for physicians to manipulate profiles by electing not to care for sick patients. However, in FIGURE 4, we see if those physicians with the worst profiles (patients with higher than expected HbA_{1c} levels) for 1991 managed to discourage the patients with the top 5% of HbA_{1c} levels (representing only 1-3 patients per physician) from returning to their panel, they would in most cases achieve a panel HbA_{1c} profile in 1992 that would be substantially improved than average. About half of this improvement occurs through regression toward the mean (determined by examining 1991 and 1992 profiles without any patient selection for the 1991 outlier physicians) but the other half was due to patient selection. Thus, the patient's HbA_{1c} levels from the previous year proved a far better predictor of what a patient's HbA_{1c} level would be in the current year, better than physician practice or our case-mix adjusters. Manipulating their patient pool, based on a patient's prior year HbA_{1c} level, is the easiest way for physicians to have a substantial improvement in their profile.

COMMENT

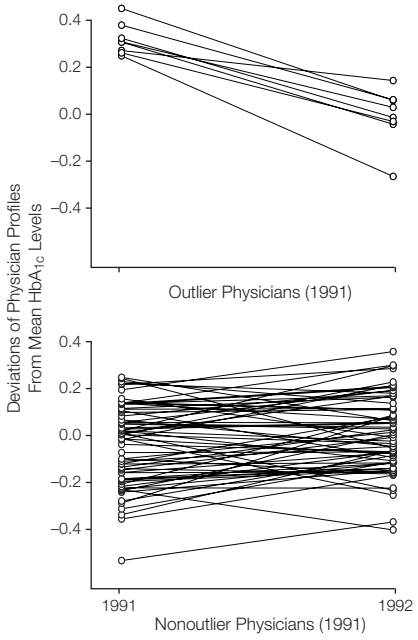
If profiling is to be successful at providing consumers and purchasers of health care services with a way to monitor the quality of health care, then pro-

Figure 3. Adjusting Physician Laboratory Utilization Profiles for Reliability at the HMO Site



The usual approach to reporting physician profiles is shown on the left. Each point represents the amount laboratory costs of patients who have diabetes deviates on average from the mean of all physicians (in US dollars per patient per year). The profile adjusted for reliability, or the signal (physician effect) to noise (patient chance variation) ratio, is shown on the right. The lines illustrate what happens to each physician's profile when adjusted for reliability.

Figure 4. Gaming the System-Profiles of HbA_{1c} Levels



The effect on the outlier physician profiles for 1992, in the upper panel, even after full case-mix adjustment, if they selectively chose not to treat, in 1992, the 1 to 3 patients with diabetes whose 1991 year hemoglobin (HbA_{1c}) levels were higher than the 95 percentile and replaced them in their panel with a patient who had an average 1991 HbA_{1c} level. The lower panel shows the profile changes between 1991 and 1992 for the physicians who were not outliers in 1991.

files must cover more than the delivery of preventive health services to healthy people, such as vaccinations and cancer screening rates, which were the early focus of such profiling efforts as the Health Plan Employer Data and Information Set.[3,5] Diabetes would seem to be an ideal chronic condition to monitor because the disease causes major morbidity, is quite common, and is expensive to treat. Furthermore, in contrast to some areas in medicine, there is good evidence that diabetes care interventions can substantially affect patient outcomes and complications.[28]

Although we found some differences between our 3 sites that were not attributable to case-mix differences, relatively little of the variation in any of the resource utilization or glycemic control measures evaluated was due to individual physician practice style variation. The usual approach to generating physician profiles (averaging case-mix–adjusted patient measures within each physician's panel) exaggerates both the magnitude of physician practice differences and the savings that could be achieved by correcting the practice of the outlier physicians to that of the average physician. Even if this amount of variation at the physician level is considered to be clinically important, the profiles of individual physicians are not very reliable. This lack of reliability is due to a combination of the small physician effect (relative to the substantial patient variations only a small portion of which is captured by even detailed case-mix measures) and the size of individual physician patient panels.

It is possible that at other sites physician practice effects might be more dramatic. We sampled patients from only 3 sites. However, these sites represent the different types of organizations providing health care now, and the patients in our study have a wide range of socioeconomic backgrounds. We did not study retinal or foot examination rates, which are considered to be important measures of quality of diabetes care. As such examinations reflect a fairly specific process of care, it is possible that they may have a larger component of physician variation.[29] Our results relate to only 1 disease. Again, it is possible that, for other diseases, differences among physicians may be larger. However, diabetes is one of the most common diseases in the United States. Apart from hypertension, it is difficult to imagine that there would be enough cases per primary care physician to construct disease-specific profiles for almost any other chronic condition. At the HMO site, none of the more than 250 primary care physicians had more than 85 patients with diabetes. For any of the measures that we examined, at least 100 patients would be needed to reach 80% reliability (often considered the minimum for making decisions about individuals).[25] Finally, our hospital and visit-rate data are by patient self-report, but the findings are similar for our other measures of use and laboratory costs, which are based on computerized records.

How do our findings for diabetes compare with the findings of other diseases, which have been examined specifically for physician-level variation?

