A Nonparametric Statistical Method That Improves Physician Cost of Care Analysis

Brent A. Metfessel and Robert A. Greene

Objective. To develop a compositing method that demonstrates improved performance compared with commonly used tests for statistical analysis of physician cost of care data.

Data Source. Commercial preferred provider organization (PPO) claims data for internists from a large metropolitan area.

Study Design. We created a nonparametric composite performance metric that maintains risk adjustment using the Wilcoxon rank-sum (WRS) test. We compared the resulting algorithm to the parametric observed-to-expected ratio, with and without a statistical test, for stability of physician cost ratings among different outlier trimming methods and across two partially overlapping time periods.

Principal Findings. The WRS algorithm showed significantly greater within-physician stability among several typical outlier trimming and capping methods. The algorithm also showed significantly greater within-physician stability when the same physicians were analyzed across time periods.

Conclusions. The nonparametric algorithm described is a more robust and more stable methodology for evaluating physician cost of care than commonly used observed-to-expected ratio techniques. Use of such an algorithm can improve physician cost assessment for important current applications such as public reporting, pay for performance, and tiered benefit design.

Key Words. Statistical methods, physician profiling, nonparametric statistics, cost-efficiency, efficiency index

In the last two decades, physician performance measurement has gained momentum. Many managed care organizations use some type of physician measurement methodology, and they typically evaluate quality and cost of care. With the rising cost of health care, as well as increasing patient responsibility for and control over health care spending, physician practice pattern assessment has become increasingly important to multiple stakeholders, including consumers, health care providers, government agencies, and health plans.

As a result, there is a strong need for robust methods to evaluate physician performance. Results of such analyses have a variety of applications, including pay-for-performance programs, consumer-directed health care, tiered benefit design, and public reporting.

Accordingly, a variety of methods have been developed to compare physicians with their peers in terms of cost of care. These methods include algorithms to control for a physician's mix of patients using methods of disease classification and severity adjustment. Typically, an "expected" cost is calculated, which is based on a weighted mean for a physician's specialty peer group after adjusting for case mix and severity, and compared with a physician's observed cost (Thomas, Grazier, and Ward 2004; Sandy, Rattray, and Thomas 2008). The observed-to-expected ratio, also known as the O : E ratio or "efficiency index," is the most well-known measure of cost of care, with ratios greater than 1.0 indicating higher costs than an average practice pattern and ratios less than 1.0 indicating lower costs (Greene, Beckman, and Mahoney 2008). The O : E ratio creates a composite measure across multiple types of episodes.

Point estimates, or "snapshots," of relative efficiency typically reflect a sample of a physician's activity in time and place. It is important then to apply a statistical test of uncertainty to indicate how likely the point estimate is to differ from the average across peers or an alternative benchmark (Adams et al. 2010a). Up to this time, when statistical tests have been employed, they are typically parametric tests such as a *t*-test or the creation of a confidence interval (e.g., Shine et al. 2001; Adams et al. 2010a).

Although assessing physician cost patterns through parametric tests is a generally accepted practice, these methods have a number of limitations. Such tests assume that the data show a distribution approximating normality. However, it is generally accepted, and can easily be shown empirically, that health care cost data are highly skewed to the right (Kilian et al. 2002; Wang et al. 2003; Nixon and Thompson 2004; Tina Shih, Xu, and Elting 2007). This is a natural consequence of costs of care having a lower bound at a low nonzero number but no upper bound.

Trimming (removing) observed costs that are above a certain threshold, such as \$25,000 or at the top 2.5 or 5 percent of costs, partially addresses issues concerning cost data distributions, but the distributions still maintain much of

Address correspondence to Brent A. Metfessel, M.D., M.S., Senior Medical Informaticist, Clinical Analytics, UnitedHealthcare, MN012-S117, 5901 Lincoln Drive, Edina, MN 55436; e-mail: Brent_a_mETFessel@uhc.com. Robert A. Greene, M.D., is with the UnitedHealthcare, Edina, MN.

their skew. Capping outliers (i.e., setting costs to a specified upper limit) has the advantage over trimming in that this method maintains some outlier information. In addition, although there are generally accepted approaches to trimming and capping (Barnett and Lewis 1994; Sheskin 2000), the particular method and threshold chosen remain a judgment call.

The need to create a composite across multiple types of cases introduces an additional issue. Composites are useful for understanding a physician's general practice pattern, to increase the effective sample size, and for consumer transparency (O'Brien et al. 2007; Scholle et al. 2009). However, both the scale and the distribution of each condition's costs may vary. For example, episodes of acute pharyngitis treated in a physician's office might typically reflect the cost of an office visit, a streptococcus test, and perhaps an antibiotic. The episodes will be relatively uniform in cost and range up to a few hundred dollars. An episode of pneumonia, however, could include multiple visits, radiological imaging, emergency room, or inpatient costs, and typically could range in cost from a few hundred to thousands of dollars.

