DOI: 10.1111/j.1475-6773.2005.00499.x

Should Episode-Based Economic Profiles Be Risk Adjusted to Account for Differences in Patients' Health Risks?

J. William Thomas

Objective. To determine whether additional risk adjustment is necessary in economic profiling of physicians when claims data are already grouped into episodes of care, and to measure effects of risk adjustment on cost efficiency rankings of physicians.
Data Sources. Four years of inpatient, outpatient, professional, and pharmacy claims data from a mixed model HMO.
Study Design. Claims data were processed through Symmetry Health Data Systems' episode treatment group (ETG) grouper to define episodes of care and Symmetry's episode risk group (ERG) software to define measures of patients' health risk scores. For each episode type (ETG), ETG-mean expected costs were calculated as the mean costs of all episodes of that type, and risk-adjusted expected costs were calculated using three alternative risk model formulations.
Data Collection. Within specialties, physicians were ranked from most cost efficient to least cost efficient, based on standardized difference between actual and expected costs. ETG-mean based rankings were compared with risk-adjusted rankings. Analyses were performed for cardiologists, family practitioners, general surgeons, and neurologists.
Principal Findings. With all three risk models, risk scores were essentially unrelated to episode costs in approximately three-fourths of episode categories (ETGs). In a sample of ETGs for which risks–costs relationships appeared to exist, split sample validation showed the relationships to be unstable or spurious in all except one ETG. Within specialties, risk-adjusted cost efficiency rankings differ little from ETG-mean adjusted rankings.
Conclusions. Depending upon the purpose for which economic profiling is performed, additional risk adjustment, beyond that already provided by episode grouping, may be unnecessary. Additional research may be needed to identify and validate ETG-level relationships between patient risks and episode costs.

Key Words. Risk adjustment, economic profiling, health care costs

After the medicare prospective payment system was introduced in the early 1980s, some hospitals—especially teaching institutions—were concerned about inadequate reimbursement for the care they delivered. This concern

was prompted by findings that: (a) severity-related cost differences existed among hospitalized patients, even after controlling for diagnosis-related group (DRG) (Horn et al. 1985; McMahon and Newbold 1986; Averill et al. 1992); and, (b) within individual DRG categories, teaching hospitals treated a more severe case mix than other hospitals (Berman et al. 1986; McNeil, Kominski, and Williams-Ashman 1988).[1] A similar question, based on similar assumptions, has been raised about possible bias of patient health status differences on economic profiles of physicians (Yi et al. 2002).

In economic profiling, health plans compare physicians' actual costs for services performed or ordered on behalf of patients to the expected costs of those services. Physicians whose actual costs are less than expected are considered cost efficient,[2] while those whose actual costs exceed the expected values are viewed as cost inefficient. On the basis of measured cost efficiency, physicians may find their compensation rates increased or decreased (Strunk and Reschovsky 2002), their placement in provider network tiers determined (Fronstin 2003), or their continued membership in provider networks terminated.

In recent years, it has become common for health plans to use *episode of care* as the unit of analysis for economic profiling. Thus, the first step in economic profiling is to process a claims database through software, such as Symmetry Health Data Systems' Episode-Treatment Group (ETG) system (Symmetry Health Data Systems 2005), that aggregates groups of claims into diagnostically and chronologically related episodes. The actual cost of each defined episode is calculated as the sum of costs associated with included claims, and episode expected costs are determined, typically as the mean cost of all episodes of the same type (e.g., ETG) in the database. Once episodes and their costs are defined, responsibility for each episode is assigned to a physician. Physicians' profiles are then constructed using average actual costs and average expected costs of attributed episodes.

In economic profiling of physicians, a concern is that both (a) differences in health status among patients may influence treatment costs within defined episode types and (b) average health status of patients treated may differ significantly among physicians. As a consequence, average cost per episode may be higher for some physicians than for others, not because of poorer cost efficiency, but rather because of the poorer health status of their patients. If episode

---

Address correspondence to J. William Thomas, Ph.D., Institute for Health Policy, Edmund S. Muskie School of Public Service, University of Southern Maine, 509 Forest Avenue, Suite 200 Portland, ME 04101-9300.

expected costs were adjusted to account for effects of differences in patient health status, the potential bias against such physicians might be removed.

The purpose of the study described here is to answer the following questions:

- Are episode costs related to patient health status, and, if so, can episode expected costs be risk adjusted to account for relationships that exist?
- Do cost efficiency rankings that are adjusted for patient health status differ from rankings that are not adjusted?

The first of these questions asks whether or not episode definitions themselves account adequately for relationships between patient health status and episode costs. If they do, then no additional risk adjustment is necessary. If they do not, then additional risk adjustment *may* be needed. Even if relationships between patient health status and episode costs exist, additional risk adjustment may not be needed if these relationships are too small and/or infrequent to influence profile rankings. This is the focus of the second question.

