The Use of an Episode Grouper for Physician Profiling in Medicare: A Preliminary Investigation

Robert L. Houchens, Ph.D.
Scott McCracken, M.B.A.
William Marder, Ph.D.
Robert Kelley, M.S.

Thomson Healthcare
5425 Hollister Avenue
Suite 140
Santa Barbara, CA 93111-2348

•

MedPAC
601 New Jersey Avenue, NW
Suite 9000
Washington, DC 20001
(202) 220-3700
Fax: (202) 220-3759
www.medpac.gov

•

The views expressed in this report
are those of the authors.
No endorsement by MedPAC
is intended or should be inferred.

A study conducted by staff from Thomson Healthcare
for the Medicare Payment Advisory Commission

June 2009 • No. 09–3

The Use of an Episode Grouper For Physician Profiling in Medicare

A Preliminary Investigation

Final Report

Prepared for

Medicare Payment Advisory Commission

601 New Jersey Avenue, NW
Suite 9000
Washington, DC 20001

Prepared by

Robert L. Houchens, Ph.D.
Scott McCracken, M.B.A.
William Marder, Ph.D.
Robert Kelley, M.S.

Thomson Healthcare
5425 Hollister Avenue, Suite 140
Santa Barbara, CA 93111–2348

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

This study assessed the feasibility of using an episode grouper to identify physicians with substantially higher than expected utilization in the treatment of Medicare patients.

The episode grouper used in this study was Thomson's Medical Episode Grouper™ (MEG), which was commercially released in 1998. This tool combines medical claims found in administrative data into coherent and distinct episodes of treatment. These episodes describe a series of related health care services for the treating each patient's spells of illness. Episodes comprise health care utilization from multiple sites of service[1].

Each episode was characterized by several factors, including patient demographics, as well as the patient's disease, stage of disease, and complexity. Patient complexity was measured by the patient's Diagnostic Cost Group relative risk score, which is predictive of a patient's level of overall medical expenditures. The main episode outcome was the sum of standardized payments for services contained in the episode. Payments were standardized to eliminate area wage variations and other local cost factors. For example, standardized diagnosis related group (DRG) payments were used for hospitalizations and standardized payments based on relative value units (RVUs) were used for physician service payments.

The data employed in this study were provided by the Medicare Payment Advisory Commission (MedPAC). They comprised all 2002 and 2003 Medicare claims[2] for patients residing in six metropolitan statistical areas (MSAs): Boston, MA; Greenville, SC; Miami, FL; Minneapolis, MN; Orange County, CA; and Phoenix, AZ. In all, there were about 75 million claims per year, which were processed by MEG to produce episode groups. Approximately 85 percent of the claims could be grouped, representing over 96 percent of total claim payments. The final analysis file contained over 6 million Medicare episodes per year.

Each episode was attributed to a single physician based on that physician's share of the evaluation and management (E&M) payments for the episode. An episode was attributed to the physician who billed the highest percentage of total E & M payments, if it was at least 35 percent. Physicians were identified by the Unique Physician Identification Number (UPIN) provided on the claims data. Physician specialties were obtained from the Medicare physician file. In total, episodes were attributed to about 37,000 physicians per year. For our analyses we selected only physicians assigned at least 20 episodes, for a total of about 25,000 physicians per year.

Our objective was to identify physicians whose average episode payment could be considered an "outlier." For this study, we defined an outlier to be an average episode payment that exceeded the expected mean payment by at least 25 percent at the .0001 level of statistical significance. The expected mean payment was adjusted for episode and patient severity and it corresponded to the mean for an "average" physician with the same specialty and in the same MSA as the physician under test. The choice of threshold, 25 percent above the expected mean, represented a substantial deviation from the expected mean, although it was somewhat arbitrary. The very low significance level, .0001, was selected because of the large number of tests that were conducted (one for each physician). Compared with more conventional significance levels, it reduced the

[1] Prescription drug claims can also be incorporated. However, they were not available for the present study.

[2] 2001 claims were also used to ensure that episodes beginning in 2001 and ending in 2002 would be completed.

overall chances that one of the many physicians under test would be erroneously labeled an outlier.

We employed two statistical methodologies to determine outlier status. First, we employed multilevel regression models, which accounted for the correlation of episodes within physicians while estimating the physician's residual (deviation of the physician's observed episode mean payment from his or her expected episode mean payment). Multilevel models have been widely used for provider profiling applications. Second, we employed randomization tests. Unlike multilevel models, these tests require no assumptions concerning statistical distributions. For the randomization tests, each physician's mean payment was compared to a distribution of mean payments estimated from a large number of random samples of episodes similar to those attributed to the physician under test.

Both methods yielded stable estimates of physician residuals. Among physicians with at least 20 episodes in both years, the correlation between the physician's 2002 and 2003 residuals was 89 percent for the multilevel model and 87 percent for the randomization test[3]. The multilevel models identified a slightly higher percentage of physician outliers compared with the randomization tests (4.4 % vs. 2.9 % in 2002, and 4.7 % vs. 3.4 % in 2003).

Both methods produced "stable" outliers in the sense that outliers in 2002 tended to have small *p*-values (large residuals) in 2003 and, likewise, outliers in 2003 tended to have small *p*-values (large residuals) in 2002. The following table quantifies this stability:

Method	**Year**	**# Outlier MDs**	**% of Outliers Adjacent Year *p*-value < .05**
Multilevel model	2002	918	90.7
	2003	972	88.6
Randomization test	2002	611	93.6
	2003	712	90.0

The multilevel method produced 918 outliers in 2002, of which 90.7 percent had a small *p*-value in 2003. In other words, over 90 percent of the 2002 outliers also showed evidence of being an outlier in 2003. We call this the "look forward" from 2002 to 2003. Similarly, we did a "look backward" from 2003 to 2002. The .05 significance level defining the "small" *p*-value for the adjacent year is justified on the grounds that we were only testing physicians for whom we already had strong evidence of being an outlier. Approximately 10 percent of the outlier physicians in one year had *p*-values larger than .05 in the adjacent year, indicating that their residual in the adjacent year was not significantly 25 percent above the expected residual. These physicians could have been truly outliers in one year and truly not outliers in the adjacent year, or they could have been erroneously identified as an outlier. In any event, the overall results are encouraging that a physician's outlier status appears to be highly persistent from year to year.

While we believe that this study establishes the feasibility of episode groupers for use in Medicare physician profiling, we note the following limitations:

[3] Residuals for the randomization test were based on the difference between the physician's observed mean episode payment and the average of the mean payment distribution generated from random samples of similar episodes.

1. Standardized payments were the basis for measuring episode resource intensity and physician "efficiency." For example, hospital payments were the same for every patient hospitalized with a given diagnosis related group. This standardization no doubt masked some true episode cost variation.
2. Each episode was attributed to the single physician that billed the highest percentage of E&M dollars (at least 35 %) for that episode. For episodes involving multiple physicians, it is possible that less than full responsibility should have been accorded to that physician.
3. Risk adjustment was based on episode severity as measured by the episode's principal disease, the stage of the principal disease, and the relative risk score. Although these factors incorporated patient diagnoses and demographics, other factors might have provided further risk adjustment.
4. Physician comparisons were based only on episodes attributed to physicians within the same specialty group and within the same MSA. There might be an argument for comparing performance across a broader spectrum of specialties and geographic areas.
5. These analyses were strictly episode-based. They only compared physicians on their average episode-level resource intensity. They did not account for the frequency of episodes. It is possible that some physicians broke up the treatment for a condition into several low-intensity episodes, while other physicians combined the treatment for a condition into a few high-intensity episodes. However, the several low-intensity episodes would need to have been widely spaced to create separate episodes using the MEG algorithms.
6. The episodes in this analysis were based on the MSA of the patient, not on the MSA of the physician. For example, all episodes for Boston physicians were based solely on patients residing in the Boston MSA. However, this excluded episodes for patients outside the Boston MSA that were treated by Boston physicians.