A statistical method for physician cost of care assessment should address the issues raised above: highly skewed data distributions, potential effects of outlier processing choices, and creation of composite measures across conditions of varying average costs. Ideally there would be a "gold standard" for physician cost measurement, but no such standard exists. Although previous research has used statistical reliability as a proxy for physician classification accuracy (Adams et al. 2010b), statistical reliability only estimates risk of misclassification and in the absence of a gold standard does not tell whether a particular physician's cost classification (such as "high cost," "low cost," or "average cost") is correct.

Instead, one can compare statistical classification methods in terms of relative robustness and stability. An improved method would be less sensitive to variations in data processing methodologies such as outlier trimming or capping and would exhibit greater stability across multiple time periods. Increased stability in physician rating systems is desirable because it increases the probability that systematic physician practice pattern variation will be uncovered, making physician measurement more useful for public reporting, quality improvement, and other initiatives.

While a recent paper demonstrated that a parametric test is superior to an absolute threshold for classifying physicians as lower or higher cost, the authors suggest that a nonparametric test would be preferable (Adams et al. 2010a). Distribution-free or "nonparametric" statistical tests have been used to analyze

skewed data with greater validity and reliability (Wang et al. 2003; Barsky, Orav, and Bates 2005; Lee et al. 2005; Visser et al. 2006; Mino et al. 2007).

One commonly used test, the Mann–Whitney *U*-test and its mathematical equivalent, the Wilcoxon rank-sum (WRS) test, is the nonparametric counterpart to the *t*-test for two independent samples. The WRS test transforms the data into relative ranks to test two groups of data for statistically significant differences. The distance of the sum of ranks of each group from the expected sum of ranks (if all values were at the median) forms the basis for statistical inference. In addition to being a test for differences in medians, the WRS also accounts for differences in distribution shape, spread, and location of data items (Hart 2001).

We adapted the WRS test to address the issues and needs described above. To create a composite across different conditions, we scaled episode costs to percentiles within a set of comparable episodes. Because the goal was statistical testing rather than an analysis of effect size, we were willing to trade off some loss of information from the native values in return for a test not bound to assumptions about the data distribution. A percentile scale established a uniform distribution that also should be less sensitive to extreme values commonly found in raw medical cost data. To test the performance of this method, we compared the WRS to observed-to-expected ratios, a common parametric statistic for physician cost-efficiency analysis, in terms of stability across time and across trimming methodologies.

The goal of this work was to determine whether in fact this adaptation of the WRS test is more robust, stable, and useful than commonly used techniques for physician cost-efficiency analysis. The improved operating characteristics we demonstrate are desirable from a policy viewpoint because they meet critically important stakeholder expectations for physician cost assessment.

METHODS

This section describes an adaptation of the WRS test for two independent samples so that case mix, severity, and other adjustments are preserved. The WRS test itself is described in detail in common reference works (e.g., Sheskin 2000). This adaptation consists of two steps. First, for each set of comparable episodes, the index physician (the physician being evaluated) and peer group episode costs are transformed into percentiles. Second, all sets of episodes containing the index physician's episodes are combined into a superset of episodes, sorted by percentile, and assigned ranks. The WRS is then applied

to the ranks to determine whether the physician's episode percentiles are significantly different from the expected rank sum of the superset as a whole.

There are a number of case mix and risk adjustment systems available for use in physician cost of care evaluation. To construct a practical example, we chose a widely utilized commercially available episode grouper, the Symmetry Episode Treatment Groups software system (*ETG*[®] v7.0, OptumInsight, Eden Prairie, MN, USA). *ETG* is an illness classification system that takes into account disease class along with complications, comorbidities, defining surgeries and treatments, and disease severity (for further details on *ETG*, see OptumInsight 2011).

In our model each *ETG* episode is assigned (attributed) to the one physician who generated the highest percentage of costs in an episode. We used only those episodes where the highest percentage of involvement (as measured by cost) by a single physician was at least 30 percent of total episode allowed amount. The attribution methodology reflects generally accepted practices (Sandy, Rattray, and Thomas 2008).

After episodes are grouped and assigned to physicians, the next step is to sort episodes into sets with common characteristics. The combination of *ETG*, severity, treatment indicator, and pharmacy benefit status creates comparable episodes that we will call a "treatment set" (Figure 1). For statistical stability purposes, in all analyses, we included only treatment sets consisting of *ETGs* relevant to internal medicine and containing at least 50 total episodes.

Next, episodes in each treatment set in which the index physician has episodes are sorted from lowest to highest dollar cost. Based on the sort order, a cost percentile is assigned to each episode with the result that the index physician's data are compared with the peer group or reference data in the treatment set.

Assigning percentiles to episode costs has two purposes. Different conditions span widely different cost ranges as discussed above. Converting all treatment sets to a uniform scale allows us to combine treatment sets into a superset to which we apply ranks. The other important function is to address high-cost outliers in a natural fashion. No case can be assigned more than 100 "units" (in practical terms, we used 99.9999 percent).