For inpatient resource utilization at 1 academic center, attending and resident physicians accounted for only about 2% of the variation in total resource utilization for hospitalized medical patients[30] and 1% of the variation in total resource utilization for vascular surgery.[31] Feinglass et al[15] reported what appears to be a large physician practice effect at an urban hospital. Indeed, they suggested that annual patient charges could be reduced by $250 000 if the 10 high resource-use physicians practiced at the average resource-use level for all physicians. However, the proportion of variance accounted for by physician appears from their tables to be at most 3.0%, and these potential savings very likely are overestimated when using the usual profiling approach. Adjusting for the low reliability of the physician effect, an estimate of the savings that could be achieved at that hospital is closer to $65 000 (and this is for patient charges, actual savings would be less). Miller et al[32] found similarly little difference in how physicians practice for several specific measures of outpatient resource utilization in the general medical clinics at a university teaching hospital.

Only for a few of the most specific clinical indicators measuring processes of care did we find good evidence in the literature of larger amounts of physician-level variation as a proportion of the total variation in a profile measure. Orav et al,[33] found that the practitioner accounted for a maximum of about 24% of the variance in a process of care score related to the management of digoxin and a minimum of 3% in process scores related to cancer screening.[33] Although there was more evidence for variation at the physician level when examining a few of these very specific processes of care, they also noted that there was essentially no correlation between provider performance on 1 guideline and their performance on any of the others. Thus, any more global quality score would average out differences between physicians and consequently would have low reliability. Unfortunately, very specific process measures will usually apply to only a small fraction of a physician's patients, which makes the process of profiling even more difficult.

In summary, most of the published evidence suggests that the individual physicians rarely account for more than 4% of the variation in common profile measures after case-mix adjustment. It might be useful to profile and control this relatively small amount of physician variation, but only if the costs of intervening are worth the expected gains. For a utilization measure, 4% of the variance may represent a great deal of money. However, given this relatively small physician-specific effect, there might be much more value in seeking factors that affect utilization, satisfaction, and clinical efficacy among all patients more substantially than the practices of their particular physician. Perhaps system and cohort effects minimize differences between providers within sites or groups of physicians, and larger differences may be found between sites or regions of the country.

Given the small amount of variation attributable to the physician and the

larger amount attributable to the patient's prior utilization or experience,[34] it will generally be much easier for physicians to change their profile results by manipulating their patient populations than by improving their efficiency or quality. Uncertainty about the consequences of managed care deselecting decisions (decisions to terminate a contract with a physician) and the role of profiles in these decisions has created an environment in which the "the soundest strategic advice for physicians has been simply to make every effort to avoid deselection."[6] Unfortunately, an easy way to ensure a good profile is to refuse to care for sick patients, those who have failed therapy, or those who do not adhere to treatment plans. Our results suggest that, at least for patients with diabetes, even the most sophisticated case-mix adjustment will not eliminate this strong perverse incentive, similar to the incentives that may encourage HMOs to attempt to select healthy populations when enrolling up capitated patients.[34-37]

What conclusions should we draw from the results of our study? Those who produce physician profiles should first make a realistic assessment of the reliability of those profiles. The reliability depends on both panel size and how much physicians vary in their practice. Some specialists may have very large panel sizes, which may allow for reliable profiling, but it is then important to ask if the differences between physicians are worth profiling. Profiling should be considered only if the physician-level variation is deemed important relative to other potential sources of variation. At that point, more complete case-mix measures will be necessary. Age- and sex-adjustment, while inexpensive, is little better than no case-mix adjustment at all. For profiling, more detailed case-mix adjustment, such as the Total Illness Burden Index and the Medical Outcomes Study Short-Form Health Survey, will be needed and adjustments for patients' past use or values may also be necessary.[38] Finally, profilers must consider that the application of profiles may foster an environment in which deselection of patients is the easiest way for physicians to avoid becoming deselected themselves. In our opinion, those who implement the profiling system would be as responsible for this result as those physicians who have succumbed to these pressures by denying care to the sickest and most vulnerable people needing medical care.

Funding/Support: This work was supported by grant HSO 6665-01 from the Agency for Health Care Policy and Research, Type II Diabetes Patient Outcomes Research Team, and the Veterans Affairs Health Services Research and Development Service, Washington, DC. Dr Hofer is supported by a Career Development Grant from the Health Services Research and Development Office of the Department of Veterans Affairs.