METHODS

Data for the project were provided by a university-owned, mixed model (group and independent practice association [IPA]) HMO in Southeast Michigan. Data included all professional, outpatient, inpatient, and pharmacy claims for members who were enrolled for the full 12 months of 1999, 2000, 2001, and/or 2002. The HMO was experiencing growth during the study years, expanding its membership, its geographic market area, and its network of physicians.

Claims data were processed through Symmetry's ETG software. The ETG system, which is widely used by health plans and other organizations for analyses of claims databases, is described by Rosen and Mayer-Oakes (1999). The resulting 4-year dataset included 658,646 completed episodes, each of which is associated with a specified member, and is characterized by: a category (ETG) and in some cases a subcategory (ETGSub), a start date, an end date, and a total cost. Episodes were divided among 693 ETG/ETGSub combinations (we will hereafter refer to these combinations simply as ETGs) and were associated with a total of 104,744 different members. Data were also processed through Symmetry's Episode-Risk Group (ERG) software to obtain

member level retrospective risk scores, which can be considered measures of member health status for the periods analyzed. Separate member-level retrospective risk scores were developed for each of the 4 data years. Over the 4-year period, scores averaged 2.31 with standard deviation of 3.11, and ranged from 0 (very healthy) to 64.44 (very unhealthy). Mean annual retrospective risk scores did not differ significantly across years.

For each type of service, a standard cost was developed using all claims in the 4-year database. For professional and outpatient claims, standard costs were calculated as arithmetic averages of actual costs of claims associated with each CPT-4 procedure code (for professional claims) or HCPCS or local revenue code (for outpatient claims). Pharmacy claims costs were standardized on the basis of national drug code (NDC) and amount dispensed, and costs of inpatient admissions were standardized on the basis of DRG. (Details of cost standardization procedures are given in Thomas, Ward, and Grazier 2004.) Once costs were standardized, episode costs were calculated by summing costs of individual claims—professional, outpatient, inpatient, and pharmacy—associated with that episode.

To control for the potentially distorting effects of very high cost or very low cost episodes on estimates of physicians' mean costs, costs within ETGs were Winsorized to the second percentile of category-specific costs for low outliers; high outlier episode costs were Winsorized at the 98th percentile (Hedges and Olkin 1985).

Four different methods were used to develop expected cost estimates for episodes:[3]

- *ETG Means:* Expected cost for an ETG was estimated as the average cost of all episodes in that ETG category. (This is the methodology commonly used by health plans and others when profiling physicians on the basis of episode costs.)
- *One Variable Regression Model:* Expected cost was estimated by regressing retrisk, (the ERG member-level retrospective risk score for the year in which the episode occurred) on episode costs. With episode as the unit of analysis, the regression equation is:

$$\hat{Y}_{ij} = a_i + b_i \times retrisk_j$$

where $\hat{Y}_{ij}$ is expected cost ETG i of member j; a_i is the intercept term for ETG i, $retrisk_j$ is the ERG retrospective risk score for member j, and b_i is the slope coefficient of $retrisk_k$.

- *Two Variable Regression Model:* Expected cost was estimated by regressing *retrisk* and $retrisk^2$ on episode costs. The regression equation is:

$$\hat{Y}_{ij} = a_i + b_i \times retrisk_j + c_i \times retrisk_j^2$$

  where $\hat{Y}_{ij}$, a_i, b_i, and $retrisk_j$ are the same as above, $retrisk_j^2$ is the square of $retrisk_j$ and c_i is its slope coefficient.

- *Dichotomized Risk Model:* Within each ETG, episodes were dichotomized using mean retrisk score, and expected costs were estimated as the average cost of all episodes within each of the two subgroups.

Three alternative risk-adjustment models were examined because it was not clear, a priori, which functional form would be most appropriate. Because two of the risk-adjustment methods involved regression, analyses (and therefore expected cost estimates) were limited to ETGs for which at least 26 episodes were available in the database. Further, because we wanted to produce full-year profiles, our analyses were limited to episodes that started and ended during the same calendar year.[4] With these two sets of restrictions, the final, 4-year analysis dataset included 595,425 episodes, representing 457 ETGs and a total of 104,335 members. Among the 457 ETGs, 106 included 1,000 or more episodes, 62 included 500–999 episodes, 160 included 100–499 episodes, and 129 included 99 or fewer episodes.

The three retrospective risk-based risk-adjustment models were compared on the basis of strength and stability of relationships between patient retrospective risk score and episode costs. Strength of relationship was indicated by R^2 value—regression-adjusted R^2 for the regression based models and squared correlation between actual and expected costs for the dichotomized risk model.[5] Relationship stability was assessed using split sample analysis for selected ETGs, with the first half of the sample used for estimating risk–cost relationships, and the second half used for validation testing of those relationships. ETG selection for these analyses was based on strength of risk–cost relationships (both one-variable and two-variable regression model $R^2 \geq 0.05$) and volume of episodes (the 12 largest volume ETGs satisfying the R^2 requirement were selected).