To partially address the second point, MedPAC has commissioned a study currently under way to test multiple physician attribution in place of single physician attribution. We also recommend that MedPAC should repeat the analyses in the present study to address the sixth limitation. If some physicians treated a large number of patients outside their own MSA, then their estimated mean episode payment could have been biased if patients outside their MSA had different treatment patterns compared with patients in their own MSA. At a minimum, the larger sample of episodes could produce a more reliable estimate of their mean episode payments.

.

INTRODUCTION

The purpose of this study was to assess the usefulness of episode groupers for profiling physicians on their treatment of Medicare patients. In particular, we developed measures related to physician efficiency. We say "related to" because we could only approximately estimate efficiency with the available data.

The episode grouper used in this study was Thomson's Medical Episode Grouper™ (MEG), which was commercially released in 1998. This tool combines medical claims found in administrative data into coherent and distinct episodes of treatment. These episodes describe a series of related health care services for the treating each patient's spells of illness. Episodes can be comprised of outpatient, inpatient, skilled nursing facility, and home health agency utilization[4]. The grouper is described in more detail in the Appendix.

The data employed in this study were provided by the Medicare Payment Advisory Commission (MedPAC). They comprised all 2002 and 2003 Medicare claims for patients residing in five metropolitan statistical areas (MSAs): Boston, MA; Greenville, SC; Miami, FL; Minneapolis, MN; Orange County, CA; and Phoenix, AZ. The data are described in the Data section of this report. The detailed results of applying MEG to the data are contained in the Appendix.

A primary objective was to identify physicians whose average episode payment could be considered an "outlier." We employed two statistical methodologies to determine outlier status. These methods and the process for identifying outliers based on them are explained in the Methods section of this report.

First, we employed multilevel regression models, which accounted for the correlation of episodes within physicians while estimating the physician's residual (deviation of the physician's observed episode mean payment from his or her expected episode mean payment). Multilevel models have been widely used for provider profiling applications (Dubois et al., 1987; Jencks et al., 1988; Thomas et al., 1994; Normand et al., 1995; Epstein, 1995; Schneider and Epstein, 1996; Morris and Christiansen, 1996; Goldstein and Spiegelhalter, 1996; Rice and Leyland, 1996; Normand et al., 1997; Leyland and Boddy, 1998; Marshall and Spiegelhalter, 2001).

Second, we employed randomization tests (Manly, 2007; Noreen, 1989). Unlike multilevel models, these tests require no assumptions concerning statistical distributions. For the randomization tests, each physician's mean payment was compared to a distribution of mean payments estimated from a large number of random samples of episodes similar to those attributed to the physician under test. We do not know of any previous applications of this methodology to provider profiling.

The Results section contains results for the two methods separately as well as comparisons between the methods. This section shows the distribution of the outliers and the underlying efficiency measures overall and for selected conditions. It also addresses the stability—the year-to-year persistence—of both the outliers and the efficiency measures.

The final section of this report, Conclusions and Recommendations, contains a broad assessment of the results, some important caveats to the study, and some considerations for future studies.

[4] Prescription drug claims can also be incorporated. However, they were not available for the present study.

DATA

MedPAC provided the study data, composed of all medical claims during the calendar years 2001 through 2004 for Medicare beneficiaries residing in the six study MSAs: Boston, Greenville, Miami, Minneapolis, Orange County and Phoenix. Table 1 displays the number of claims, by year, for each claim source.

Table 1: Number of Medicare Claims, by Source and Year.

Claim Source	Year				Total
	2001	2002	2003	2004	
HHA	147,523	159,901	178,903	197,674	**684,001**
MEDPAR	575,519	591,412	618,358	633,141	**2,418,430**
Physician	47,342,026	51,054,090	55,980,215	57,916,665	**212,292,996**
Outpatient	14,961,933	16,035,609	16,855,777	18,030,835	**65,884,154**
Total	**63,027,001**	**67,841,012**	**73,633,253**	**76,778,315**	**281,279,581**

Key Variables in the Raw Data

The following data elements, which are necessary for episode creation, were extracted from the raw data files and placed in a uniform format:

- Patient ID – a unique and encrypted patient identifier.
- UPIN – a unique physician identification number.
- Diagnosis Codes – the reconfigured claims records contained up to 11 diagnosis codes assigned using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis coding system.
- Procedure Codes – Current Procedural Terminology (CPT) procedure codes, Healthcare Common Procedure Coding System (HCPCS) procedure codes, and ICD-9-CM procedure codes were extracted from the original data. Each claim record contained one procedure code.
- MSA – the patient's metropolitan statistical area.
- Standardized Payment – as described below, claims payment amounts were standardized to remove local market payment differences among episodes.
- Age – patient age, in years.
- Gender – patient gender.
- Date of Service – the date of outpatient service or the date of admission.
- Claim Number – a unique record identification number.
- Length of Stay- inpatient length of stay.

Standardized Payments

MedPAC established methods for standardizing payments for physician profiling applications with episode groupers (Medicare Payment Advisory Commission, 2006). Briefly, for each type of claim, MedPAC standardized payments as follows:

- Hospital inpatient services — A standardized amount was created for each Diagnosis Related Group (DRG) for each year and applied to all records uniformly.
- Skilled Nursing Facility (SNF) services— SNF Medicare Provider Analysis and Review records were merged to the DataPro SNF Stay file. This information was combined with specific standardized amounts of resource utilization groups from CMS.
- Long-term care hospital services —For discharges that occurred on or after October 1, 2002, a standardized amount for each DRG was applied. For discharges prior to this date, local area wage-index adjustments from each hospital's payment were backed-out, assuming local area wage indexes acted as a proxy for underlying costs.
- Rehabilitation/psychiatric hospital services —Total Medicare payments and total length of stay were calculated for each DRG, a DRG-level per diem amount was created and then multiplied by the length of stay for each record.
- Home health — the home health case-mix weight on each claim was multiplied times the base payment rate for the appropriate fiscal year.
- Physician services — the relative value unit (RVU) was determined for each record by matching the HCPCS code and modifier on the record to the physician fee schedule RVU file. The RVU was multiplied by the units of volume for each record by the conversion factor for the appropriate year and reduced the standardized payment for multiple surgical procedures on the same claim and for services provided by physician assistants and assistants at surgery.
- Ambulatory Surgical Center (ASC) services — HCPCS codes were used to match records to ASC payment rate files. Consistent with Medicare payment rules the payment rate was reduced for multiple surgical procedures on the same claim.
- Clinical laboratory services — A record was classified as a clinical lab service if the HCPCS for a record on the carrier file matched a HCPCS code on the clinical lab fee schedule. The standardized payment rate for each lab record is the national limitation amount (NLA) for the service.
- Anesthesia services —The base and the time units were summed for each anesthesia record and multiplied by the anesthesia conversion factor for the appropriate year. Certified registered nurse anesthetists were assigned an amount that was half of the full amount, consistent with Medicare payment rules.
- Hospital outpatient services — HCPCS codes were used to match outpatient records to an outpatient prospective payment system payment rate file and a standardized payment amount was assigned to each record.

In this study, the total payment for an episode is the total of the standardized payments for the claims contained in that episode. Throughout this report the term "payment" is shorthand for "standardized payment."

Berenson-Eggers Type of Service (BETOS)[5]

The BETOS coding system was developed primarily for analyzing the growth in Medicare expenditure. The coding system assigns each and every HCPCS codes to a single BETOS code, which represents a readily understood clinical category. BETOS codes were added to professional and outpatient claims.

BETOS codes are broadly classified under seven major categories:

1. Evaluation and Management

[5] See www.cms.hhs.gov/HCPCSReleaseCodeSets/20_BETOS.asp (last accessed 9/9/2007) for more information on BETOS categories.

2. Procedures
3. Imaging
4. Tests
5. Durable Medical Equipment
6. Other
7. Exceptions/Unclassified

The category of Evaluation and Management (E&M) played a special role in the assignment of episodes to physicians, as explained in the Appendix. We also used these as descriptive payment categories to “drill down” on total episode payments to better understand outpatient utilization patterns.

METHODS

The Appendix contains a description the Medical Episode Grouper (MEG^{TM}), which was used to produce episodes for our analyses. It also explains the method we used to attribute episodes to physicians.

