Methods

Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures

J. William Thomas, Kyle L. Grazier, and Kathleen Ward

Objective. To investigate whether different risk-adjustment methodologies and economic profiling or "practice efficiency" metrics produce differences in practice efficiency rankings for a set of primary care physicians (PCPs).

Data Source. Twelve months of claims records (inpatient, outpatient, professional, and pharmacy) for an independent practice association HMO.

Study Design. Patient risk scores obtained with six profiling risk-adjustment methodologies were used in conjunction with claims cost tabulations to measure practice efficiency of all primary care physicians who managed 25 or more members of an HMO.

Data Collection. For each of the risk-adjustment methodologies, two measures of "efficiency" were constructed: the standardized cost difference between total observed (standardized actual) and total expected costs for patients managed by each PCP, and the ratio of the PCP's total observed to total expected costs (O/E ratio). Primary care physicians were ranked from most to least efficient according to each risk-adjusted measure, and level of agreement among measures was tested using weighted kappa. Separate rankings were constructed for pediatricians and for other primary care physicians.

Findings. Moderate to high levels of agreement were observed among the six risk-adjusted measures of practice efficiency. Agreement was greater among pediatrician rankings than among adult primary care physician rankings, and, with the standardized difference measure, greater for identifying the least efficient than the most efficient physicians. The O/E ratio was shown to be a biased measure of physician practice efficiency, disproportionately targeting smaller sized panels as outliers.

Conclusions. Although we observed moderate consistency among different risk-adjusted PCP rankings, consistency of measures does not prove that practice efficiency rankings are valid, and health plans should be careful in how they use practice efficiency information. Indicators of practice efficiency should be based on the standardized cost difference, which controls for number of patients in a panel, instead of O/E ratio, which does not.

Key Words. Physician profiling, economic profiling, practice efficiency, risk-adjustment, observed-to-expected ratio, standardized cost difference

Managed care plans in the United States are increasingly using paid claims records to monitor and judge the physicians who provide services to plan members. The motivation for economic, or "practice efficiency," profiling, as this is termed, is primarily financial—physicians identified as being inefficient are considered to be wasteful of health plan resources, and these physicians can be encouraged to change their practice styles or they can be dropped from the plan's provider network (Sandy 1999; Nickerson and Rutledge 1999; Litton, Sisk, and Akins 2000). Vendors and clients (primarily health plans) of profiling systems consider a physician to be efficient if total claims costs of services provided to managed patients are no greater than costs expected for those patients, given the patients' demographic characteristics and health conditions. "Inefficient" physicians are those for whom actual claims costs exceed the expected amounts.[1]

When several physicians are involved in providing care to a patient, it is often quite difficult to determine from claims records which of the physicians ordered a particular service, prescribed a particular drug, or even admitted the patient to a hospital. Because of this attribution problem, economic profiling appears to be used most frequently in managed care plans that require primary care physicians (PCPs) to serve as gatekeepers. In these plans, the PCPs are assumed responsible for *all* costs incurred on behalf of the patients they manage, regardless of whether they themselves performed or even ordered the services. And with this assumption, attribution of responsibility is not an analytic problem.

Even though dozens of firms offer profiling software and services to health plans, there are relatively few methodologies that can be used for risk-adjusting physician profiles, that is, for estimating patients' expected costs. And nearly all profiling vendors use one or more of these established methodologies. In the project reported here, our purposes were to determine whether some risk-adjustment methodologies used for physician profiling are more accurate than others, and whether risk-adjustment differences lead to

---

The study was supported by grant no. 36874 from the Robert Wood Johnson Foundation Health Care Financing and Organization (HCFO) Program, and by Grant 243-II/99 from the Blue Cross/Blue Shield of Michigan Foundation.

Address correspondence to J. William Thomas, Ph.D., 356 Shore Road, Bremen, ME 04551. Dr. Thomas is Professor Emeritus of Health Management and Policy, University of Michigan, and Professor of Health Policy and Management, University of Southern Maine, Portland. Kyle L. Grazier, Dr. P.H., is Professor of Health Management and Policy, University of Michigan, Ann Arbor, MI. Kathleen Ward, M.P.A., is Health Programs Analyst, M-CARE Quality Improvement, Ann Arbor, Michigan.

different judgments about PCP practice efficiency. Our findings on accuracy of risk-adjustment methodologies are presented in Thomas, Grazier, and Ward (2002). Here, we describe findings related to levels of agreement among PCP practice efficiency rankings when, for a common set of PCPs, patients' expected costs are estimated using different risk-adjustment methodologies and two different measures of practice efficiency.

For our analyses, we used the membership and claims databases of an independent practice association HMO that serves the five counties of Southeast Michigan. Six profiling system vendors/developers agreed to work with us on the project, to either make their risk-adjustment software available, or to process our data through their software and return the risk-adjusted results to us. A detailed description of each of the participating methodologies is provided in our project report (Thomas, Grazier, and Ward 2002). The methodologies are:

Adjusted Clinical Groups (ACGs Version 4.5, 2000) from Johns Hopkins University. Adjusted clinical groups cluster health plan members having similar comorbidities into groups that have similar resource requirements and clinical characteristics. The ACG Case-Mix System then uses a branching algorithm to place each patient into one of 82 discrete, mutually exclusive categories based on the mix of clinical groups experienced during the time period under study (Johns Hopkins University 2000).