Many episodes involved services and charges, from multiple physicians. But profiling requires that, when possible, responsibility for each episode be assigned to a single physician. Health plans have used a variety of decision

rules for determining which physician, among those participating in an episode of care, should be assigned overall responsibility for episode associated costs. In this study, we used the following rule: responsibility for each episode's actual and expected costs was attributed to the physician who accounted for 50 percent of more of episode-related professional and prescribing costs. If no one physician accounted for at least 50 percent of professional and prescribing costs, the episode was not assigned.

After episode responsibility was assigned, physicians were ranked within specialties from most cost efficient to least cost efficient on the basis of *standardized cost difference,* which is the standardized difference between average actual cost and average expected cost for the sample of episodes managed by the physician. Using Z_{kj} to represent the standardized cost difference for the *k*th physician according to the *j*th model,

$$Z_{kj} = \frac{y_k - \hat{y}_{kj}}{\sigma_j / \sqrt{N_k}}$$

where y_k is average actual costs associated with the *k*th physician's set of episodes, $\hat{y}_{kj}$ is the average expected costs associated with these episodes according to the *j*th model, σ_j is the standard deviation of episode expected costs according to the *j*th model, and N_k is the number of episodes assigned to the *k*th physician. As we have shown (Thomas, Grazier, and Ward 2004), this measure is less likely than those that do not adjust for sample size to incorrectly identify smaller sized panels as high or low outliers.

As a measure of agreement between alternative model rankings, we used the *weighted* κ statistic proposed by Landis and Koch (1977), who indicate that appropriate interpretation of weighted κ values, which vary from 0.0 to 1.0, would be: 0–20 percent, slight agreement; 21–40 percent, fair agreement; 41–60 percent, moderate agreement; and 61 percent and greater, substantial agreement. For our weighted κ analyses, we partitioned physician rankings into quintiles, and measured level of agreement between pairs of quintile rankings.

Analyses included all physicians within a specialty who satisfied a specified minimum episode sample size criterion. In this paper, we show findings for economic profile rankings in four clinical specialties: cardiology, family practice, general surgery, and neurology. Based on results provided elsewhere (Thomas 2005), we profiled and ranked cardiologists having 20 or more attributed episodes, family practitioners with at least 125 episodes, general surgeons and neurologists with at least 25 episodes.[6]

RESULTS

Table 1 presents data on strength of relationships between episode costs and patient retrospective risk scores. R^2 values shown in this table are adjusted R^2 for each of the two regression models and squared correlations between actual and expected costs for the dichotomized model. (For the first model, R^2 would be, by definition, equal to 0.) The data show that none of the three risk models is explain more than a marginal amount of within-ETG cost variation ($R^2 \leq 0.025$) for approximately three-fourths of ETG categories. Summing the last two rows for each model shows that, with one-variable regression models, in only 6 percent of ETGs are R^2 values greater than 0.10; with two-variable regression models, R^2 values are greater than 0.10 for 8 percent of ETGs, and for dichotomous models, R^2 exceeds 0.10 in 4 percent of ETGs. The relatively small proportions of ETGs exhibiting even moderate risks–costs relationships

Table 1: Number of ETGs with Risk Adjustment R^2 in Specified Range, by Risk-Adjustment Model

				Number of Episodes per ETG		
Risk-Adjustment Model	*R^2 Range*	*Number of ETGs*	*Percent of ETGs*	*Minimum*	*Average*	*Maximum*
One-variable regression	≤ 0.01	271	0.59	25	1,400.8	70,426
	$\leq 0.01 \leq 0.025$	85	0.19	28	1,567.5	18,208
	$\leq 0.025 \leq 0.05$	49	0.11	28	1,388.6	14,745
	$\leq 0.05 \leq 0.075$	19	0.04	30	380.7	3,170
	$\leq 0.075 \leq 0.10$	4	0.01	46	245.0	800
	$\leq 0.10 \leq 0.15$	15	0.03	38	340.3	1,627
	≤ 0.15	14	0.03	26	54.6	114
Two-variable regression	≤ 0.01	228	0.50	25	1,477.1	70,426
	$\leq 0.01 \leq 0.025$	91	0.20	32	1,745.3	21,979
	$\leq 0.025 \leq 0.05$	67	0.15	30	1,225.3	14,745
	$\leq 0.05 \leq 0.075$	26	0.06	28	310.6	3,170
	$\leq 0.075 \leq 0.10$	5	0.01	53	357.0	1,331
	$\leq 0.10 \leq 0.15$	20	0.04	28	302.7	1,627
	≤ 0.15	20	0.04	26	68.1	279
Dichotomous risk score	≤ 0.01	224	0.49	25	1,775.8	70,426
	$\leq 0.01 \leq 0.025$	116	0.25	25	1,359.2	18,208
	$\leq 0.025 \leq 0.05$	58	0.13	28	544.3	4,514
	$\leq 0.05 \leq 0.075$	30	0.07	25	212.6	1,627
	$\leq 0.075 \leq 0.10$	11	0.02	30	61.2	149
	$\leq 0.10 \leq 0.15$	8	0.02	28	66.0	104
	≤ 0.15	10	0.02	26	38.5	59