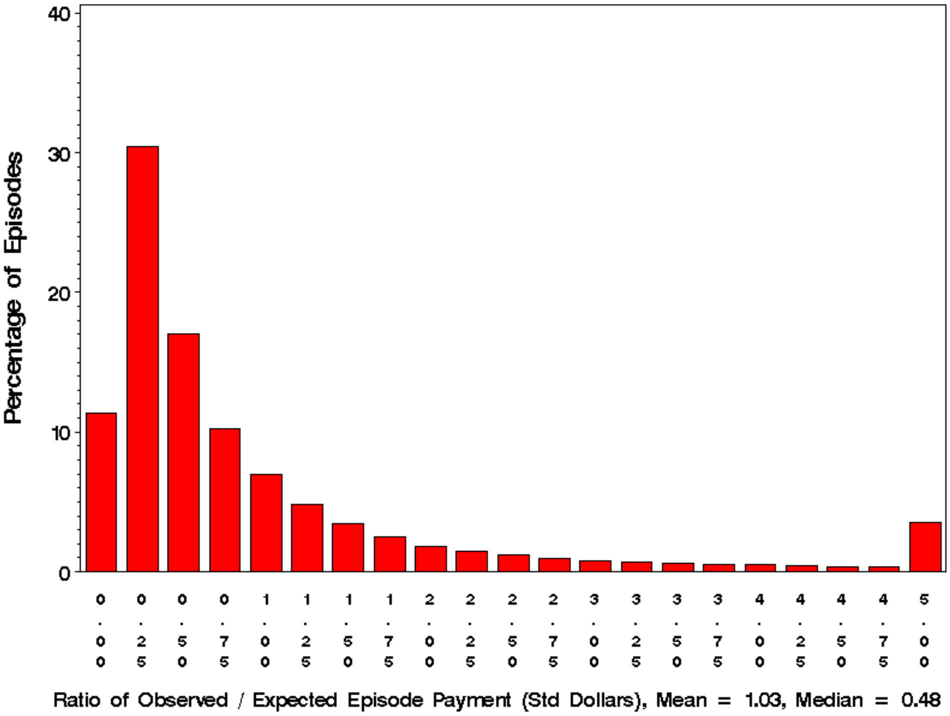
Below, we discuss the statistical methods that were employed to identify physician outliers. We used two approaches. First, we fit a multilevel model, which is a regression model suitable for nested data. Second, we used an approximate randomization test, which is a more transparent method that makes fewer statistical assumptions than the multilevel model. In both cases, we tested whether each physician is an outlier in terms of mean episode payments. The *p*-value for these tests was set at a very small value to account for the large number of tests performed.

Multilevel Models

Multilevel models are often used and recommended for physician and hospital profiling applications (Dubois et al., 1987; Jencks et al., 1988; Thomas et al., 1994; Normand et al., 1995; Epstein, 1995; Schneider and Epstein, 1996; Morris and Christiansen, 1996; Goldstein and Spiegelhalter, 1996; Rice and Leyland, 1996; Normand et al., 1997; Leyland and Boddy, 1998; Marshall and Spiegelhalter, 2001). These regression models—also called hierarchical models or mixed effects models—are designed for nested or grouped data such as we have with episodes nested within physicians. Specifically, these models take into account the correlation of episodes within physicians, unlike standard regression methods that assume the observations are uncorrelated.

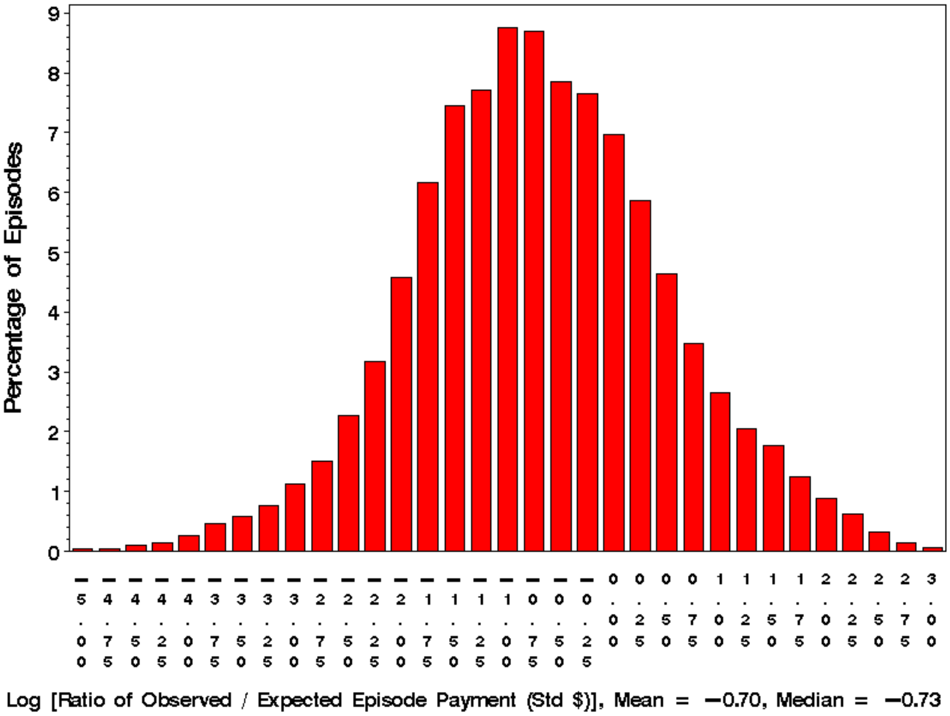
Throughout this report, the term “payment” is shorthand for “standardized payment.” We analyzed the payment ratio = (observed payment) / (expected payment) calculated for each episode. This ratio is highly skewed (Figure 1), with a long right tail. Consequently, we modeled the logarithm of the payment ratio, which has a more nearly normal distribution (Figure 2), helping to satisfy one assumption for the multilevel regression models we fit.

Figure 1: Distribution of Payment Ratio per Episode.



Data source: All episodes for Medicare patients from six MSAs during 2002.

Figure 2: Distribution of Log(Payment Ratio) per Episode.



Data source: All episodes for Medicare patients from six MSAs during 2002.

We begin by considering a simple multilevel model. Define

O_{ij} = observed payment for episode i attributed to physician j.
E_{ij} = expected payment for episode i attributed to physician j.

Consider the regression model:

$$\begin{aligned}
&\ln\left(\frac{O_{ij}}{E_{ij}}\right) = \beta_{0j} + e_{ij} \\
&\beta_{0j} = \beta_0 + u_j \\
&u_j \sim N\left(0, \sigma_u^2\right) \\
&e_{ij} \sim N\left(0, \sigma_e^2\right)
\end{aligned} \qquad \text{(Model 1)}$$

Model 1 assumes that the logarithm of the payment ratio is distributed as normal with a specific mean for physician j, denoted by β_{0j}. The physician means are assumed to be distributed as normal with an overall mean β_0, which is the mean for an “average” physician.