Burden of Illness Score (BOI Version PRS 4.6, 2001) from MEDecision, Inc. This system is based on MEDecision's Practice Review System (PRS), which partitions care into episodes of illness and assigns services, severity levels, and medications to these episodes. The BOI Score is a linear-scaled measure that indicates relative health care cost risks associated with the particular mix of episodes experienced by a patient during a defined time period (Anderson and Gilbert 2002).

Clinical Complexity Index (CCI Version 3.6, 1997) from Solucient, Inc. The CCI methodology considers age, severity, comorbidity, hospital admissions, and categories of diagnoses (acute, chronic, mental health, and pregnancy) to assign patients into mutually exclusive CCI risk categories. Although the system provides for 1,418 different categories, 95 percent of patients fall into just 45 of these (Solucient 1999).

Diagnostic Cost Groups (DCGs Version 5.1, 2000) from DxCG, Inc. The DCG system includes a whole family of multiple linear regression models. For this study, we utilized the all encounter, hierarchical model (DCG/HCC) for a commercial population, which uses data on age, sex, and all diagnoses—inpatient and outpatient—to explain patients' health care expenditures for the

period under study. In our analyses, we used the DCG retrospective risk measure, together with patients' age/sex categories (DxCG Inc. 2002).

Episode Risk Groups (ERGs Version 4.2, 2001) from Symmetry Health Systems, Inc. Like BOI Score, ERGs are episode-based. The episodes underlying ERGs are created using Symmetry's Episode Treatment Groups (ETG™) methodology, a basic illness classification system that uses a series of clinical and statistical algorithms to combine related services into more than 600 mutually exclusive and exhaustive categories. For a given patient, episodes experienced during a time period are mapped into 119 Episode Risk Groups, and then a risk score is determined based on age, gender, and mix of ERGs. For our analyses, we used the ERG retrospective risk score (Symmetry Health Data Systems 2001).

General Diagnostic Groups (GDGs Version 1.0, 2000) from Allegiance LLC. General Diagnostic Groups were developed using the Agency for Health Care Policy and Research's Clinical Classification Software (CCS). CCS aggregates individual ICD-9-CM codes identified on health care claims into 260 broad diagnosis categories for statistical analysis and reporting. The GDG system then lumps together CCS categories considered to be clinically similar and to have similar associated per-patient charges into 57 diagnostic categories. These 57 diagnostic categories are used as dummy variables in a multiple regression model for predicting health care costs (Cowen et al. 1998).

Although CCI and the two episode-based methodologies include utilization data as independent predictors of risk, our analyses (not reported here) found that these systems were not different from the other three systems in terms of accuracy of member cost predictions (Thomas, Grazier, and Ward forthcoming).

METHODS

Data

The data used in this study included Blue Care Network of Michigan member and claims files (inpatient, outpatient/professional, and pharmacy) for July 1, 1997, to June 30, 1998. For this study period, the HMO had 156,280 continuously enrolled members, all having the same benefit package. Of this total, 127,004 (81.3 percent) had at least one claim; and 115,856 members (74.1 percent) had at least one provider encounter that resulted in an identified diagnosis. Because the participating risk-adjustment systems treat nonusers and pharmacy-service-only users differently,[2] we chose to limit our analyses to

the members who had at least one identified diagnosis during the study period and could therefore be classified by all of the risk-adjustment methodologies.

Member Standardized Actual Costs

For each member who had at least one identified diagnosis during the study period—the fundamental unit of analysis for the study—we calculated total health care costs as the sum of submitted amounts on the member's pharmacy, professional and outpatient services claims, and the sum of paid amounts on the member's inpatient services claims. Because the health plan uses a diagnosis related group (DRG)-based payment system for hospital services, hospital payments were considered to be a more appropriate measure for these calculations than hospital charges.

Claims costs reflect not only quantity and mix of services delivered, but also the prices paid for those services. Hence, some of the cost variation in our data base was associated with price differences among providers. Since the purpose of our study was to investigate effects of differences in risk-adjustment, price related variation in the data was considered to represent "noise." To enhance our ability to detect true differences among risk-adjustment methodologies, we developed standard prices, such that a given service would be priced at the same level across all providers. The price standardization methodology is described in Thomas, Grazier, and Ward (2002). After summing standardized claims costs for each member, member costs were truncated, or "top-coded," at $25,000 to minimize potentially distorting effects of high cost outlier patients on physician profiles.[3] In the discussion below, we refer to these top-coded totals as members' "standard cost."

Risk-Adjustment

HMO membership and claims data were processed through the risk-adjustment software of each of the participating systems. System output files yielded the following member specific risk categorization data for members included in our analyses:

- For ACGs, members were assigned to 78 discrete ACG categories;
- For BOI, members were assigned a linear scaled risk score that ranged from 5 to 40,585 with a mean value of 792;
- For CCI, members were assigned to 327 discrete CCI categories, 36 of which accounted for 90 percent of HMO members;

- For DCGs, each member was assigned a retrospective risk score mapped into 25 ordinal scaled categories; members were also assigned to 31 age/sex categories;
- For ERGs, members were assigned a retrospective risk score mapped into 26 ordinal scaled retrospective risk categories;
- For GDGs, members were assigned to one or more of 57 diagnostic category variables.