ETG, episode treatment group.

suggest that Symmetry's ETG grouper performs well at defining episode categories in which patient health status exerts relatively little influence on diagnostic and treatment costs. Nevertheless, patient health status appears to be related to episode costs in some ETGs. Table 1 indicates that retrospective risk has at least a slight relationship ($R^2 > 0.025$) with episode costs in 22 percent of ETGs for one-variable regressions, 26 percent of ETGs for dichotomized model risk adjustment, and for 30 percent of ETGs for two-variable regressions. These ETGs account for 13.7, 6.6, and 16.7 percent, respectively, of all episodes in the study population.

To test stability of the risk-adjustment models, we selected 12 ETGs according to the criteria specified above. The ETGs are shown in Table 2. In each of these ETG categories, half of the episodes were used to estimate relationships between retrospective risk and episode costs. Estimation R^2 values suggest that modeled retrospective risk relationships are modestly predictive of episode costs in every ETG category considered. Weighted average difference between estimation and validation R^2 values is smallest for the two-variable regression model. However, even with this model, differences within individual ETGs are quite large—greater than 30 percent in 10 of the 12 ETGs listed. We cannot attribute these relatively poor validity results to the small size of the database used for the project, as ETG volume appears to be unrelated to risk model validation results. In the absence of validity evidence, we must consider that many, if not all, of the ETG-level risks–costs relationships are spurious.

Even if our analyses had validated all of the estimated relationships, the question of whether or not the frequency and strength of such relationships are sufficient to affect physicians' profile rankings would remain. If rankings are unaffected, then additional risk adjustment, beyond episode grouping, would be unnecessary. If risks–costs relationships exist and do affect judgments about physicians' relative cost efficiency performance, additional risk adjustment in profiling analyses should be performed. While we are unable to validate the existence of most ETG-level risks–costs relationships, we can investigate implications of those relationships, under the assumption that they do in fact exist.

Differences among risk-adjusted cost efficiency rankings are shown in Table 3 for four specialties in each of 4 years[7]. In all cases, weighted κ values in this table indicate substantial agreement between the rankings based on ERG risk scores and the rankings based on ETG means. Although agreement between pairs of rankings in most cases is not perfect (i.e., weighted $\kappa < 1.0$), weighted κ values are high in the substantial agreement range for all risk-adjustment models, indicating that for these four specialties risk-score-based and ETG-mean-based rankings are very similar.

Table 2: Split Sample Stability Analysis of Three Risk-Adjustment Models for Selected ETGs

		One-Variable Regression Models			*Two-Variable Regression Models*			*Dichotomized Models*		
ETG	*Total Number of Episodes*	*Estimation R^2*	*Validation R^2*	*Percent Difference (%)*	*Estimation R^2*	*Validation R^2*	*Percent Difference (%)*	*Estimation R^2*	*Validation R^2*	*Percent Difference (%)*
(a) Comparison of estimation and validation R^2 values										
0005	523	0.136	0.080	41	0.134	0.082	39	0.061	0.019	69
0032	377	0.161	0.022	86	0.157	0.022	86	0.023	0.014	38
0076	3,170	0.052	0.072	− 38	0.053	0.069	− 30	0.036	0.016	56
0310	284	0.046	0.181	− 293	0.050	0.032	36	0.038	0.012	69
0373	662	0.171	0.111	35	0.169	0.111	34	0.038	0.103	− 171
0374	1,627	0.147	0.059	60	0.149	0.069	54	0.046	0.065	− 39
0399	370	0.077	0.056	27	0.080	0.058	27	0.034	0.047	− 39
0450	1,331	0.065	0.073	− 12	0.066	0.100	− 52	0.031	0.036	− 14
0645	582	0.029	0.104	− 259	0.099	0.000	100	0.027	0.034	− 25
0737	279	0.11	0.168	− 50	0.140	0.249	− 78	0.010	0.047	− 384
0799	423	0.188	0.084	55	0.235	0.026	89	0.032	0.011	64
0812	800	0.100	0.096	4	0.098	0.096	2	0.033	0.077	− 133
Weighted average percent difference				− 17			9			− 16