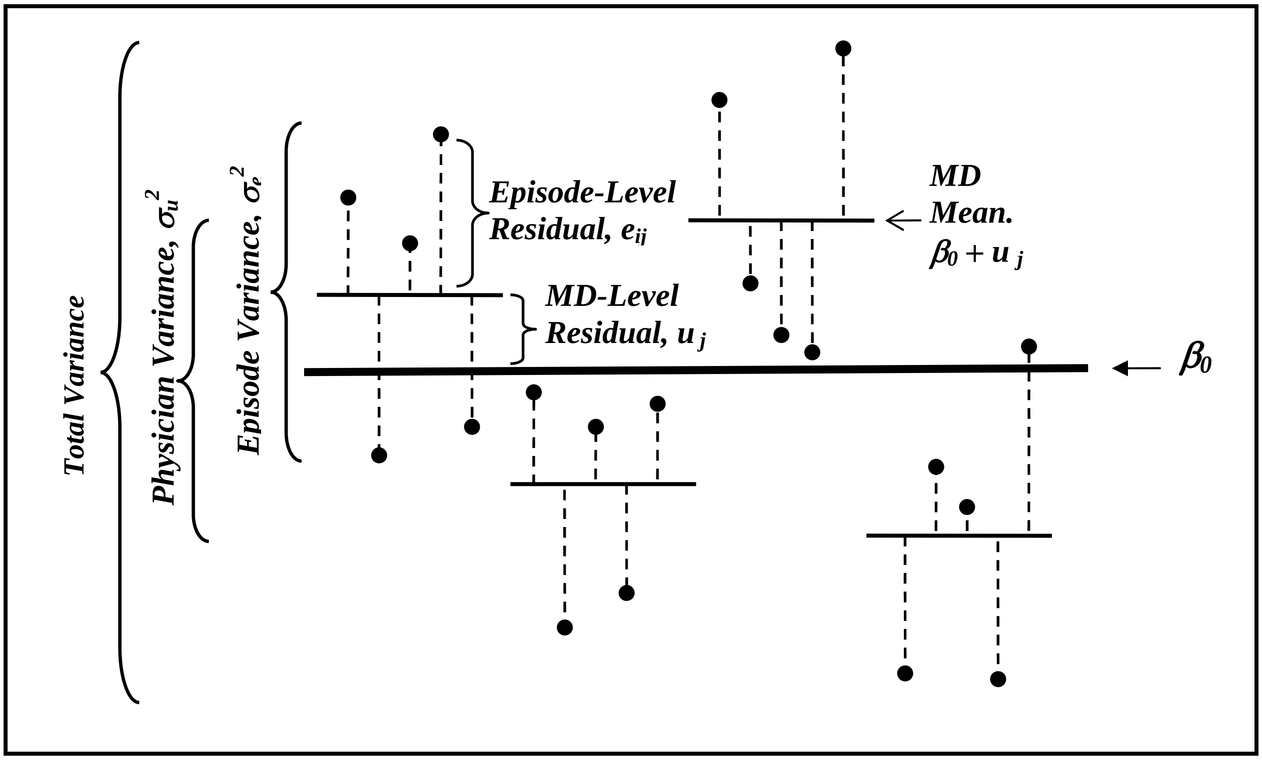
We are interested in the value of u_j, the residual deviation of physician j from the overall average. If this residual is positive (negative), then the average payment per episode for physician j tends to be higher (lower) than that of the average physician. This residual forms the basis for each physician’s estimated “efficiency” score. The episode-level residual for episode i treated by physician j is denoted by e_{ij}. In this simple version of the model, this episode-level residual is assumed to be normally distributed with a mean of zero and a constant variance, σ_e^2.

In Figure 3, this model is illustrated graphically for four physicians. The thin, short horizontal lines represent episode means for each of four physicians. Each physician-level residual is equal to the difference between the physician’s mean and the overall mean. The overall mean is represented by the thick horizontal line through the center of the graph, labeled β_0. Each episode’s residual is equal to the difference between the episode’s observed response, represented by a dot, and the physician’s mean. The total residual is the sum of the episode-level residual and the physician-level residual. The physician variance is represented by the spread of the physician-level means around the overall mean. The episode variance is represented by the spread of the episode-level responses around the physician means. In our regression, the response is equal to ln(Observed Payment).

This model description above is sufficient for understanding the analyses reported below. However, we actually fit a slight modification of the model designed to account for variance heterogeneity, as explained in the Appendix.

SAS PROC MIXED was used to fit the multilevel models, and to estimate the value of each physician residual, u_j, and its standard error. Subsequently, these estimates were used to identify outlier physicians—those with especially large positive residual values.

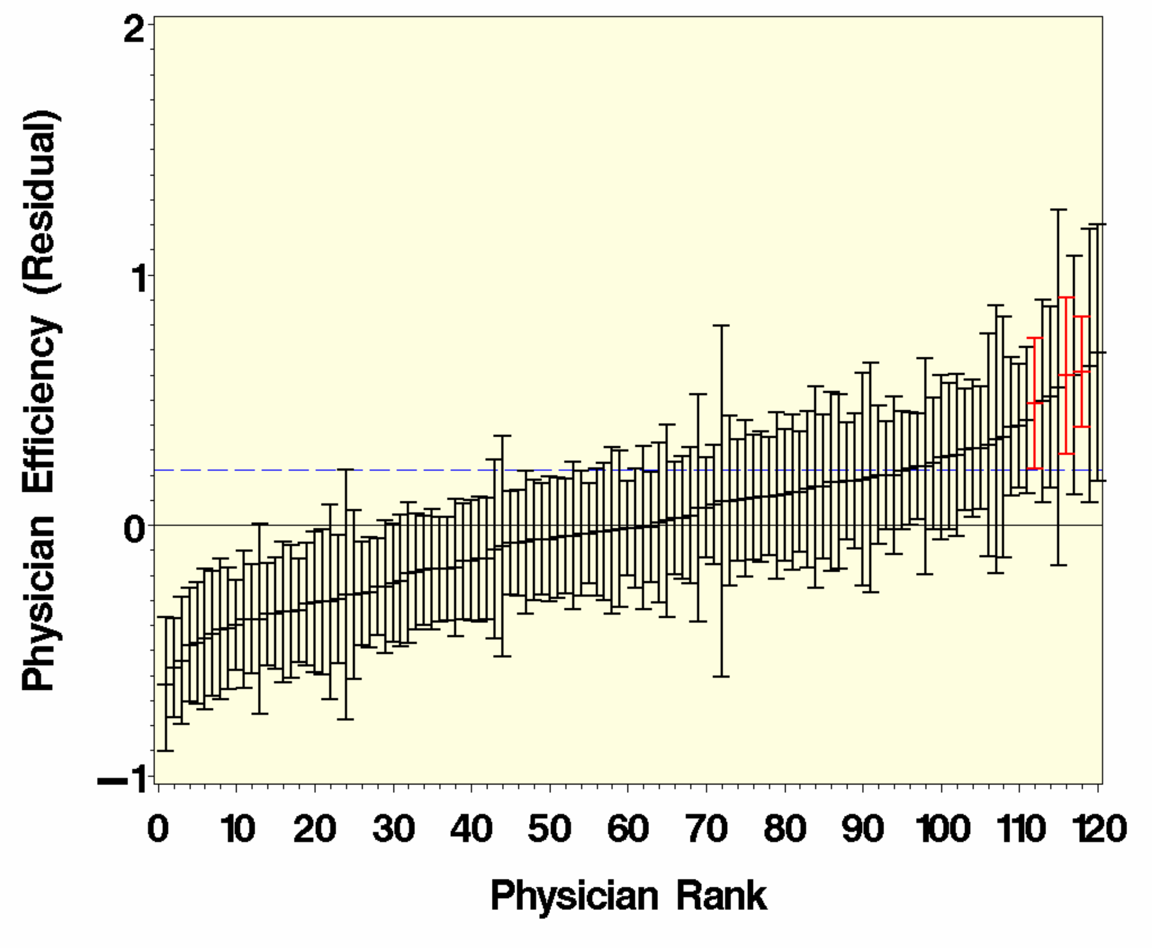
Figure 3: Illustration of Episode-level Residuals and Physician-level Residuals.



We tested whether each physician was significantly above the average by at least 25 percent. In so doing, we assumed that the physician residual variance was fairly small and that we were mainly identifying residuals that lie outside a narrow range (Ohlssen, et al., 2006). Since we performed many hypothesis tests, we declared significance only when $p_j < 0.0001$, where p_j is a one-sided *p*-value for the null hypothesis that the residual for physician j is equal to zero versus the alternative hypothesis that the residual for physician j is greater than the mean by at least 25 percent. The threshold of 0.0001 was selected to reduce the probability of identifying a false outlier among the large number of physicians being tested.

An example residual plot is shown in Figure 4, which shows the estimated residuals and 99.98 % confidence limits for 120 physicians, ranked from smallest residual (most efficient) to largest residual (least efficient). Each red confidence bar is completely above the dashed blue line and corresponds to an “outlier” physician whose residual is 25 percent higher than that of the average physician at the .0001 significance level.

Figure 4: Example Physician Residual Plot.