PCP-Level Analyses

Primary care physician panels were the primary unit of inference in the study. We limited our analyses to those that contained at least 25 patients, and these were divided into two groups:

- Mostly adult panels (661) managed by family practitioners, general practitioners, and internal medicine specialists who served a combined total of 79,959 patients (we refer to these as adult care PCPs), and
- Pediatricians (143) serving a total of 20,796 patients.

For each of the 100,755 members served by these two groups of physicians, risk categories/scores were used to develop six expected costs estimates according to procedures described elsewhere (Thomas, Grazier, and Ward 2002). With these expected costs, we then calculated, for each risk-adjustment methodology, two alternative measures of practice efficiency for each PCP. The first measure was the ratio of observed to expected costs, and it is calculated as:

$$O/E_{ki} = \frac{y_k}{\hat{y}_{ki}}$$

where y_k is average standardized actual costs for physician k, $\hat{y}_{ki}$ is average expected cost for physician k according to the i^{th} risk adjustment methodology, and O/E_{ki} is the observed-to-expected ratio for the k^{th} PCP according to the i^{th} risk-adjustment system. Although, to our knowledge there are no published data to document this, our discussions with profiling vendors and clients suggest that O/E ratio is the measure typically used to characterize practice efficiency in physician profiling systems.

The second measure of PCP practice efficiency was the standardized difference between average standardized actual cost and average expected cost for patients managed by the physician. Using Z_{ki} to represent the

standardized cost difference (SCD) for the k^{th} physician according to the i^{th} risk-adjustment measure,

$$Z_{ki} = \frac{y_k - \hat{y}_{ki}}{\sigma_i / \sqrt{N_k}}$$

where y_k and $\hat{y}_{ki}$ are as defined above, σ_i is the standard deviation of average expected costs associated with the i^{th} risk-adjustment system across all panels, and N_k is the number of patients in the panel of the k^{th} PCP. The concept underlying this measure is that panels may be considered to consist of random samples of patients drawn from a population having health care costs with mean equal to $\hat{y}_i$ and standard deviation equal to σ_{i}. According to the Central Limit Theorem, the distribution of sample means can be considered approximately normal with a mean of $\hat{y}_i$ and standard deviation of $\sigma_{i./}\sqrt{N_k}$, with N_k representing sample size. Thus a panel's SCD score measures its mean cost in terms of number of standard deviations above or below the population mean. For both adult care physicians and pediatricians in our sample, SCD values were found to average 0.0 ± 0.1 with standard deviations ≤ 2.0 for all risk-adjusted measures. Had the sample mean distributions been perfectly normal, we would have expected mean values of 0.0 and standard deviations of 1.0.

With each of these two measures, adult care PCPs and pediatric PCPs were ranked from most efficient (lowest Z_{ki} and O/E_{ki}) to least efficient (highest Z_{ki} and O/E_{ki}) according to each of the risk-adjustment systems. High (low) outliers were defined arbitrarily as the 10 percent having the highest (lowest) scores. We also investigated outlier thresholds of 5 percent and 20 percent. Separately for adult care PCPs and pediatric PCPs and each pair of risk-adjustment systems, we calculated:

- Percentage agreement on low-outlier PCPs, defined as the fraction of low-outliers identified by one system that are also identified by the other system,
- Percentage agreement on high-outlier PCPs, defined as the fraction of high-outliers identified by one system that are also identified by the other system, and
- Weighted kappa measure of agreement among PCP decile rankings of the two systems. To calculate weighted kappa, PCP rankings (1 to 661 for adult care PCPs and 1 to 143 for pediatricians) were recoded into deciles, and for each pair of recoded rankings analysis of variance was used to determine the proportion of variance (weighted

kappa) in the one ranking explained by the second ranking (Landis and Koch 1977).

RESULTS

Results of pair-wise level-of-agreement analyses are summarized in Table 1.[4] For SCD- and O/E-based measures, the table shows average pair-wise levels of agreement between each risk-adjusted ranking and other risk-adjusted rankings. Average percentage agreement is shown for high outlier and low outlier identification, and average weighted kappa values indicate level of agreement across the 10-decile range.[5] For internal medicine, general medicine, and family medicine physicians, levels of agreement on identification of high outliers are higher when based on SCD rankings, while agreement on low outlier identification is higher with O/E-based rankings. For these adult care physicians, average weighted kappa values indicate that overall agreement between pairs of rankings is moderate, according to criteria of Landis and Koch (1977), and that overall levels of agreement are similar for the two types of rankings. These same patterns hold for pediatric practice efficiency rankings. However, average levels of agreement on identification of high outliers are higher—with both SCD- and O/E-based measures—than those for adult care physicians. Overall agreement between pairs of pediatric rankings is also greater; average weighted kappa values are all in what Landis and Koch (1977) consider the "substantial agreement" range.