Continued

Table 2: *Continued*

ETG	ETG Name	Episode Cost Range ($)		
		Minimum	*Mean*	*Maximum*
(b) Selected high volume ETGs				
0005	Major infectious disease except HIV and septicemia, w/o comorbidity	13	396	4,204
0032	Benign endocrine disorders of the pancreas	4	299	11,193
0076	Non-neoplastic blood disease, minor	5	456	6,549
0310	Other diseases of the veins	10	243	2,702
0373	Bacterial lung infections, with comorbidity	40	1,624	16,974
0374	Bacterial lung infections, w/o comorbidity	40	635	6,037
0399	Other inflammatory lung disease, w/o surgery	21	1,334	22,617
0450	Other infectious diseases of the intestines and abdomen	21	346	3,425
0645	Malignant neoplasm of the female genital tract, w/o surgery	34	1,835	45,456
0737	Closed fracture or dislocation of trunk, w/o surgery	32	795	7,871
0799	Minor specific procedures not classifed elsewhere	10	295	6,952
0812	Poisonings and toxic effects of drugs	21	756	7,392

No episodes used in Table 3 analyses were in ETGs for which full sample estimation $R^2 \geq 0.05$ for any of the risk models. However, a number of episodes used in these profiles were in ETGs for which patient risk was marginally related to episode cost (ETGs with risk model $0.025 < R^2 \leq 0.05$). It might reasonably be assumed that lower weighted κ values would be associated with larger proportions of physicians' episodes in ETGs with higher R^2 values. However, Table 3 demonstrates that this assumption is incorrect. For all three types of models, there appears to be no relationship between weighted κ scores and percentages of cases in ETGs with R^2 values > 0.025.

To better understand the meaning of "substantial agreement" between risk-score adjusted and ETG-mean adjusted cost efficiency rankings, in Table 4 we show data for the 14 cardiologists who had 20 or more attributed episodes during 2001. Several interesting patterns can be observed in these data. First, it is clear that the physicians identified as most cost efficient with ETG-mean adjusted expected costs remain as most cost efficient when expected costs are based on risk score calculations. Rankings of several of the six most cost efficient cardiologists change when expected costs are risk adjusted, but these six continue to be identified as the most cost efficient of the 14 physicians profiled. At the bottom of the cost efficiency rankings, five of the six cardiologist identified as least cost efficient remain at the bottom of the distribution when expected episode costs are adjusted to account for patient risk. The sole exception is physician K who is ranked 11th with ETG-mean adjustment and is replaced by physician G in all risk-adjusted cost efficiency rankings.

Table 4 also shows that physicians' average actual episode costs are not predictive of cost efficiency rankings, regardless of risk adjustment methodology. Average actual costs for physicians C and D, who are ranked second, third, or fourth depending upon risk-adjustment model, are among the lowest for the 14 cardiologists, while average actual cost for physician A, who is ranked first in cost efficiency by all models, is relatively high compared with those of the other physicians. Neither are average risk scores determinant of physicians' cost efficiency rankings. Physicians M and N, ranked last (13th and 14th) by all models, have higher average risk scores than 8 and lower average risk scores than 4 of the other profiled physicians. It is not actual costs or risk scores, but rather relationships between expected and actual episode costs that determine physicians' cost efficiency rankings. Table 4 shows that expected cost estimates based on ETG means are generally similar to those based on the various risk score models. With all risk-adjustment methodologies, physician A has the largest and physician H the smallest average expected costs. For two of the risk-adjustment methodologies—ETG-means and dichotomized

Table 3: Level of Agreement between ETG-Mean Adjusted and Risk-Adjusted Cost Efficiency Rankings for Four Specialties, by Year

			Risk Adjusted with One-Variable Regression Model		*Risk Adjusted with Two-Variable Regression Model*		*Risk Adjusted with Dichotomized Model*	
*Specialty**	*Year*	*Number of Physicians*	*Percent Cases in ETGs with $R^2 > .025$ (%)†*	*Weighted Kappa‡*	*Percent Cases in ETGs with $R^2 > .025$ (%)†*	*Weighted κ‡*	*Percent Cases in ETGs with $R^2 > .025$ (%)†*	*Weighted κ‡*
Cardiology ($N \geq 20$)	1999	4	0.0	1.00	0.0	1.00	0.0	1.00
	2000	11	2.3	0.68	2.3	0.76	2.3	0.76
	2001	14	5.3	0.80	5.3	0.80	5.3	0.90
	2002	11	0.7	0.87	0.7	0.87	0.7	0.87
Family Practice ($N \geq 125$)	1999	18	14.6	0.88	14.6	0.88	0.0	0.88
	2000	30	15.9	0.69	15.9	0.65	3.0	0.64
	2001	41	14.4	0.75	14.4	0.70	2.9	0.75
	2002	45	13.9	0.80	13.9	0.76	0.7	0.80
General Surgery ($N \geq 25$)	1999	2	25.9	1.00	25.9	1.00	25.9	1.00
	2000	9	14.6	0.94	14.6	0.94	14.6	0.94
	2001	9	9.6	0.94	9.6	0.94	9.6	0.94
	2002	9	14.2	1.00	14.2	1.00	14.2	1.00
Neurology ($N \geq 25$)	1999	4	1.1	1.00	1.1	1.00	1.1	1.00
	2000	10	6.2	0.70	7.7	0.70	6.2	0.70
	2001	11	3.5	0.97	5.0	0.97	3.2	0.97
	2002	11	2.7	0.76	3.0	0.76	1.4	0.76