Once the percentiles in each relevant treatment set are established, percentile values pertaining to physician and peer group episode costs are combined across all relevant treatment sets forming a final superset for establishment of ordinal ranks. For the purposes of this discussion, we will term the superset the "combined cost set." Preserving the original treatment set percentiles in the combined cost set maintains the integrity of case mix and risk adjustment through the process. Once the combined cost set is estab-

Figure 1: Example of the Two-Step Ranking System Used to Create Rank Sums for the Wilcoxon Rank-Sum Test for a Physician (Dr. A)

Calculation of Wilcoxon statistic for Dr. A:

- Sum the observed ranks for Dr. A:
 $\text{RankSum}_{\text{Dr. A}} = 3 + 10 + 13.5 + 15 + 16 + 19 + 21 + 22 + 23 + 24 + 26 = 192.5$
- Calculate expected values
 - Expected Rank Sum for Dr. A:
 $\text{RankSum}_{\text{Exp}} = (n_{\text{Dr. A}} * (n_{\text{Tot}} + 1)) / 2 = 148.5$
 - Standard Deviation of Expected Rank Sums:
 $\text{Stdev}_{\text{Exp}} = \text{Sqrt}((n_{\text{Dr. A}} * n_{\text{Peer}} * (n_{\text{Tot}} + 1)) / 12) = 19.27$
- Calculate Z-score for Dr. A
 $\text{Z-score}_{\text{Dr. A}} = (\text{RankSum}_{\text{Dr. A}} - \text{RankSum}_{\text{Exp}}) / \text{Stdev}_{\text{Exp}} = 2.28$
 $(p < 0.05, 2\text{-tailed test})$
- Interpretation: Dr. A is less cost-efficient than the peer (reference or comparison) group with similar case-mix

Each row represents one episode of care. For further details, see Methods section

lished, the percentiles are sorted in ascending order and each percentile is then assigned an ordinal rank in the superset, to which the WRS test is applied. Ties in percentiles are given the mean of the ordinal ranks as is standard for prepar-

ing data before applying the WRS test (Sheskin 2000). Figure 1 illustrates the process of assigning percentiles and then moving episodes into a combined cost set, with calculation of the WRS statistic for an example physician.

The algorithm generates a z -score when the normal approximation for the distribution of rank sums is used, which in practice requires 10 or more observations in each of the two groups to be considered applicable. Although the physician and reference (peer group) episodes are combined for the ordinal ranking step, the ranks for the physician being evaluated are summed separately from the peer group episodes to calculate the observed rank sum. The observed rank sum is compared with the expected rank sum (i.e., if all the physician's episodes were at the median percentile) and a z -score is calculated. Even though the dollar cost data have a skewed distribution toward high-cost values, the assigning of ranks to the data results in a transformation to a normally distributed variable, which in this case is the sampling distribution of the sums of the ranks.

To illustrate the practical significance of the WRS methodology, its performance was compared with the O : E ratio method of physician cost analysis in three areas: a simulation analysis to characterize the native distributions from the two methods, a comparison of the two methods for within-physician stability between different outlier processing or mitigation strategies, and a comparison of the methods for within-physician stability over time.

For the O : E method we proceeded along the lines described in the introduction. We summed the observed costs for each physician. We created peer group means for each treatment set, and assigned that as the expected cost for each observed episode cost in that treatment set. We then divided the total observed by the sum of their treatment set means to calculate the O : E ratio. The methodology is explained in further detail elsewhere (Greene, Beckman, and Mahoney 2008; Sandy, Rattray, and Thomas 2008).

Data Characteristics

The dataset consists of episodes generated by *ETG* version 7.0 from a large metropolitan area for the specialty of internal medicine. We chose internal medicine because internists handle a wide variety of conditions across a wide range of costs.

Cost values for services were based on health plan allowed amounts. Episodes with allowed amounts below an estimated 2.5 percent level for each *ETG*, based on benchmarks supplied by OptumInsight with the *ETG* software, were excluded from this analysis, since these episodes often contain systematic data anomalies.

Only complete and full year episodes were used. The simulation analysis, representing 453,435 episodes, covered the time period from December 1, 2004, to February 19, 2008.

For the within-physician stability analysis over time, two partially overlapping time periods or “reporting periods” were used, duplicating common practice for systems that update data yearly. Time period 1 contained episodes for the 2-year period from August 1, 2007, to July 31, 2009. Time period 2 contained episodes for the 2-year period from August 1, 2008, to July 31, 2010. For within-physician stability across outlier methodologies, we used episodes from time period 2. For all within-physician stability analyses, we used actual physician attributed episodes. Physicians needed to have at least 30 episodes for both time periods to be included, consistent with National Committee for Quality Assurance (NCQA) requirements (NCQA 2008). This is also the sample size at which the normal distribution becomes a good approximation for the t -distribution. There were 729 physicians who met these criteria. There were 333,153 episodes in time period 1 and 317,115 in time period 2, corresponding to an average of 457 episodes per physician in time period 1 and 435 episodes per physician for time period 2.