In the analyses presented in Table 1, high outliers were defined as those above the 90th percentile of the practice efficiency distribution, and low outliers as those below the 10th percentile. In Table 2, we show how outlier threshold definition influences levels of agreement between pairs of practice efficiency measures. In general, for both adult care and pediatric PCPs, with both SCD and O/E measures of practice efficiency, average levels of agreement on identification of outlier physicians are lower when outlier thresholds are defined at 5 percent than when thresholds are defined at 10 percent, and they are lower when thresholds are defined at 10 percent than when defined at 20 percent. This is true both for high outlier agreement and low outlier agreement. For pediatricians, there are two exceptions to this general pattern: levels of SCD measure agreement for high outlier pediatricians are not consistently different when outlier thresholds are defined at 10 percent and at 20 percent; and levels of O/E measure agreement for low outlier pediatricians are actually somewhat higher when outlier thresholds are

Table 1: Average Levels of Agreement (%) of PCP Practice Efficiency Rankings: by Risk-Adjustment System, Physician Type, and Practice Efficiency Measure*

		Practice Efficiency Measure					
		SCD Score			*O/E Ratio*		
Physician Specialty	*Risk-Adjustment System*	*Agreement High Outliers*	*Agreement Low Outliers*	*Weighted Kappa*	*Agreement High Outliers*	*Agreement Low Outliers*	*Weighted Kappa*
Family Medicine,	ACGs	54	49	0.56	44	53	0.56
General Medicine,	BOI Score	55	47	0.56	46	58	0.57
and Internal Medicine	CCI	54	44	0.52	44	53	0.52
	DCGs	58	53	0.59	48	57	0.59
	ERGs	57	47	0.59	45	56	0.60
	GDGs	56	42	0.51	48	52	0.52
Pediatrics	ACGs	77	49	0.70	65	57	0.73
	BOI Score	67	44	0.66	53	58	0.66
	CCI	69	47	0.67	67	51	0.68
	DCGs	71	37	0.67	61	59	0.69
	ERGs	71	46	0.63	60	59	0.67
	GDGs	69	34	0.61	51	46	0.62

*Average pair-wise levels of agreement between each risk-adjusted ranking and other risk-adjusted rankings. High- and low-outlier thresholds defined at 10%.

Table 2: Average Levels of Agreement (%) of PCP Practice Efficiency Rankings: by Physician Type, Practice Efficiency Measure, Risk-Adjustment System, and Cost Outlier Threshold*

			High Outlier Threshold			*Low Outlier Threshold*		
Physician Specialty	*Practice Efficiency Measure*	*Risk-Adjustment System*	*5%*	*10%*	*20%*	*5%*	*10%*	*20%*
Family Practice, General Practice, and Internal Medicine	Standardized Cost Difference	ACGs	46	54	64	44	49	62
		BOI Score	49	55	63	42	47	59
		CCI	45	54	62	38	44	61
		DCGs	52	58	63	45	53	62
		ERGs	52	57	63	42	47	61
		GDGs	45	56	63	35	42	55
	O/E Ratio	ACGs	35	44	60	44	53	67
		BOI Score	32	46	59	53	58	66
		CCI	35	44	55	48	53	66
		DCGs	37	48	60	48	57	69
		ERGs	36	45	62	50	56	68
		GDGs	42	48	59	48	52	62
Pediatrics	Standardized Cost Difference	ACGs	53	77	73	40	49	61
		BOI Score	55	67	69	46	44	63
		CCI	50	69	75	31	47	63
		DCGs	55	71	72	43	37	62
		ERGs	60	71	66	34	46	64
		GDGs	58	69	72	23	34	49
	O/E Ratio	ACGs	50	65	74	60	57	68
		BOI Score	55	53	73	63	58	66
		CCI	55	67	74	57	51	66
		DCGs	61	61	71	66	59	67
		ERGs	60	60	72	63	59	69
		GDGs	46	51	70	46	46	59

*Average pair-wise levels of agreement between each risk-adjusted ranking and other risk-adjusted rankings.

defined at 5 percent than when defined at 10 percent. Also, the pattern observed in Table 1, that SCD measured practice efficiency is associated with higher levels of agreement on high outlier identification and lower levels of agreement on low outlier identification, continues to hold for the alternative outlier threshold definitions considered here.

To investigate influence of practice efficiency metric on high- and low-outlier designations, for each risk-adjustment methodology we partitioned adult PCP outliers into (a) those identified by both types of practice efficiency measure, (b) those identified only on the basis of SCD score, and (c) those