Approximate Randomization Tests

The multilevel model makes critical assumptions concerning statistical distributions and the form of the model. Using that approach, physician outliers are identified based on the physician-level residuals estimated from the model. In contrast, approximate randomization tests are non-parametric, making very few assumptions about the data (Manly, 2007; Noreen, 1989). The idea is to test whether the observed average episode payment for each physician's sample is consistent with the complete distribution of average episode payments *for similar samples* drawn at random from the collection of all physicians' episodes. Using this approach, physician outliers are identified based on how unlikely the physician's observed average episode payment is compared with the distribution of average episode payments for similar samples of randomly-drawn episodes.

Perhaps the simplest way to understand the approximate randomization approach is through an example. In Table 2, the fifth column labeled "MD Sample" contains the observed payment for 22 episodes attributed to an example physician, and we want to test whether the average episode payment of $1,521 makes this physician an outlier. Scanning down that column, the observed payments were $114 for the physician's first episode, $334 for the physician's second episode, and so on. The average observed payment for all 22 episodes was $1,521, shown at the bottom of column five. For each episode, the second column ("MEG"), third column ("Stage"), and fourth column ("RRS Group") indicate the episode's group number, stage of disease, and relative risk score group, respectively. For example, the first five episodes have MEG = 180, Stage = 1, and RRS Group = 1.

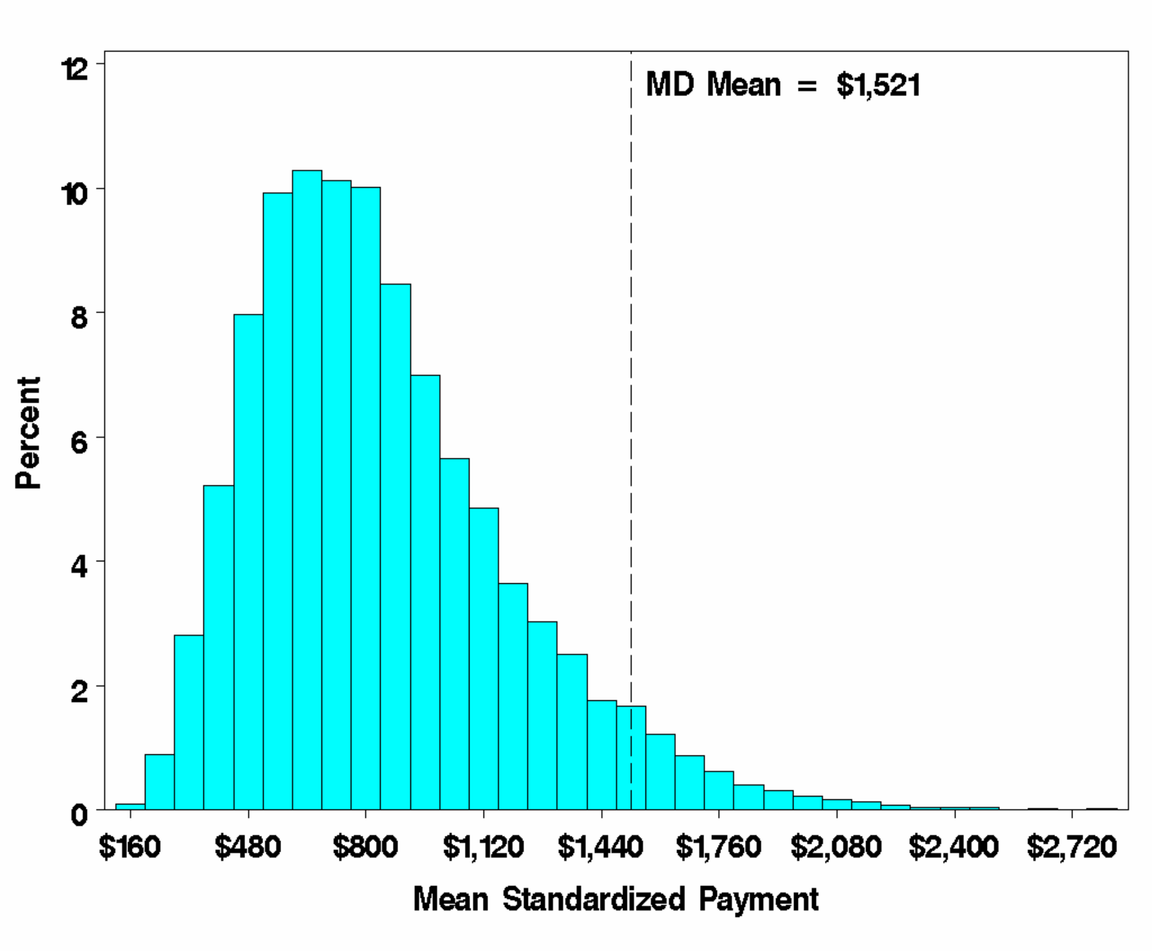
The columns labeled "Sample *m*" for *m* = 1, 2,…, 5, contain observed payments for randomly drawn episodes from within each category of MEG, Stage, and RRS. The mean payment for each sample is shown at the bottom of the table. For example, the mean payment was $1,343 for sample 1, and it was $801 for sample 4. The mean payments for the five samples ranged from a low of $579 (sample 5) to a high of $1,375 (sample 2). These mean payments are for samples containing the same number of episodes and with the same case-mix as the subject physician's sample of episodes because the episodes in each of the five samples are matched on MEG, Stage, and RRS. Based on these five sample means, the physician's observed mean payment of $1,521 appears high. Of course, five is too small a sample on which to judge whether the physician's mean payment is really an "outlier." Therefore, we drew 10,000 random samples of episodes and compared the physician's mean payment to the distribution of 10,000 sample mean payments.

Table 2: Five Monte Carlo Samples Matched to an MD Sample of 22 Episodes

Episode #	MEG	Stage	RRS Group	MD Sample	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
1	180	1	1	$ 114	$ 812	$ 55	$ 301	$ 655	$ 197
2	180	1	1	334	1,003	1,221	66	51	84
3	180	1	1	96	4,256	80	3,140	41	55
4	180	1	1	151	2,135	44	704	51	1,544
5	180	1	1	55	4,521	55	52	139	99
6	181	1	1	141	600	120	235	87	55
7	184	1	1	3,475	851	528	5,499	2,822	106
8	184	1	1	623	3,562	15,141	3,363	1,123	681
9	184	1	1	5,680	154	154	8,605	7,218	4,119
10	192	1	1	625	168	58	51	336	2,129
11	192	1	1	527	2,037	81	876	51	1,073
12	192	1	1	1,188	577	168	454	279	183
13	193	1	1	110	84	51	89	55	73
14	331	1	1	96	55	416	263	641	799
15	331	1	1	78	236	177	387	125	192
16	331	1	1	171	111	66	243	456	70
17	331	1	3	264	4,664	157	83	1,538	158
18	331	1	3	68	92	80	171	69	225
19	331	2	3	3,995	3,030	5,690	4,060	584	390
20	336	1	1	625	218	41	169	172	195
21	336	2	1	14,900	339	5,384	51	212	195
22	339	1	1	151	51	487	139	907	120
			Mean	**$ 1,521**	**$ 1,343**	**$ 1,375**	**$ 1,318**	**$ 801**	**$ 579**

Figure 5 shows the distribution of 10,000 sample mean payments for the example physician. About 5 percent of the 10,000 sample means exceeded this physician's observed mean payment. This percentage can be considered a *p*-value, the probability of observing a mean payment as large or larger from a random sample of payments for similar episodes[6]. A *p*-value of 0.05 is too large to reject the null hypothesis. Consequently, the physician's mean is not significantly different from the average.