*N refers to minimum number of episodes required for physician profiles.

†Full sample estimation R^2 values.

‡Measuring level of agreement between risk-adjusted rankings and ETG-mean adjusted rankings.

ETG, episode treatment group.

Table 4: Cost Efficiency Rank, by Risk-Adjustment Method: Cardiologists with ≥ 20 Episodes, Year = 2001

				Risk Adjusted with ETG Means			*Risk Adjusted with One-Variable Regression Model*			*Risk Adjusted with Two-Variable Regression Model*			*Risk Adjusted with Dichotomized Model*		
Provider	*Number of Episodes*	*Avg. Risk*	*Avg. Actual Cost*	*Avg. Exp. Cost*	*Std. Cost Dif.*	*Eff. Rank*	*Avg. Exp. Cost*	*Std. Cost Dif.*	*Eff. Rank*	*Avg. Exp. Cost*	*Std. Cost Dif.*	*Eff. Rank*	*Avg. Exp. Cost*	*Std. Cost Dif.*	*Eff. Rank*
A	37	4.2	$1,604	$3,538	−4.381	1	$2,379	−2.553	1	$2,440	−2.756	1	$2,454	−2.808	1
B	32	3.7	$1,135	$2,101	−2.033	2	$1,285	−0.458	4	$1,306	−0.523	4	$1,324	−0.581	3
C	23	2.9	$854	$1,945	−1.947	3	$1,175	−0.835	2	$1,176	−0.837	2	$1,157	−0.789	2
D	70	3.1	$740	$1,041	−0.940	4	$872	−0.600	3	$891	−0.686	3	$831	−0.416	4
E	27	4.1	$1,450	$1,824	−0.723	5	$1,259	0.538	6	$1,261	0.532	6	$1,361	0.253	5
F	28	4.8	$1,467	$1,768	−0.591	6	$1,362	0.301	5	$1,407	0.172	5	$1,316	0.436	6
G	33	2.2	$1,485	$1,420	0.140	7	$961	1.631	10	$978	1.580	10	$1,048	1.365	9
H	71	2.1	$800	$731	0.216	8	$576	1.022	7	$583	0.988	7	$590	0.960	7
I	20	2.0	$1,366	$1,177	0.314	9	$754	1.483	9	$744	1.506	9	$761	1.470	10
J	61	1.9	$1,248	$1,139	0.317	10	$818	1.817	11	$796	1.910	11	$820	1.814	11
K	24	2.5	$1,098	$799	0.545	11	$693	1.075	8	$703	1.047	8	$697	1.067	8
L	51	2.6	$1,838	$1,345	1.310	12	$956	3.410	12	$953	3.423	12	$938	3.489	12
M	60	3.2	$2,017	$1,440	1.664	13	$1,045	4.077	14	$1,055	4.039	14	$977	4.376	14
N	70	3.2	$2,049	$1,454	1.855	14	$1,158	4.038	13	$1,184	3.922	13	$1,110	4.268	13

Avg, average; Std. Cost. Dif., standardized cost difference; Eff., efficiency; Exp., expected.

model—five of the six largest average expected cost values are associated with physicians A through F, while for each of the two regression-based risk-adjustment methodologies four of the six largest average expected cost values are associated with these physicians.

Although not presented here, ranking statistics for family practitioners, general surgeons, and neurologists show similar patterns to those in Tables 4.[8] Adjustment for relationships between retrospective risk and episode costs leads to small alterations in cost efficiency rankings with all models, but no major shifts in relative rankings occur.

Discussion and Conclusion

In this study, we examined four different methodologies for estimating episode expected costs. Although all risk-adjusted rankings were in substantial—and in some cases perfect—agreement with ETG-mean based rankings, differences did exist. Of the risk-adjustment methodologies considered, two-variable regression produced the strongest relationships between retrospective risk and episode costs. However, even with the two-variable regression model, essentially no relationship was found between patient health status and episode costs in approximately three-quarters of symmetry's ETG categories. In less than 10 percent of ETGs do risks–costs relationships produce R^2 values that exceed 0.10. Furthermore, with all models, split sample validation indicates instability, and possibly spuriousness, in most modeled relationships.