identified only on the basis of O/E ratio.[6] In Table 3, we show descriptive statistics for adult PCP panels in each of these categories. The same data for adult care PCP panels identified as low outliers are shown in Table 4. Table 3 shows that with the ACG risk-adjusted measures, 51 of the 67 adult PCPs (76 percent) identified as high outliers on the basis of SCD score were also identified as high outliers with O/E ratios. Although agreement between the two types of measures on high-outlier identity was lower with the other risk-adjustment systems studied, even the lowest rate of agreement (with ERGs) exceeded 64 percent. However, the most interesting aspect of the data in Table 3 is not the high rate of agreement between SCD score and O/E ratio. Rather, it is the characteristics of the PCP panels on which the two measures differ. Compared to panels identified as high outliers only with O/E ratio, those identified with SCD score include larger numbers of patients, and have larger average standardized costs and larger expected costs. Across all 661 adult PCP panels included in this analysis, the mean number of patients per panel is 121.0, with standard deviation equal to 87.7. Sizes of panels identified as high outliers by both measures are not very different from this overall average—mean panel sizes range from 119.6 for ACG high outliers to 94.8 for ERG high outliers. But the mean size of high outlier panels identified only with SCD score range from a low of 159.6 for BOI Score risk adjustment up to 201.6 with ERGs. For O/E ratio-identified high outliers, on the other hand, mean panel sizes are in the 44.2 to 57.5 range. Why are these panel sizes so different? The SCD measure describes a panel's mean standardized actual cost in terms of number of standard deviations above or below the mean of the sampling distribution. While cost ratios (right-most column of Table 3) of panels identified as high outliers on the basis of O/E ratio are higher than those of panels identified using SCD score, the standardized actual costs of these panels are not as deviant—in terms of number of standard deviations from the sampling distribution mean—as those of panels identified as high outliers using SCD score. For example, for ACG-identified high outliers, those identified only on the basis of SCD score have standardized actual costs that are, on average, 3.11 standard deviations above the mean, while standardized actual costs of those identified only on the basis of O/E ratio are 2.12 standard deviations above the mean. With the other risk-adjustment systems (except for BOI Score) mean standardized actual costs of panels identified as high outliers on the basis of O/E ratio only are less than 2.0 standard deviations above the mean.

Table 4 presents characteristics of panels identified as low outliers. This table shows less agreement between the two types of measures on low-outlier

Table 3: Characteristrics of Family Medicine, General Medicine, and Internal Medicine Panels Identified as High Outliers, by Risk-Adjustment System and Practice Efficiency Measure*

Risk-Adjustment System	*Practice Efficiency Measure*	*Number PCP Panels Identified*	*Number Patients per Panel*	*Standardized Cost per Patient*	*Expected Cost per Patient*	*SCD Score*	*O/E Ratio*
ACGs	Both	51	119.6	$3,835	$2,699	4.01	1.43
			(65.7)	($844)	($619)	(1.23)	(0.16)
	SCD Score Only	16	191.0	$4,306	$3,564	3.11	1.21
			(136.7)	($1,080)	($857)	(0.45)	(0.04)
	O/E Ratio Only	16	57.5	$3,151	$2,346	2.12	1.35
			(18.3)	($442)	($380)	(0.34)	(0.07)
BOI Score	Both	45	112.0	$3,896	$2,748	3.87	1.42
			(66.6)	($882)	($612)	(1.12)	(0.11)
	SCD Score Only	22	159.6	$4,151	$3,363	3.20	1.23
			(85.9)	($910)	($723)	(0.62)	(0.04)
	O/E Ratio Only	22	45.0	$3,451	$3,068	2.08	1.37
			(20.7)	($597)	($801)	(0.39)	(0.06)
CCI	Both	40	111.2	$4,253	$3,205	3.80	1.33
			(65.3)	($839)	($621)	(0.97)	(0.07)
	SCD Score Only	27	184.3	$3,868	$3,257	2.89	1.19
			(89.6)	($821)	($727)	(0.44)	(0.04)
	O/E Ratio Only	27	48.5	$3,438	$2,682	1.91	1.28
			(20.3)	($631)	($503)	(0.39)	(0.03)
DCGs	Both	50	114.5	$4,163	$3,090	3.70	1.35
			(68.5)	($890)	($654)	(0.89)	(0.12)
	SCD Score Only	17	186.5	$4,070	$3,432	2.93	1.19
			(92.0)	($668)	($559)	(0.46)	(0.03)
	O/E Ratio Only	17	46.6	$3,236	$2,445	1.93	1.33
			(12.1)	($440)	($405)	(0.33)	(0.07)
ERGs	Both	43	94.8	$4,411	$3,344	3.35	1.32
			(46.5)	($830)	($659)	(0.67)	(0.09)
	SCD Score Only	24	201.6	$3,885	$3,267	2.92	1.19
			(85.9)	($583)	($498)	(0.56)	(0.02)
	O/E Ratio Only	24	44.2	$3,529	$2,766	1.72	1.28
			(13.9)	($849)	($706)	(0.34)	(0.05)
GDGs	Both	45	115.4	$3,875	$2,882	3.25	1.35
			(74.1)	($845)	($619)	(0.78)	(0.12)
	SCD Score Only	22	198.2	$3,965	$3,337	2.86	1.19
			(113.3)	($698)	($600)	(0.52)	(0.03)
	O/E Ratio Only	22	49.0	$3,143	$2,411	1.69	1.31
			(23.8)	($743)	($605)	(0.39)	(0.09)

*High-outlier thresholds defined at 10%. Sample consisted of 661 PCP panels, with average cost per panel of $3,038 (standard deviation = $801) and average number of patients per panel of 121 (standard deviation = 88).