Simulation Analysis

The central limit theorem states that for a valid statistical test, given a large enough sample size, the sampling distribution of the mean will asymptotically approach normality. In general, the better a statistical test handles deviations from its assumptions, the lower the sample size needed to approach normality (robustness of the test).

To evaluate this, we compared the native distribution of the O : E ratios and the O : E ratio t -test results (using the standard error calculation outlined in Adams et al. 2010a) to the WRS z -score output using a basic simulation algorithm. For each run, 300 simulated physician panels were created, consisting of randomly selected episodes. Panel sizes were held constant within each run. Multiple runs were performed to study how varying the panel sizes (between 10 and 3,000 episodes per panel, with the remaining episodes used in the reference group) affects the sampling distribution. The distributions were evaluated using the following four goodness-of-fit tests for normality:

1. Shapiro–Wilk (Shapiro and Wilk 1965)
2. Kolmogorov–Smirnov (Sheskin 2000)
3. Cramer–von Mises (Sheskin 2000)

4. Anderson–Darling (Anderson and Darling 1952)

All four normality tests needed to show a p -value $\geq .05$ to qualify as consistent with a normal distribution. Separate runs with capped (at the upper 95 percent level for each treatment set, also known as “modified Winsorization”) and noncapped data were compared as well.

Stability Analysis

For the stability analysis, the WRS algorithm was compared with the O : E method under two different high outlier processing strategies commonly practiced in the industry after data are grouped: (1) outlier trimming (exclusion) at the 95th percentile for each treatment set, (2) outlier capping (modified Winsorization) at the 95th percentile for each treatment set. The methods were compared with no outlier exclusion or capping and to each other. To maintain the integrity of comparisons, the parametric O : E ratio (without a statistical test) was compared with the WRS observed-to-expected rank-sum ratio (WRS O : E ratio) without a statistical test, and the O : E t -test results were compared with the WRS z -score.

The stability of physician scores across the outlier processing methods (time period 2), and across the two reporting time periods, was evaluated using Spearman rank-order correlation coefficients. Stability of physician categorical cost ratings (“high cost,” “average cost,” “low cost”) across time periods was evaluated using a weighted kappa agreement statistic, where disagreement across two categories (i.e., “high cost” in one time period and “low cost” in the other period) are weighted more heavily than disagreement across a single category (“high cost” or “low cost” in one time period to “average cost” in the other period). Overall concordance and discordance was evaluated using the z -test for two independent proportions.

All data analyses were performed using SAS 9.1[®] and SAS Enterprise[®] Guide 4 (Cary, NC, USA).

RESULTS

Simulation Analysis

For the O : E ratio, the native distribution was not consistent with normality up to 3,000 episodes per simulated physician when no outlier capping was used. With capping at the 95th percentile, the native distribution became

consistent with normality at $>1,100$ episodes per simulated physician, and the O : E ratio t -test distribution became consistent with normality at about 1,000 episodes per physician. The WRS algorithm was consistent with normality at 10 episodes per simulated physician on both the noncapped and the capped datasets.

Stability of the Measures across Varying Outlier Processing Strategies

Table 1 shows the physician measurement correlations between no outlier processing, trimming (removal) of outliers at the 95 percent level, and capping at the 95 percent level. Correlations of measures between the outlier processing strategies were significantly higher for the WRS O : E ratio than for the parametric O : E ratio ($p < .0001$ for all correlations). The WRS normal approximation (z -score) showed significantly higher correlations across outlier processing strategies compared with the O : E ratio with a t -test ($p < .0001$ for all correlations). Furthermore, for the WRS, there was essentially no difference between using the raw data and using either of the outlier processing methodologies (Spearman coefficients 0.988 or higher for comparisons).

Stability of the Measures across Time

Table 2 shows the correlations between time period 1 and time period 2 for the statistical methodologies in question using three different outlier processing strategies (no processing, capping at the 95 percent level, removal at the 95 percent level). All correlations were higher for the WRS O : E ratio than for the parametric O : E ratio ($p < .0001$), as well as being higher for the WRS z -score compared to the parametric O : E ratio t -test ($p < .0001$).

For stability over time analysis, the ratings were adjusted so that each of the three ratings categories had the same proportions for both statistical tests for both time periods (adjusted ratings). Physician movement between rating categories could then be compared between the two time periods on a level playing field. Therefore, for both time periods, physicians were given ratings of "high cost" (25 percent of physicians), "average cost" (50 percent of physicians), and "low cost" (25 percent of physicians), similar in fashion to Adams et al. (2010a) based on sorted and ranked lists of the z -scores from the WRS and the t -test scores for the O : E ratio, using the 95th percentile capped datasets.