Other findings of this study can be summarized as follows:

- Symmetry Health Data System's ETG grouper classifies health care claims into episode groupings in which patient health status exerts relatively little influence on diagnostic and treatment costs. For all three of the risk models considered in this study, none of the economic profiles in any of the four specialties included ETG categories for which risk–cost model R^2 values were 0.05 or higher.
- No relationship was found between percentage of episodes in ETG categories with nonzero risks–costs relationships and the likelihood that risk adjustment will cause changes in physicians' cost efficiency rankings.
- Physicians' cost efficiency rankings are not biased with respect to average retrospective risk score of patients treated. This is true even when episode expected costs are based on ETG means. Thus, the

"my patients are sicker" argument cannot be used credibly to explain substandard cost efficiency performance.

- Risk-adjusted cost efficiency rankings agree substantially with ETG-means based rankings, for all years and all four specialties examined. In several cases, there was perfect agreement between pairs of rankings.
- When changes in rankings occurred because of risk adjustment, changes were small. Specific changes in rankings differed with different risk models.

When evaluating cost efficiency performance of network physicians, health plans typically compare actual costs of services provided or ordered by each physician to the costs expected for those services, given the types of episodes being managed. Expected costs of episodes are calculated as average actual costs of all episodes of the same type (i.e., ETG) in the claims dataset being analyzed. With nominal cost standards such as these, performance assessments depend not only on the claims for which physicians are themselves responsible, but also on other claims in the dataset, as these other claims help determine episode expected costs. Use of nominal cost standards might also suggest that if additional risk adjustment is to be performed, ETG-level risk–costs relationships should be estimated in the dataset being analyzed. However, we found that most of the risks–costs relationships estimated with a 4-year database from a small HMO are spurious or unstable, in that, if they exist at all, the relationships almost certainly are sensitive to inclusion or exclusion of small numbers of episodes in profiling analyses. To protect against such sensitivity, risk-adjustment, if performed, should utilize validated risk models, and this suggests that models must be estimated and validated with datasets larger than that used in this study. If externally estimated and validated models are used in when profiling physicians, episode expected costs will still be nominal because models will be recalibrated to the dataset being analyzed.

A possible concern related to generalizability of our findings is that our analyses were based on only one patient risk measure, the ERG retrospective risk score. We have shown in an earlier study (Thomas, Grazier, and Ward 2004) that this measure produces similar profiling results to those obtained with other commonly used risk measures (e.g., Adjusted Clinical Groups [ACGs] from Johns Hopkins University and Diagnostic Cost Groups [DCGs] from D_xCG Inc.). For the current study, in addition to developing health plan members' ERG retrospective risk scores, we also processed the database thorough DCG software from D_xCG Inc. For each of the four data years, we

found ERG and DCG retrospective risk measures to be highly correlated (average $r = 0.75$ over the 4-year period), and we are confident that our findings on risk adjustment would have been similar had we utilized the DCG risk measure instead of the ERG risk measure in our analyses.

Another issue regarding the generalizability of out findings—and this is an important caveat—is that our analyses were performed on data from a single, university-owned health plan. We do not know whether or not the strength and stability of within-ETG risks–costs relationships would be similar if estimated in different datasets. Similarly, it is possible that effects of such relationships on cost efficiency rankings could be greater or lesser than shown here. We believe that a priority for future research should be investigation of these relationships in large datasets.

ACKNOWLEDGMENT

This study was supported by Grant 047789 from the Robert Woods Johnson Foundation Health Care Financing and Organization (HCFO) Program.

NOTES

1. Other research has suggested that, controlling for DRG, teaching institutions do not, in fact, treat more severely ill patients than nonteaching hospitals (Welch 1987; Goldfarb and Coffey 1987).
2. A consensus conference convened in September 2005 by the Ambulatory Care Quality Alliance and National Committee for Quality Assurance determined that the relative-resource-use measure described in this paper should be termed cost efficiency, and that this should be differentiated from efficiency, a term long used by economist to refer to the cost of resources utilized in achieving a given outcome or benefit to the patient.
3. To control for skewness in episode cost distributions, in preliminary analyses, we estimated risk models using log transform of episode costs as the dependent variable. Findings from subsequent analyses using transformed actual and expected cost estimates did not differ greatly from those based on untransformed data, and as a result are not presented here. Results of these analyses are available from the author upon request.
4. The 595,425 single year episodes represented 89.1 percent of the total 668,234 episodes in the database. Among episodes that began in one year and ended in a different year (10.8 percent of the total), the most frequently occurring conditions were benign hypertension (ETG 0281), hyperlipidemia (ETG 0047), and minor depression (ETG 0096). Including these cross-year episodes in our analyses would have required estimating cross-year retrospective risk scores, for example by

calculating simple or time-weighted averages of single year risk scores. Because within-ETG relationships between these calculated risk scores and episode costs might be different than risk adjustment relationships for single year single year episodes, we chose to exclude cross-year episodes from our analyses.