Table 4: Characteristics of Family Medicine, General Medicine, and Internal Medicine Panels Identified as Low Outliers, by Risk-Adjustment System and Practice Efficiency Measure*

Risk-Adjustment System	*Practice Efficiency Measure*	*Number PCP Panels Identified*	*Mean (Standard Deviation)*				
			Number Patients per Panel	*Standardized Cost per Patient*	*Expected Cost per Patient*	*SCD Score*	*O/E Ratio*
ACGs	Both	31	136.9	$2,154	$3,044	−3.42	0.71
			(74.9)	($396)	($599)	(1.18)	(0.07)
	SCD Score Only	35	238.1	$2,571	$3,090	−2.71	0.83
			(102.8)	($489)	($546)	(0.50)	(0.02)
	O/E Ratio Only	35	52.3	$2,057	$2,746	−1.70	0.75
			(20.6)	($461)	($539)	(0.29)	(0.05)
BOI Score	Both	42	142.9	$2,149	$2,989	−3.30	0.72
			(81.7)	($507)	$595	(0.84)	(0.06)
	SCD Score Only	24	252.0	$2,805	$3,374	−3.00	0.83
			(114.3)	($602)	$681	(0.44)	(0.02)
	O/E Ratio Only	24	52.2	$1,893	$2,581	−1.69	0.73
			(18.1)	($445)	$556	(0.40)	(0.05)
CCI	Both	45	136.5	$2,305	$3,320	−4.03	0.70
			(84.0)	($496)	($669)	(1.26)	(0.08)
	SCD Score Only	21	228.9	$2,647	$3,205	−2.87	0.83
			(130.9)	($428)	($507)	(0.48)	(0.03)
	O/E Ratio Only	21	54.0	$1,973	$2,681	−1.84	0.73
			(28.3)	($514)	($632)	(0.44)	(0.05)
DCGs	Both	46	139.1	$2,125	$2,948	−3.23	0.72
			(94.8)	($403)	($531)	(1.02)	(0.06)
	SCD Score Only	20	259.5	$2,602	$3,132	−2.96	0.83
			(104.9)	($397)	($445)	(0.63)	(0.03)
	O/E Ratio Only	20	46.9	$1,998	$2,690	−1.64	0.74
			(19.0)	($482)	($570)	(0.31)	(0.05)
ERGs	Both	46	137.3	$2,124	$2,928	−2.96	0.73
			(93.7)	($434)	($642)	(1.05)	(0.06)
	SCD Score Only	20	266.9	$2,495	$2,965	−2.53	0.84
			(119.2)	($440)	($500)	(0.44)	(0.02)
	O/E Ratio Only	20	55.0	$1,896	$2,494	−1.47	0.75
			(27.7)	($481)	($531)	(0.23)	(0.05)
GDGs	Both	43	127.3	$2,437	$3,365	−3.28	0.73
			(75.4)	($614)	($805)	(1.02)	(0.06)
	SCD Score Only	23	209.0	$3,137	$3,698	−2.65	0.85
			(99.1)	($792)	($853)	(0.44)	(0.03)
	O/E Ratio Only	23	53.3	$1,861	$2,542	−1.68	0.73
			(22.0)	($391)	($471)	(0.42)	(0.05)

*Low-outlier thresholds defined at 10%. Sample consisted of 661 PCP panels, with average cost per panel of $3,038 (standard deviation = $801) and average number of patients per panel of 121 (standard deviation = 88).

identification (from 47 percent [ACGs] to 70 percent [DCGs and ERGs]) than seen in Table 3 for high-outlier identification. However, the very strong panel-size pattern seen in Table 3 is not only present in Table 4, it is even stronger. For low outlier panels, mean numbers of patients for panels identified only on the basis of SCD score range from 209. 0 (GDGs) to 266.9 (ERGs), while for those identified using O/E ratio only mean sizes vary between 46.9 (DCGs) and 55.0 (ERGs). Again, the O/E ratio measure is more likely to identify smaller panels as cost outliers. And patterns of mean SCD scores and mean O/E ratios are similar to those in Table 3 as well. Although mean O/E ratios of panels identified as low outliers on the basis of O/E ratio are smaller than those of panels targeted by SCD score, when sample size is taken into account the degree to which standardized actual costs for the O/E-identified panels deviate from their expected values is less than that of SCD-identified low outlier panels.

The pattern shown in Tables 3 and 4 is strong, and it is consistent for all of the risk-adjusted rankings. This pattern is not an artifact of our data or of our methodology. Rather, it is a straightforward consequence of the statistical property that the standard deviation of a sampling distribution is equal to the population standard deviation divided by the sample size. The SCD calculations reflect this property; the O/E ratio calculations do not.

Finally, in Table 5 we look at the combined effects of choice of risk-adjustment methodology and choice of practice efficiency measure on consistency of (a) high-outlier and (b) low-outlier identification with 10 percent outlier thresholds. In both parts of the table, percentages shown on the diagonals represent proportions of outliers identified by both types of practice efficiency measures with the designated risk-adjustment systems. Off the diagonal, percentages indicate proportions of outliers identified by row-designated risk-adjusted O/E ratio measures that are also identified by the column-designated risk-adjusted SCD measures. For high outliers, off-diagonal percentages range from 32.8 percent to 50.7 percent, and for low outliers the range is 25.8 percent to 47.0 percent. Thus, a health plan using its claims data to profile primary care physicians with GDG-adjusted O/E ratios might identify as high outliers fewer than 33 percent of the PCPs than it would identify if it were to use ERG-adjusted SCD scores. Further, 67 percent of the high outlier PCPs identified by the GDG-O/E ratio measure are not high outliers according to the ERG-SCD measure. And only about a quarter of low outliers identified on the basis of ACG risk-adjustment and SCD practice efficiency measurement would be identified if CCI-adjusted O/E ratios were to be used with the same claims data.