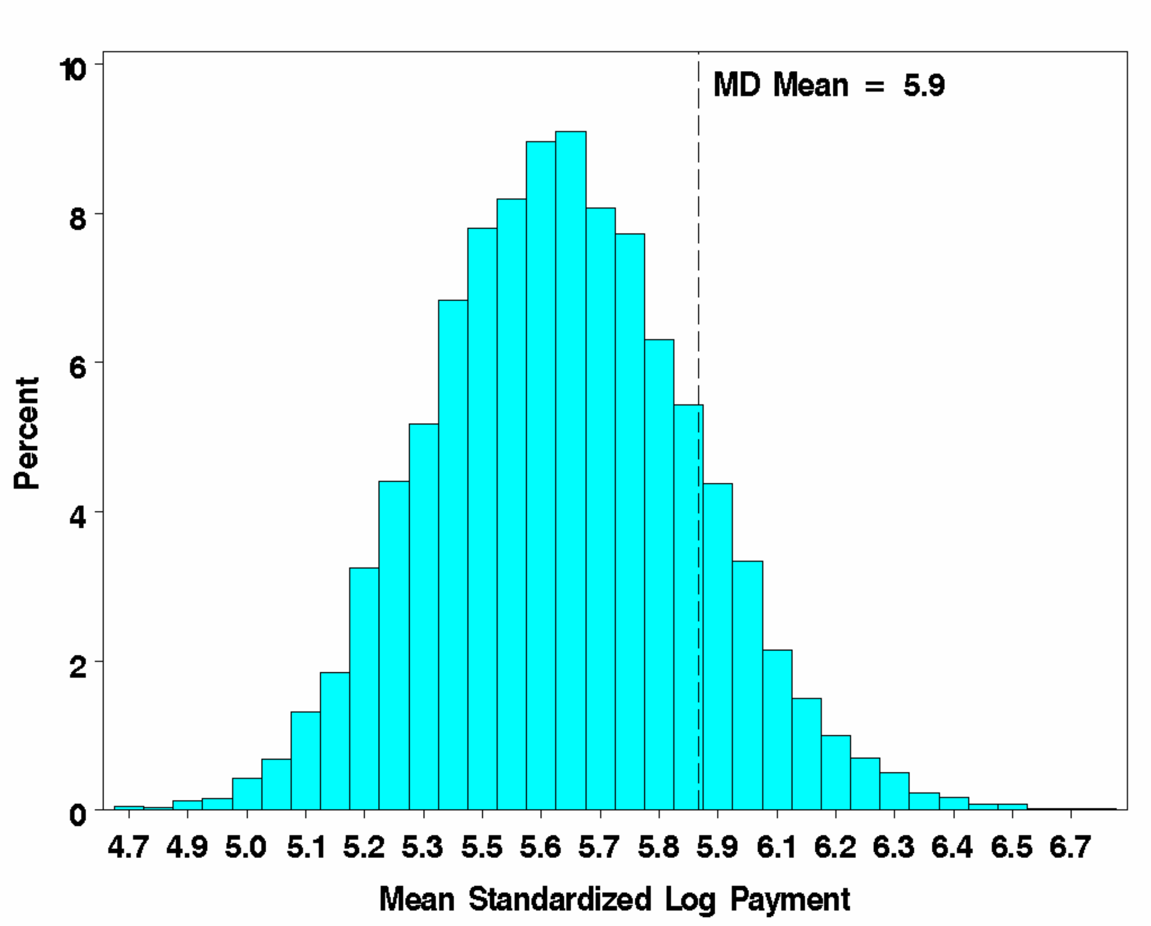
Figure 5: Distribution of Mean Payments from 10,000 Monte Carlo Samples for an MD.



We actually conducted the randomization tests based on log(payments) rather than payments for two reasons. First, the log transformation reduces the influence of episode-level payment outliers. Second, this approach is consistent with the multilevel models, which employed log(payment) as the dependent variable. The log(payment) distribution is shown in Figure 5 corresponding to the payment distribution shown in Figure 5. The physician's observed log(payment) is 5.9 and the corresponding p-value is 0.14, indicating that this physician's mean log(payment) is not significantly different from the average log(payment).

[6] Technically, the *p*-value is calculated as (g+1) / 10,001, where g is the number of sample means with a value greater than the physician's observed mean.

Figure 6: Distribution of Mean Log(Payments) from 10,000 Monte Carlo Samples for an MD.



Finally, rather than test whether each physician's mean was above average, we tested whether each physician's mean exceeded the expected mean by at least 25 percent. To accomplish this, we added log(1.25) to each of the 10,000 means on the log scale. For example, in Figure 6 the entire distribution is shifted to the right by 0.223 (=log(1.25)), while the physician's average stays at 5.9. Using approximate randomization tests to identify outliers entails multiple comparisons. Therefore, just as for the multilevel modeling approach, we identified as outliers those physicians with *p*-values under .0001.

While we conducted the randomization tests on the logarithm of payments to identify outlier physicians, we recommend plotting the distribution of means on the dollar scale, as shown in Figure 5, for descriptive purposes. There are advantages to describing the outlier physician's payments on the original dollar scale. In particular, it allows for an easy "drill down" on total payments by disaggregating episode payments into several service categories (e.g., inpatient payments, office payments, imaging payments, etc.). The 10,000 sample average payments can be calculated by type of service and Monte Carlo distributions can be constructed for each payment category. The means for these payment categories will sum to the mean for total payments. An example will be shown later in this report.

For each physician, SAS PROC SURVEYSELECT was used to obtain 10,000 random samples stratified on MEG (disease), Stage of Disease, and Relative Risk Group, taken from the "population" of episodes in that physician's specialty and MSA. Therefore, each physician's "peer group" was considered to be physicians in the same specialty and in the same MSA. For example, for an endocrinologist in Boston, a random sample of diabetes episodes would be taken from diabetes episodes attributed to Boston endocrinologists, but not from diabetes episodes attributed to internists, general practitioners, and other specialties, and not from endocrinologists in MSAs other than Boston. This is consistent with the multilevel model approach, wherein each regression was limited to episodes in a given MSA for physicians in a given specialty.

Identifying Physician Outliers

We are interested in identifying *outlier physicians*, and not merely physicians with mean payments above that of the average physician, which in reality is probably about half of all physicians. Therefore, we tested whether each physician's mean payment was at least 25 percent *above the average* at the .0001 significance level. For an analysis involving N physicians, the overall type I error rate (probability of identifying at least one false outlier) is approximately $1 - .9999^N$. Table 3 shows the overall *p*-values for representative values of N:

Table 3: Overall *p*-values for a One-sided *p*-value = .0001 for Each Test.

Number of Physicians, N	Overall *p*-value
25	.0025
50	.0050
100	.0100
250	.0247
500	.0488

For example, in an analysis involving 100 physicians, if we identify outliers as physicians with values of $p \leq .0001$, then there is a 1 percent chance that at least one physician will be classified as an outlier who is not truly an outlier. Likewise, in an analysis involving 500 physicians, there is a nearly 5 percent chance that at least one physician will be declared an outlier by mistake. On the other hand, setting such a low *p*-value increases the risk of failing to identify true outliers.

Year-to-year Stability

To measure the "stability" of the efficiency measures, we analyze physicians' *p*-values between 2002 and 2003. In the absence of efforts to change behavior, we expect most outlier physicians to remain outlier physicians during adjacent years. By changing practice patterns, it is certainly possible for a physician to be a true outlier during one year and not an outlier during either the previous year or the following year. It is also possible for a physician to be "unlucky" in the sense that his or her episodes tend to come from the high end of the "normal" payment distribution for one year, incorrectly causing him or her to be declared an outlier in that year. In that case, it is unlikely that the same physician would be "unlucky" twice, because the physician's average episode payment would tend to "regress" to the mean in another year, yielding an unremarkable *p*-value. Still another reason for inconsistent results between years could be a larger sample of patients in one year than in another year. A larger sample can turn a statistically insignificant difference into a statistically significant one.

To test stability, we identify physician outliers in 2002 and then look "forward" at their *p*-values in 2003. If the 2002 outlier physicians are "true" outliers, then most of them also should have low *p*-values in 2003. Likewise, we identify physician outliers in 2003 and then look "backward" at their *p*-values in 2002. Again, if the 2003 outlier physicians are “true” outliers, then most of them should also have low *p*-values in 2002. For this purpose, we regard a *p*-value as “low” if it is less than .05. This threshold is somewhat arbitrary. However, it is a conventional threshold for statistical testing, and we are only considering physicians for whom we have reason to believe that their mean will substantially deviate from the overall mean because they had extremely low *p*-values (< .0001) in the adjacent year.