Data for the concordance of physician ratings across time periods are shown in Table 3. The weighted kappa (agreement between time period 1 and

Table 1: Stability of Within-Physician Results for the Parametric O : E Ratio and Wilcoxon Rank-Sum Algorithms between Different Outlier Processing Methods (Time Period 2, Spearman Rank-Order Correlation Coefficients)

Outlier Processing Method		Parametric O : E Ratio (Raw)		WRS O : E Rank-Sum Ratio (Raw)		Parametric O : E Ratio (t-Test)		WRS Z-Score	
Method 1	Method 2	Spearman Coefficient	CI	Spearman Coefficient	CI	Spearman Coefficient	CI	Spearman Coefficient	CI
No processing	Removed at 95%	0.724	0.688–0.757	0.988	0.986–0.990	0.755	0.722–0.785	0.988	0.986–0.990
No processing	Capped at 95%	0.909	0.895–0.920	0.999	0.999–0.999	0.910	0.896–0.921	0.999	0.999–0.999
Capped at 95%	Removed at 95%	0.878	0.861–0.894	0.988	0.986–0.990	0.904	0.890–0.916	0.988	0.986–0.990

Table 2: Stability of Within-Physician Results for the Parametric O : E Ratio and Wilcoxon Rank-Sum Algorithms between Two Time Periods

Outlier Processing Method	Parametric O : E Ratio (Raw)		WRS O : E Rank-Sum Ratio (Raw)		Parametric O : E Ratio (t-Test)		WRS Z-Score	
	Spearman Coefficient	CI	Spearman Coefficient	CI	Spearman Coefficient	CI	Spearman Coefficient	CI
No processing	0.620	0.573–0.662	0.896	0.881–0.909	0.682	0.641–0.719	0.906	0.891–0.918
Capped at 95%	0.676	0.635–0.714	0.896	0.881–0.910	0.727	0.690–0.759	0.906	0.891–0.918
Removed at 95%	0.634	0.588–0.675	0.890	0.873–0.904	0.689	0.649–0.725	0.897	0.882–0.910

Table 3: Stability of Physician Ratings for the Parametric O : E Ratio (*t*-Test) and Wilcoxon Rank-Sum Algorithms (Z-Score) between Two Time Periods Based on Ratings Proportions (729 Internal Medicine Physicians, Data Capped at the 95th Percentile Level)*

<i>Physician Ratings Set 1</i>	<i>Physician Ratings Set 2</i>	<i>Physician Count, WRS</i>	<i>% of Physicians, WRS</i>	<i>Physician Count, O : E Ratio</i>	<i>% of Physicians, O : E Ratio</i>
Low cost	Low cost	141	19.3	122	16.7
Low cost	Average cost	40	5.5	58	8.0
Low cost	High cost	1	0.1	2	0.3
Average cost	Low cost	41	5.6	57	7.8
Average cost	Average cost	287	39.4	233	32.0
Average cost	High cost	37	5.1	75	10.3
High cost	Low cost	0	0.0	3	0.4
High cost	Average cost	38	5.2	74	10.2
High cost	High cost	144	19.8	105	14.4
Total concordance		572	78.5	460	63.1
Totals†		729	100.0	729	100.0

*For time periods 1 and 2, 25 percent of physicians were rated “high cost,” 50 percent were rated “average cost,” and 25 percent were rated “low cost.”

†Due to rounding, percentages may not total exactly 100 percent.

time period 2) was 0.498 for the parametric O : E ratio with a *t*-test (“moderate agreement” according to the classification by Landis and Koch 1977) and 0.711 for the WRS *z*-score (substantial agreement). When ratings comparisons between the two time periods for a statistical algorithm were classified as concordant (the physician had the same categorical rating for both time periods) and discordant (the physician had different categorical ratings between time periods), the WRS *z*-score had greater ratings concordance than did the O : E *t*-test, and the difference was highly significant (for the WRS *z*-score, 78.5 percent of physicians were concordant and 21.5 percent were discordant; for the parametric O : E *t*-test, 63.1 percent of physicians were concordant and 36.9 percent were discordant; *z*-test for two independent proportions = 6.45, *p* < .0001).

Agreement between the Parametric O : E t-Test and WRS Z-score for Physician Ratings

Table 4 illustrates the differences in categorical adjusted ratings between the parametric O : E ratio *t*-test and WRS *z*-scores when each physician was rated using both methods. The 95th percentile capped dataset for time period 2 was used. The weighted kappa was 0.458 (moderate agreement) between the two

Table 4: Concordance of Physician Ratings between the Parametric O : E Ratio *t*-Test and Wilcoxon Rank-Sum Z-Score (Data Winsorized at the 95th Percentile Level, 729 Internal Medicine Physicians, Time Period 2)*

<i>Physician Ratings</i> WRS Z-score	<i>Physician Ratings</i> Parametric O : E Ratio <i>t</i> - test	<i>Physician</i> Count	<i>% of</i> <i>Physicians</i>
Low cost	Low cost	105	14.4
Low cost	Average cost	71	9.7
Low cost	High cost	6	0.8
Average cost	Low cost	66	9.1
Average cost	Average cost	234	32.1
Average cost	High cost	65	8.9
High cost	Low cost	11	1.5
High cost	Average cost	60	8.2
High cost	High cost	111	15.2
Total concordance		450	61.7
Totals [†]		729	100.0

*For time periods 1 and 2, 25 percent of physicians were rated "high cost," 50 percent were rated "average cost," and 25 percent were rated "low cost."