Table 5: Degree of Agreement (%) on Outlier Designation for Family Medicine, General Medicine, and Internal Medicine Physicians by Risk-Adjustment System and Type of Practice Efficiency Measure*

(a) Designated High Outliers

Risk-Adjustment System and Practice Efficiency Measure		*SCD Score*					
		ACG	*BOI*	*CCI*	*DCG*	*ERG*	*GDG*
O/E Ratio	ACG	76					
	BOI	39	67				
	CCI	36	46	60			
	DCG	51	51	46	75		
	ERG	40	40	45	49	64	
	GDG	43	45	37	37	33	67

(b) Designated Low Outliers

Risk-Adjustment System and Practice Efficiency Measure		*SCD Score*					
		ACG	*BOI*	*CCI*	*DCG*	*ERG*	*GDG*
O/E Ratio	ACG	47					
	BOI	38	64				
	CCI	26	35	68			
	DCG	36	41	47	70		
	ERG	39	45	33	41	70	
	GDG	29	30	44	39	39	65

*High- and low-outlier thresholds defined at 10%.

Discussion And Conclusion

With SCD score and outlier thresholds set at 10 percent, different risk-adjusted rankings agreed on identities of 50 percent to 60 percent of the least efficient family medicine, general medicine, and internal medicine physicians. With outlier thresholds of 20 percent, agreement between pairs of risk-adjusted rankings was consistently above 60 percent. For pediatricians, high-outlier levels of agreement were in the 65 percent to 75 percent range with both 10 percent and 20 percent outlier thresholds. Are these levels of agreement good? Are they satisfactory? From a purely statistical standpoint, 65 percent to 75 percent agreement is quite good. Over the whole range of pediatric PCP practice efficiency rankings, agreement among the six different

risk-adjustment systems represented what Landis and Koch (1977) consider to be "substantial agreement." But whether "substantial agreement" is good, or good enough, or even satisfactory, depends entirely on how the ranking information is used. The problem with lack of agreement, *any* lack of agreement, is that—as noted above—we do not have information on the validity of any of the rankings. If different risk-adjustment methodologies identify the same set of PCPs as high outliers, the set possesses "concurrent validity," and this agreement gives us confidence that the identified physicians indeed are the least efficient. If, however, different systems identify different physicians, we cannot have similar confidence in any of the identified sets, because there is no way for us independently to know which of the systems' rankings, if any, is correct. Moreover, physician profiling vendor clients are unlikely to use multiple systems and produce multiple rankings in order to establish concurrent validity. Instead, clients will normally license and use a single profiling methodology. And even though our analyses suggest that 50 percent to 60 percent of adult PCPs identified by their system as being high outliers are likely to be identified by other profiling systems as well, the client has no way to know which of the identified outliers are the ones that multiple systems would agree on. Thus the profiling client must deal with practice efficiency rankings knowing that, in all likelihood, 40 percent to 50 percent of PCPs identified as high outliers are actually not among the least efficient 10 percent of primary care physicians. Is this level of information of adequate quality for providing confidential feedback to physicians on their own practice efficiency performance? In our opinion, the information is sufficient for this purpose. Is the information adequate for taking punitive action against the low efficiency physicians, perhaps dropping them from the health plan's provider panel? We think that it is not.

While our results do not allow us to identify any particular risk-adjusted ranking as being more or less valid than other rankings, we are able to make a definite statement about the type of measure that should be used for profiling primary care physicians. Standardized cost difference provides a more accurate measure of PCP practice efficiency than the more commonly used O/E ratio. When physician rankings are based on the ratio of observed to expected costs, smaller panels are more likely and larger panels less likely to be identified as outliers than they would be if panel size were taken into account.[7] Depending upon risk-adjustment methodology, 24 percent to 40 percent of high outliers and 30 percent to 53 percent of low outliers identified using an O/E ratio of PCP practice efficiency must be considered to be false positives. This specificity problem occurs because the degree of random

variation present in panel average costs varies inversely with the square root of panel size, and the O/E ratio measure fails to consider this relationship. Further, it compounds measurement problems associated with choice of risk-adjustment methodology, leading to potential false positive rates in excess of 60 percent for high outlier PCPs and 70 percent for low outliers. Because all needed data—average standardized actual cost, average expected costs, and number of patients—are readily available, panel SCD score can be calculated as easily and quickly as O/E ratio, and this source of potential error in PCP rankings can be eliminated. Although the SCD is more valid and reliable than O/E ratio, it is likely to be more difficult to explain to policymakers and those who use the efficiency results. Nevertheless, the extra work and creativity required in devising explanations are justified by improvements in decisions.

Potential generalizability of our findings may be influenced by several factors. Our analyses were performed with data from a single health plan, and there is no assurance that these data are typical of other plans. However, neither is there reason to believe that these data are in any way unique, or that characteristics of the plan, provider network, or patients would have influenced the findings. On the other hand, it is likely that several aspects of our methodology may have affected levels of agreement observed between pairs of risk-adjusted rankings:

- Prior to our analyses, we standardized claims costs in order to remove variability associated with service pricing differences. Had these costs not been standardized, levels of agreement between rankings might have been different than those presented in our tables.
- We limited our analyses to HMO members who were enrolled for the entire 12-month study period. Had we included part-year members, it is likely that expected costs estimates would have been less reliable, and that levels of agreement between rankings would have been different than those presented in our tables.
- We restricted our analyses to members who during the study period had at least one provider encounter that resulted in a diagnosis. Had we included nonusers in our analyses, or included members who used only pharmacy services, levels of agreement would have been lower than shown here because different risk-adjustment systems treat these two member categories in different ways.