REFERENCES

1. Kerr EA, Mittman BS, Hays RD, Siu AL, Leake B, Brook RH. Managed care and capitation in California: how do physicians at financial risk control their own utilization? *Ann Intern Med*. 1995;123:500-504.
2. Wennberg DE. Variation in the delivery of health care: the stakes are high [editorial]. *Ann Intern Med*. 1998;128:866-868.
3. Spoeri RK, Ullman R. Measuring and reporting managed care performance: lessons learned and new initiatives. *Ann Intern Med*. 1997;127(8 pt 2):726-732.
4. Kassirer JP. The use and abuse of practice profiles. *N Engl J Med*. 1994;330:634-636.
5. Corrigan JM, Nielsen DM. Toward the development of uniform reporting standards for managed care organizations: the Health Plan Employer Data and Information Set (Version 2.0). *Jt Comm J Qual Improv*. 1993;19:566-575.
6. Liner RS. Physician deselection: the dynamics of a new threat to the physician-patient relationship. *Am J Law Med*. 1997;23:511-537.
7. Welch HG, Miller ME, Welch WP. Physician profiling: an analysis of inpatient practice patterns in Florida and Oregon. *N Engl J Med*. 1994;330:607-612.
8. Blum JD. The evolution of physician credentialing into managed care selective contracting. *Am J Law Med*. 1996;22:173-203.
9. Gauging quality regulation's impact on premium costs. *Med Health*. 1997;51:1.
10. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. *N Engl J Med*. 1995;332:1229-1232.
11. Greenfield S, Kaplan SH, Silliman RA, 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):32-39.
12. Hayward RA, Manning WG, Kaplan SH, Wagner EH, Greenfield S. Starting insulin therapy in patients with type 2 diabetes: effectiveness, complications, and resource utilization. *JAMA*. 1997;278:1663-1669.
13. McMahon LF Jr, Creighton FA, Bernard AM, Pittinger WB, Kelley WN. The integrated inpatient management model: a new approach to clinical practice . *Ann Intern Med*. 1989;111:318-326.
14. 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(suppl 4):AS47-AS55.
15. Feinglass J, Martin GJ, Sen A. The financial effect of physician practice style on hospital resource use. *Health Serv Res*. 1991;26:183-205.
16. Goldstein H. *Multilevel Statistical Models*. 2nd ed. New York, NY: Halstead Press; 1995.
17. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. New York, NY: Chapman & Hall; 1989.
18. Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the 3rd Berkeley Symposium Mathematics, Statistics, and Probability*. Vol 1. Berkeley: University of California Press; 1956:137.
19. James W, Stein C. Estimation with quadratic loss. *Proceedings of the 4th Berkeley Symposium of Mathematics, Statistics, and Probability*. Vol 1. Berkeley, Calif: University of California Press; 1961:361.
20. O'hagan A. Bayesian inference. In: Green PJ, Little RJ, Ord JK, Scott AJ, Weisberg S, eds. *Kendall's Advanced Theory of Statistics*. Vol 2B. New York, NY: Halstead Press; 1994.
21. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian *Data Analysis*. New York, NY: Chapman & Hall; 1995.
22. Gleason JR. sg65: computing intraclass correlations and large ANOVAs. In: Newton HJ, ed. *The Stata Technical Bulletin Reprints*. Vol 6. College Station, Tex: Stata Corp; 1996:167-176.
23. Carmines EG, Zeller RA. Reliability and validity assessment. In: Lewis-Beck MS, ed. *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Newbury Park, Calif: Sage Publications; 1979. Series no. 07-001.
24. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. *J Clin Epidemiol*. 1991;44:381-390.
25. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. 2nd ed. New York, NY: Oxford University Press; 1996.
26. *Stata Statistical Software* [computer program]. Release 5.0. College Station, Tex: Stata Corp; 1997.
27. *MLwiN* [computer program]. Release 1.02.002. London, England: Multilevel Models Project, Institute of Education; 1998.
28. Vijan S, Stevens DL, Herman WH, Funnell MM, Standiford CJ. Screening, prevention, counseling, and treatment for the complications of type II diabetes mellitus: putting evidence into practice. *J Gen Intern Med*. 1997;12:567-580.
29. Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ*. 1995;311:793-796.
30. Hayward RA, Manning WG, Mcmahon LF, Bernard AM. Do attending or resident physician practice styles account for variations inhospital resource use. *Med Care*. 1994;32:788-794.
31. Kuczynski YT, Stanley JC, Rosevear JS, McMahon LF Jr. Vascular surgeons' resource use at a university hospital related to diagnostic-related group and source of admission. *J Vasc Surg*. 1997;26:193-198.
32. Miller ME, Hui SL, Tierney WM, McDonald CJ. Estimating physician costliness: an empirical Bayes approach. *Med Care*. 1993;31(suppl 5):YS16-YS28.
33. Orav EJ, Wright EA, Palmer RH, Hargraves JL. Issues of variability and bias affecting multisite measurement of quality of care. *Med Care*. 1996;34 (suppl 9):SS87-SS101.
34. Newhouse JP, Manning WG, Keeler EB, Sloss EM. Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Rev*. 1989;10:41-54.
35. Porell FW, Turner WM. Biased selection under an experimental enrollment and marketing Medicare HMO broker. *Med Care*. 1990;28:604-615.
36. Lichtenstein R, Thomas JW, Watkins B, et al. HMO marketing and selection bias: are TEFRA HMOs skimming? *Med Care*. 1992;30:329-346.
37. Hellinger FJ. Selection bias in HMOs and PPOs: a review of the evidence. *Inquiry*. 1995;32:135-142.
38. Newhouse JP. Patients at risk: health reform and risk adjustment. *Health Aff (Millwood)*. 1994;13: 132-146.