[†]Due to rounding, percentages may not total exactly 100 percent.

algorithms, with 61.7 percent overall concordance and 38.3 percent discordance.

DISCUSSION

We have described a method of compositing episode of care data by converting episode cost data to percentiles and applying a nonparametric test. The method we describe is consistent with the central limit theorem down to small sample sizes, is insensitive to common methodology choices required by parametric tests, and is more stable over time than usual parametric techniques.

Simulation testing showed that the parametric tests evaluated needed at least a sample size of 1,000 episodes per simulated physician to be consistent with normality, much larger than typical sample sizes available. In contrast, the WRS *z*-score distribution was consistent with normality down to a sample size of 10, the smallest sample size typically recommended for the normal approximation of the WRS. We believe there are two causes for the difficulty of achieving normality for typical parametric testing: first, that standard methodology combines data with different absolute sizes (e.g., \$100 episodes of pharyngitis and \$10,000 episodes of pneumonia), and second that the underly-

ing data are highly skewed and sensitive to outliers. The new method we describe in this paper addresses both issues.

As mentioned in the introduction, there are varying outlier processing strategies to adjust for the skew associated with health care cost data when performing statistical testing. These strategies require choosing where to trim or cap outliers. Accordingly, we tested our nonparametric algorithm against the parametric $O : E$ ratio for stability across different outlier mitigation strategies including no outlier processing. We chose to trim and cap outliers at the 95th percentile rather than the 97.5th percentile since the 95th percentile would create a greater change in the data and thus would be a better comparison group to illustrate statistical stability.

Our results show that such choices make a difference in physician cost rankings when common parametric techniques are applied. Although some may choose to customize, transform, or adjust the data to make parametric testing work more reasonably, applying the WRS test to percentiles removes the need for such further methodological choices. The technique described here showed no performance difference between using the raw data and applying an outlier methodology.

Data transformation into a distribution that more closely approximates normality, such as a logarithmic transformation, is another method of handling skew in the underlying data. However, log transformation can lead to difficulties when comparing distributions of dissimilar variances (Gianfrancesco et al. 2002), the results may be difficult to explain to physicians who do not think in terms of $\log(\text{cost})$, and it still may not produce a normal distribution especially with smaller sample sizes.

The WRS also performed significantly better than the parametric $O : E$ ratio statistics for stability over time. The Spearman rank-order correlation coefficients across time for the WRS $O : E$ ratio and for the WRS z -score was approximately 0.90, with the WRS z -score showing slightly higher Spearman coefficients than the WRS $O : E$ ratio. For the parametric $O : E$ t -test, the highest Spearman coefficient (0.727) came from the 95th percentile capped dataset, consistent with previous work that suggest Winsorization as the preferred outlier processing strategy (Thomas and Ward 2006).

A robust statistical test is one that continues to provide reasonably valid results even though some of its assumptions are violated, such as the assumption of normality for parametric tests (Sheskin 2000). Different physician cost evaluation systems may use different outlier processing strategies, and even the same system may change outlier strategies over time. This may cause changes in ratings even without changes in physician practice patterns. The

method we describe is significantly more stable across methodological strategies and over time. Our results show that this method is more robust in several dimensions than prior approaches and is therefore preferable to them. Using the WRS on cost percentiles eliminates the need to tinker with outlier processing.

Some key attributes of physician performance measures include appropriate case mix and severity adjustment, potential to aid care process and quality improvement, reflection of true variations in care, and transparency of methodology (AAFP 2007). By removing some of the biases and instability of parametric approaches, variations in physician cost patterns reflected in the above statistical approach are more likely to depict actual systematic differences in practice patterns, and thus increase the likelihood of generating viable action items for cost of care improvement. In our experience and in the reported experience of colleagues, we have found that a root cause of practice variation when common parametric tests are used is often due to a few high-cost outliers with low potential for actionability (Greene, Beckman, and Mahoney 2008) rather than true systematic variation in practice patterns.