RESULTS

The Appendix contains descriptive results relating to the application of MEG to the entire database of Medicare claims provided by MedPAC. In what follows, we illustrate the outlier identification methodologies by presenting results in total (all physicians), and selected results for physicians in two specialties, urology and cardiology, for all six MSAs: Boston, Greenville, Miami, Minneapolis, Orange County, and Phoenix. As discussed in the methods section, each physician is tested for having a high mean episode payment (one-sided test) relative to episode payments for physicians in the same MSA and in the same specialty. A physician with a *p*-value < .0001 is considered an outlier in all analyses.

Table 4, Table 5 and Table 6 show the total number of physicians, the number of urologists, and the number of cardiologists, respectively, for each MSA, along with information on the range of physician sample sizes (episodes per physician) for 2002 and 2003. For our analyses, we include only physicians with at least 20 episodes. The numbers of physicians are lowest in Greenville and highest in Boston. Within each MSA, the number of physicians and the number of episodes per physician is similar between the two years.

Table 4: Counts Physicians and Episodes per Physician.

MSA	Total Physicians		Physicians with at least 20 Episodes		Episodes per Physician (Among Physicians with at Least 20 Episodes)							
					Mean		10th percentile		Median		90th percentile	
	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003
Boston	12,619	13,126	7,606	7,960	187	199	30	30	118	121	425	455
Greenville	1,958	2.054	1,558	1,642	374	365	54	47	292	280	805	819
Miami	4,870	5,104	3,511	3,653	231	229	32	32	140	138	538	521
Minneapolis	7,311	7,615	4,689	4,898	159	158	30	30	105	104	347	342
Orange Co.	4,763	4,922	3,216	3,451	205	205	30	31	128	128	484	481
Phoenix	5,943	6,356	4,027	4,287	208	212	30	31	124	126	486	490

Table 5: Counts of Urologists and Episodes per Urologist.

MSA	Total Urologists		Urologists with at least 20 Episodes		Episodes per Urologist (Among Urologists with at Least 20 Episodes)							
					Mean		10th percentile		Median		90th percentile	
	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003
Boston	140	142	120	126	280	306	82	72	273	296	481	549
Greenville	41	42	36	37	475	479	272	271	446	452	735	748
Miami	98	98	87	84	261	265	72	74	215	211	522	593
Minneapolis	78	80	65	63	266	264	50	103	256	244	490	486
Orange Co.	75	79	65	69	264	264	53	28	221	234	518	526
Phoenix	84	89	76	81	301	311	78	63	236	254	517	518

Table 6: Counts of Cardiologists and Episodes per Cardiologist.

MSA	Total Cardiologists		Cardiologists with at least 20 Episodes		Episodes per Cardiologist (Among Cardiologists with at Least 20 Episodes)							
					Mean		10th percentile		Median		90th percentile	
	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003	2002	2003
Boston	450	467	352	376	246	246	37	29	177	164	563	569
Greenville	61	63	53	54	398	399	142	201	413	400	567	555
Miami	212	219	197	197	324	324	81	62	270	268	626	632
Minneapolis	188	195	162	165	109	111	36	33	91	85	203	205
Orange Co.	165	168	140	149	335	326	112	55	294	281	578	581
Phoenix	222	245	186	200	280	286	59	50	234	232	538	615

The average and median number of episodes per physician included in the analyses are much higher in Greenville than they are in other MSAs. The variation across MSAs in episodes per physician might be an artifact of the data. Recall that the data contain only episodes from patients residing in the six MSAs and they exclude episodes from patients residing in other MSAs. For example, Boston physicians might have fewer episodes in these data because Boston physicians might serve a larger proportion of patients outside the Boston MSA than, say, Greenville physicians serve outside the Greenville MSA.

Overall and for cardiologists (but not urologists), Table 4 and Table 6 show that the average and median numbers of episodes per physician are dramatically lower in Minneapolis compared with other MSAs. This should be kept in mind for any comparisons between Minneapolis and other MSAs, especially for cardiologist episodes.

Results for Multilevel Models

We begin with some residual plots for urologists and cardiologists. Plots for other specialties are similar. We then describe results more generally.

Figure 7 displays the physician efficiency estimates for urologists in each of the six MSAs based on 2002 data. For each plot, the horizontal axis contains the urologist ranks, ranging from 1 to the number of urologists in the MSA, where the urologists are ordered from lowest to highest residual (from most efficient to least efficient). In the body of each plot there is one vertical bar per urologist. The middle of the bar represents the urologist’s estimated residual, and the bar endpoints represent the endpoints of a 99.98% confidence interval for the urologist’s residual. The blue horizontal bar represents an efficiency level that is 25 percent above average. Consequently, if the urologist’s bar is completely above the blue horizontal line, then the urologist’s residual is significantly above the 25 percent threshold at the .0001 significance level (one-sided). These red bars belong to the “outlier” physicians. We note that we did not find any outlier urologists in either Greenville or Minneapolis.

Notice that some black bars (non-outliers) are situated between red bars (outliers). These represent physicians who have estimated residuals as large or larger than the residuals for some outlier physicians, but with wider confidence intervals, either because they have smaller samples or because they have episodes with larger payment variances, or some combination of the two.

It is important to recognize that these graphs cannot be used to compare physicians to one another. Comparisons between physicians are invalid for at least two reasons. First, it is not valid to conduct a hypothesis test by comparing the degree of overlap between confidence intervals shown in this version of the graph. Second, each physician has his or her own mix of episodes, which the model effectively compares against a standard based on that particular mix of episodes. As an extreme example, physician 1 might only have episodes from disease A, while physician 2 might only have episodes from disease B. The two physicians cannot be compared directly. However, they can each be compared to standards based on diseases A and B, respectively.

Figure 8 displays the physician efficiency estimates for cardiologists in each of the six MSAs based on 2002 data. Each MSA has more cardiologists than urologists. Thus, there are more confidence bars plotted in Figure 8 compared with Figure 7. In Figure 8, the difference between the plots for Miami and Minneapolis is striking. The percentage of cardiologist outliers (in red) is higher for Miami than for Minneapolis. This is largely due to the higher average number of patients per cardiologist in Miami (324) compared with Minneapolis (109), shown earlier in Table 6. As a result, the Miami confidence intervals tend to be much shorter than the Minneapolis confidence intervals, in part leading to proportionately more significant residuals in Miami.

Figure 7: Urologist Physician Efficiency Estimates with 99.98% Confidence Limits, 2002.