Finally, although we also have used the terms economic profiling and practice efficiency, we must repeat that the definition of efficiency used in this paper

reflects the use of the term by profiling vendors and clients, and it is not the common definition accepted by health economists. Economists consider "efficiency" to be the relationship between quantity and mix of health care inputs and final health outcomes obtained (Palmer and Torgerson 1999). In contrast, "efficiency" in the physician profiling context refers to the relationship between the costs of inputs used to care for patients and the expected costs of those inputs, given the patients' identified diagnoses. Patient outcomes are not considered in this definition.

NOTES

1. We recognize that economists consider this definition of "practice efficiency" inappropriate (Palmer and Torgerson 1999). However, we use this definition here because it reflects the current terminology of dozens of commercial software vendors and hundreds of health plans.
2. The ACG system classifies nonusers and pharmacy-only users into one common ACG category; the DCG system assigns them a common risk score; both the CCI system and the GDG system ignore them; the BOI system assigns risk scores to pharmacy-only users, but ignores nonusers; and the ERG system assigns non-zero risk scores to some pharmacy-only users, but not to others.
3. We also looked at two other methods for dealing with high-cost outlier patients: top-coding costs at $50,000, and trimming from our analyses all patients whose costs exceeded $20,000. Analytical results obtained when using these methods did not differ from those presented here. Tables of results for the top-coded at $50,000 and trimmed at $20,000 analyses are available from the authors upon request.
4. With the exception of Table 2, all results presented in this section relate to outlier thresholds defined at 10 percent.
5. Detailed tables showing pair-wise agreement statistics between each risk-adjusted ranking and each of the other rankings are available from the authors upon request.
6. We performed these analyses only for adult PCP rankings, since there was no reason to suspect that the influence of practice efficiency metrics would differ between adult and pediatric primary care physicians. In these analyses, the outlier threshold was 10 percent.
7. Although we did not perform the analysis, it likely we could have obtained the same results by using hierarchical Bayesian modeling (Bronskill et al. 2002; Austin et al. 2001) to risk adjust panels while simultaneously controlling for panel size.

REFERENCES

Anderson, W. R., and K. Gilbert. 2002. *MEDecision Burden of Illness Methodology.* Santa Barbara, CA: MEDecision.

Austin, P. C., C. D. Naylor, and J. V. Tu. 2001. "A Comparison of Bayseian vs. a Frequentist Method of Profiling Hospital Performance." *Journal of Evaluation in Clinical Practice* 7 (1): 35–45.

Bronskill, S. E., S. L. Normand, M. B. Landrum, and R. A. Rosenheck. 2002. "Longitudinal Profiles of Health Care Providers." *Statistics in Medicine* 21 (8): 1067–88.

Cowen, M. E., D. J. Dusseau, B. B. Toth, C. Guisinger, M. X. Zodet, and Y. Shyr. 1998. "Casemix Adjustment of Managed Care Claims Using the Clinical Classification for Health Policy Research Method." *Medical Care* 36 (7): 1108–13.

Johns Hopkins University School of Hygiene and Public Health, Health Services Research and Development Center. 2000. *The Johns Hopkins ACG Case-Mix System Documentation and Application Manual. PC (DOS/WIN/NT) and Unix Version 4.5.* Baltimore, MD: Johns Hopkins University.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrica* 33 (2): 159–74.

Litton, L. M., F. A. Sisk, and M. E. Akins. 2000. "Managing Drug Costs: The Perception of Managed Care Pharmacy Directors." *American Journal of Managed Care* 6 (7): 805–14.

Nickerson, C., and R. W. Rutledge. 1999. "A Methodology for Choosing a Physician Profiling System." *Journal of Health Care Finance* 25 (2): 5–13.

Palmer, S., and D. J. Torgerson. 1999. "Definitions of Efficiency." *British Medical Journal* 318 (7191): 1136.

Sandy, L. G. 1999. "The Future of Physician Profiling." *Journal of Ambulatory Care Management* 22 (3): 1–16.

Solucient (formerly HCIA Health Chex). 1999. *The Clinical Complexity Index: Case Mix Methodology for the Peer-A-Med System.* Evanston, IL: Solucient.

Symmetry Health Data Systems. 2001. *Episode Risk Groups, ERG Users Guide.* Phoenix, AZ: Symmetry Health Data Systems.

Thomas, J. W., K. L. Grazier, and K. Ward. 2002. *A Comparative Evaluation of Risk-Adjustment Methodologies for Profiling Physician Practice Efficiency: Report to the Robert Wood Johnson Foundation*, HCFO Grant #36874. Ann Arbor, MI: Department of Health Management and Policy, University of Michigan.

———. Forthcoming. "Comparing Accuracy of Risk-Adjustment Methodologies Used in Economic Profiling of Physicians." *Inquiry.*