The WRS method, which treats the impact of low-cost and high-cost conditions similarly, is intended primarily for statistical significance testing compared with peers in the setting of physician cost-efficiency ratings. Potential instability over time and the multiple outlier trimming methods used among different programs, as we have demonstrated, decreases trust in a cost assessment. The method of handling outliers makes a difference, as we have demonstrated. The incentive to change practice patterns based on a parametric analysis is less powerful if physicians do not trust the methodology. The policy implication, therefore, is that our method based on the WRS is preferable because it is more stable over time and independent of outlier methodology choices. These stabilities are important for physician engagement, optimizing network design, and consumer transparency. More robust assessment allows all stakeholders to concentrate on practice pattern improvement.

Limitations of the WRS test include the transformation of dollar cost values into a uniform distribution (percentiles), which may result in the loss of some information. Our WRS application gives the same weight to less costly episodes, such as pharyngitis, as more costly episodes, such as pneumonia. This limitation mainly would apply where the actual dollar magnitude of cost variation is critical such as in actuarial analysis. However, this limitation may be an advantage in several areas. Current claims-based risk adjustment methods are not perfect, and part of the variance in cost is still due to variables exogenous to the physician. This is most likely for higher cost or more

complex conditions, where patient factors may contribute more to variation than in simpler cases. Thus, the greater stability of the WRS algorithm discussed may be particularly helpful for physicians who treat high-cost or complex patients, since the statistic is less sensitive to highly variable exogenous factors. However, a dollar measure of cost variance (based on a parametric method) may still be helpful as a supplement for reporting purposes, to aid in complex patient management, or to help root cause analysis. Furthermore, because the dataset need not be trimmed, outlier data are preserved for additional uses in parallel to the improved physician assessment.

Another issue is that physician performance may change over time in response to practice improvement initiatives, which adds an unknown variable to the stability over time analysis. This issue is partially mitigated by the use of overlapping time periods and is controlled for by the experimental design, since both the parametric and nonparametric tests were subject to the same time periods. The purpose was to compare the statistical methods rather than measuring physician variability over time per se.

Other issues include the potential loss of information when creating composite metrics, the assignment of a single “responsible” physician for an episode when the episode contains others’ costs, and the potential limitations of particular episode grouping software, such as *ETG*, or the episode of care concept itself. These topics are beyond the scope of this paper. We have tried here to duplicate common industry practices so as to isolate and evaluate the utility and improvements of the new statistical method. Because health care data in general are highly skewed, we expect our results will generalize to the output of other episode groupers and to other cost data.

One of the assumptions of the WRS test is that each observation (episode of care) is independent of the other observations. For example, a member could appear more than once in the combined cost set since an individual member may have multiple episodes in different *ETGs* or may have more than one episode in a treatment set if they have recurrent episodes of an acute illness, such as pharyngitis (member-level clustering). Also, patients seeing a particular physician may have traits in common that are not detected by the case mix adjuster (physician-level clustering). There may also be interactions between physicians and treatment sets, with a physician performing systematically better for one treatment set or illness condition as compared with another. Since the purpose of the paper was to compare the WRS test to other tests commonly used in practice, using the same data for both tests, the experimental design controls for clustering and interactions in that regard, noting that the same issues apply to parametric tests. An actual dollar value may still

be useful here as an aid to determine which types of conditions (i.e., treatment sets) have the most variance. Dealing with clustering effects and interaction effects in the context of physician cost-efficiency statistical testing is an important area for further research.

In conclusion, physician cost of care evaluation using nonparametric testing based on the WRS offers a number of advantages over parametric methods currently in wide use. The Wilcoxon rank-sum test can be implemented so as to maintain the integrity of case mix and severity adjustment. The process we describe enables physician cost assessment in a robust fashion and addresses real-world issues relating to stability of the results in the face of skewed data.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors would like to thank Steve Wickstrom and Lucia Savage for their review and comments on the manuscript.

Disclosures: None.

REFERENCES

- AAFP (American Academy of Family Physicians). 2007. "Performance Measures Criteria" [accessed on August 24, 2011]. Available at http://www.aafp.org/online/en/home/policy/policies/p/performance_measures.html.
- Adams, J. L., E. A. McGlynn, J. W. Thomas, and A. Mehrotra. 2010a. "Incorporating Statistical Uncertainty in the Use of Physician Cost Profiles." *BMC Health Services Research* 10: 57–63.
- Adams, J. L., A. Mehrotra, J. W. Thomas, and E. A. McGlynn. 2010b. "Physician Cost Profiling—Reliability and Risk of Misclassification." *New England Journal of Medicine* 362 (11): 1014–21.
- Anderson, T. W., and D. A. Darling. 1952. "Asymptotic Theory of Certain 'Goodness-of-Fit' Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics* 23: 193–212.
- Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. Chichester, England: John Wiley and Sons.
- Barsky, A. J., E. J. Orav, and D. W. Bates. 2005. "Somatization Increases Medical Utilization and Costs Independent of Psychiatric and Medical Comorbidity." *Archives of General Psychiatry* 62 (8): 903–10.