5. For the dichotomized risk model, R^2 is calculated as $\sum (y_i - \hat{Y}_i)^2 \Big/ \sum (y_i - \bar{y})^2$ where y_i is actual cost of episode i, $\hat{Y}_i$ is expected cost of episode i, and $\bar{y}$ is mean actual cost of all episodes within the ETG.
6. Analyses were also performed with other minimum sample sizes for each specialty. Results were similar to those reported below for the minimum sample sizes specified here.
7. Number of episodes per physician for profiles represented in Table 3 ranged from 20 to 105, with mean = 47.1 for cardiologists; from 125 to 418, with mean = 190.5 for family practitioners; from 25 to 89, with mean = 44.7 for general surgeons; and from 25 to 163, with mean = 54.1 for neurologists.
8. Tables are available from the authors upon request.

REFERENCES

Averill, R. F., T. E. McGuire, B. E. Manning, D. A. Fowler, S. D. Horn, P. S. Dickson, M. J. Coye, D. L. Knowlton, and J. A. Bender. 1992. "A Study of the Relationship between Severity of Illness and Hospital Cost in New Jersey Hospitals." *Health Services Research* 27 (5): 587–606.

Berman, R. A., J. Green, D. Kwo, K. F. Safian, and L. Botnick. 1986. "Severity of Illness and the Teaching Hospital." *Journal of Medical Education* 61 (1): 1–9.

Fronstin, P. August 2003. *Tiered Networks for Hospital and Physician Health Care Services. EBRI Issue Brief Number 260.* Washington, DC: Employee Benefits Research Institute.

Goldfarb, M. G., and R. M. Coffey. 1987. "Case-Mix Differences between Teaching and Nonteaching Hospitals." *Inquiry* 24 (1): 68–84.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis.* Orlando, FL: Academic Press.

Horn, S. D., G. Bulkley, P. D. Sharkey, A. F. Chambers, R. A. Horn, and C. J. Schramm. 1985. "Inter-Hospital Differences in Severity of Illness. Problems for Prospective Payment Based on Diagnosis-Related Groups (DRGs)." *New England Journal of Medicine* 313 (1): 20–4.

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

McMahon, L. F. Jr, and R. Newbold. 1986. "Variation in Resource Use within Diagnosis-Related Groups. The Effects of Severity of Illness and Physician Practice." *Medical Care* 24 (5): 388–97.

McNeil, B. J., G. F. Kominski, and A. Williams-Ashman. 1988. "Modified DRGs as Evidence of Variability in Patient Severity." *Medical Care* 26 (1): 53–61.

Rosen, A. K., and A. Mayer-Oakes. 1999. "Episodes of Care: Theoretical Frameworks versus Current Operational Realities." *Journal on Quality Improvement* 25 (3): 111–28.

Strunk, B. C., and J. D. Reschovsky. 2002. "Kinder and Gentler: Physicians and Managed Care, 1997–2001." Center for the Study of Health System Change (CSHSC). Tracking Report: Results from the Community Tracking Study. No. 5.

Symmetry Health Data Systems. 2005. "Episode Treatment Groups: An Illness Classification and Episode Building System." Available at http://www.symmetry-health.com/ETGTut_Desc1.htm

Thomas, J. W. 2005. "Sample Size Considerations in Economic Profiling of Physician Specialists." Portland, ME: Institute for Health Policy, Muskie School of Public Service, University of Southern Maine.

Thomas, J. W., K. L. Grazie, and K. Ward. 2004. "Economic Profiling of Primary Care Physicians: Consistency among Risk Adjusted Measures." *Health Services Research* 39: 985–1004.

Thomas, J. W., K. L. Grazier, and K. Ward. 2004. "Comparing Accuracy of Risk Adjustment Methodologies Used in Economic Profiling of Physicians." *Inquiry* 41: 218–31.

Thomas, J. W., K. Ward, and K. L. Grazier. 2004. "Using Physician Profiling Software to Evaluate the Practice Efficiency of Physician Specialists." Final Report to the Robert Wood Johnson Foundation Health Care Organization and Financing Program Grant #047789. Portland, ME: Institute for Health Policy, Muskie School for Public Service, University of Southern Maine.

Welch, W. P. 1987. "Do All Teaching Hospitals Deserve an Add-On Payment under the Prospective Payment System?" *Inquiry* 24 (3): 221–32.

Yi, R., J. Haughton, F. Cecere, and M. Rubin. 2002. "Integration of Episode Analysis and Risk Assessment for Provider Profiling." Boston: AdvanceMed. Inc. and DxCG Inc.