- Gianfrancesco, F., R. H. Wang, R. Mahmoud, and R. White. 2002. "Methods for Claims-Based Pharmacoeconomic Studies in Psychosis." *Pharmacoeconomics* 20 (8): 499–511.
- Greene, R. A., H. B. Beckman, and T. Mahoney. 2008. "Beyond the Efficiency Index: Finding a Better Way to Reduce Overuse and Increase Efficiency in Physician Care." *Health Affairs (Millwood)* 27 (4): w250–9.
- Hart, A. 2001. "Mann-Whitney Test Is not Just a Test of Medians: Differences in Spread Can Be Important." *British Medical Journal* 323 (7309): 391–3.
- Kilian, R., H. Matschinger, W. Loeffler, C. Roick, and M. C. Angermeyer. 2002. "A Comparison of Methods to Handle Skew Distributed Cost Variables in the Analysis of the Resource Consumption in Schizophrenia Treatment." *Journal of Mental Health Policy and Economics* 5 (1): 21–31.
- Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.
- Lee, W. C., M. S. Hoffmann, S. Arcona, J. D'Souza, Q. Wang, and C. L. Pashos. 2005. "A Cost Comparison of Alternative Regimens for Treatment-Refractory Partial Seizure Disorder: An Econometric Analysis." *Clinical Therapeutics* 27 (10): 1629–38.
- Mino, Y., S. Shimodera, S. Inoue, H. Fujita, and K. Fukuzawa. 2007. "Medical Cost Analysis of Family Psychoeducation for Schizophrenia." *Psychiatry and Clinical Neurosciences* 61 (1): 20–4.
- NCQA (National Committee for Quality Assurance). 2008. "Standards and Guidelines for the Certification of Physician and Hospital Quality" [accessed on August 24, 2011]. Available at http://www.ncqa.org/Portals/0/PolicyUpdates/Corrections,%20Clarifications%20and%20Policy%20Changes/2008_PHQ_Front_12_5_08.pdf.
- Nixon, R. M., and S. G. Thompson. 2004. "Parametric Modelling of Cost Data in Medical Studies." *Statistics in Medicine* 23 (8): 1311–31.
- O'Brien, S. M., D. M. Shahian, E. R. DeLong, S. L. Normand, F. H. Edwards, V. A. Ferraris, C. K. Haan, J. B. Rich, C. M. Shewan, R. S. Dokholyan, R. P. Anderson, and E. D. Peterson. 2007. "Quality Measurement in Adult Cardiac Surgery: Part 2—Statistical Considerations in Composite Measure Scoring and Provider Rating." *Annals of Thoracic Surgery* 83 (4 suppl): S13–26.
- OptumInsight. 2011. "Transparency: Measuring and Improving the Quality and Efficiency of Health Care" [accessed on January 11, 2011]. Available at <http://www.ingenix.com/transparency/>.
- Sandy, L. G., M. C. Rattray, and J. W. Thomas. 2008. "Episode-Based Physician Profiling: A Guide to the Perplexing." *Journal of General Internal Medicine* 23 (9): 1521–4.
- Scholle, S. H., J. Roski, D. L. Dunn, J. L. Adams, D. P. Dugan, L. G. Pawlson, and E. A. Kerr. 2009. "Availability of Data for Measuring Physician Quality Performance." *American Journal of Managed Care* 15 (1): 67–72.
- Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52: 591–611.

- Sheskin, D. J. 2000. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Shine, D., S. Beg, J. Jaeger, D. Pencak, and R. Panush. 2001. "Association of Resident Coverage with Cost, Length of Stay, and Profitability at a Community Hospital." *Journal of General Internal Medicine* 16 (1): 1–8.
- Thomas, J. W., K. L. Grazier, and K. Ward. 2004. "Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures." *Health Services Research* 39 (4 Pt 1): 985–1003.
- Thomas, J. W., and K. Ward. 2006. "Economic Profiling of Physician Specialists: Use of Outlier Treatment and Episode Attribution Rules." *Inquiry* 43 (3): 271–82.
- Tina Shih, Y. C., Y. Xu, and L. S. Elting. 2007. "Costs of Uncontrolled Chemotherapy-Induced Nausea and Vomiting among Working-Age Cancer Patients Receiving Highly or Moderately Emetogenic Chemotherapy." *Cancer* 110 (3): 678–85.
- Visser, J. J., M. R. van Sambeek, M. G. Hunink, W. K. Redekop, L. C. van Dijk, J. M. Hendriks, and J. L. Bosch. 2006. "Acute Abdominal Aortic Aneurysms: Cost Analysis of Endovascular Repair and Open Surgery in Hemodynamically Stable Patients with 1-Year Follow-Up." *Radiology* 240 (3): 681–9.
- Wang, F., A. B. Schultz, S. Musich, T. McDonald, D. Hirschland, and D. W. Edington. 2003. "The Relationship between National Heart, Lung, and Blood Institute Weight Guidelines and Concurrent Medical Costs in a Manufacturing Population." *American Journal of Health Promotion* 17 (3): 183–9.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.