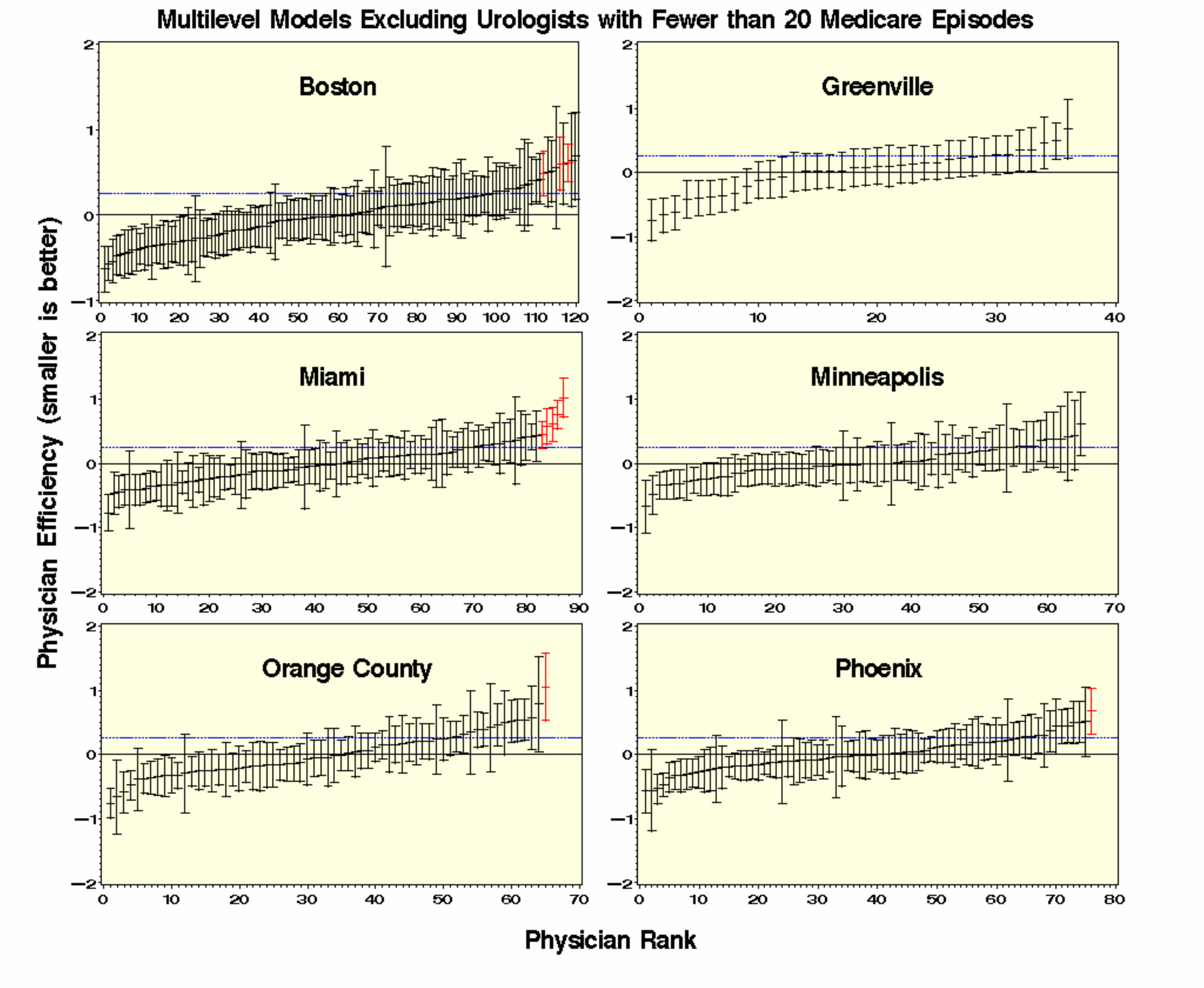


Figure 8: Cardiologist Physician Efficiency Estimates with 99.98% Confidence Limits, 2002.

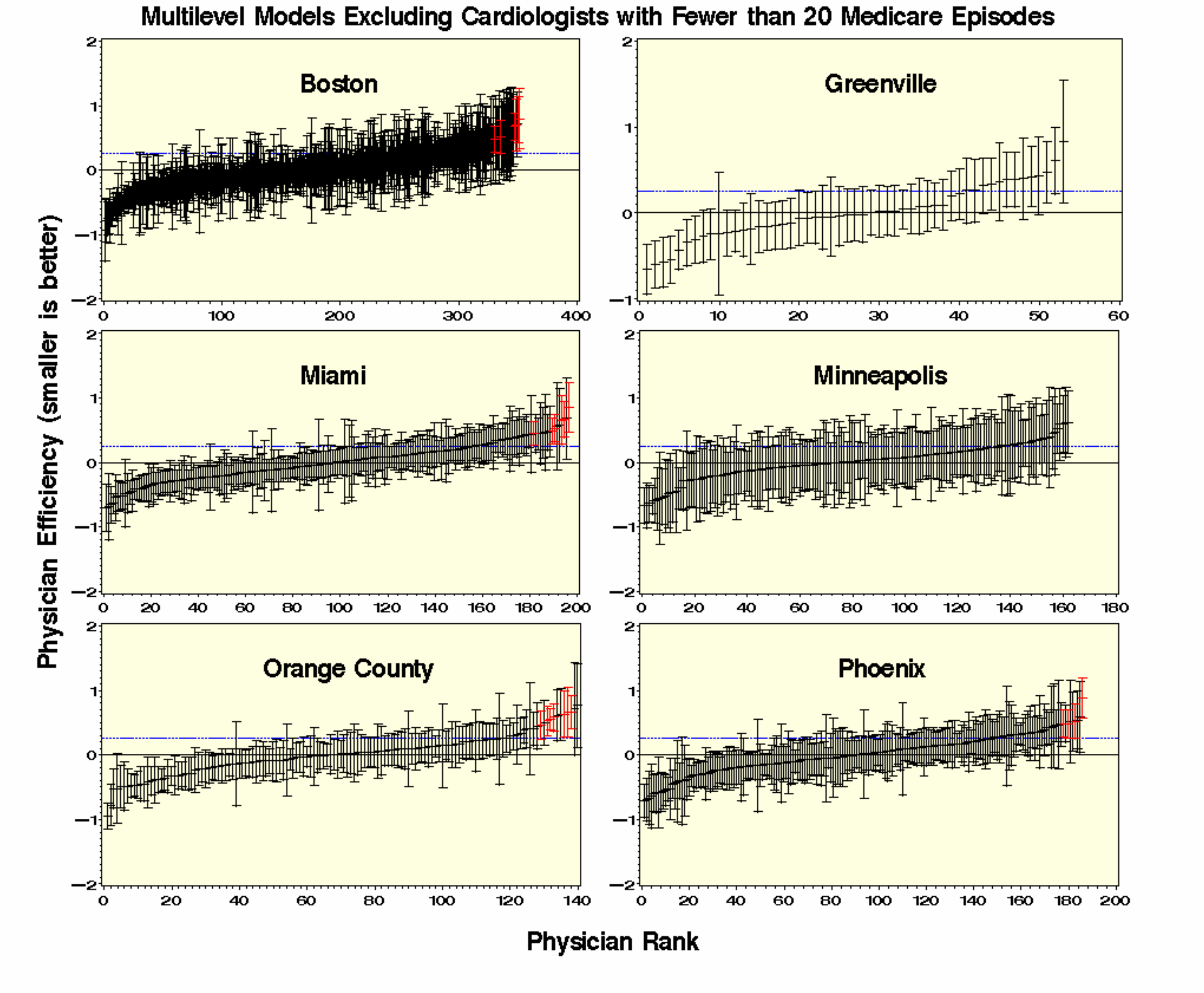


Figure 7 and Figure 8 are both based on 2002 data. They are meant only to illustrate the results of the underlying methodology. The plots for these two specialties based on 2003 data (not shown) are similar.

We now turn to a comparison of the 2002 outliers with the 2003 outliers, based on physicians who were present in the data in both years. Among physicians with at least 20 episodes in either year, about 85 percent had at least 20 episodes in both years. The percentage was higher—about 90 percent—for urologists and cardiologists.

Correlations between 2002 and 2003 efficiency scores, weighted by each physician's average number of episodes per year, are shown for urologists, cardiologists, and all physicians in Table 7. These correlations are quite high, indicating good year-to-year stability in the efficiency scores based on multilevel regressions. Physicians with high (low) efficiency scores in 2002 also tended to have high (low) scores in 2003.

Table 7: Correlation Between 2002 and 2003 Scores, Multilevel Models.

MSA	All MDs	Urologists	Cardiologists
Boston	0.90	0.89	0.88
Greenville	0.91	0.95	0.88
Miami	0.88	0.93	0.88
Minneapolis	0.86	0.85	0.79
Orange County	0.89	0.88	0.87
Phoenix	0.90	0.86	0.87
Total	0.89	0.89	0.87

To illustrate this correlation, Figure 9 contains plots of 2002 versus 2003 efficiency scores (residuals) for all physicians who had at least 20 episodes in both years. Clearly, the 2002 and 2003 efficiency scores are highly correlated. There is one graph per MSA. The upper right quadrant in each graph contains physicians whose efficiency score was positive in both years (higher than average payments). The lower left quadrant in each graph contains physicians whose efficiency score was negative in both years (lower than average payments). The other quadrants represent physicians whose efficiency score was positive in one year and negative in the other year.

In Figure 9, the light green circles represent physicians who were not declared outliers in either year. Although, not completely evident from the graph, they represent the vast majority of physicians (the green circles are highly concentrated near the center of the graph and they are partially overwritten by red and blue circles). In both years these physicians' residuals were not significantly 25 percent above average at the .0001 significance level. The red circles represent physicians who were labeled as an outlier in at least one year (*p*-value < .0001) and whose *p*-value was less than .05 in the other year. If a physician had a *p*-value under .0001 in either 2002 or 2003, and that same physician also had a *p*-value under .05 in the other year, then we have some confidence that the physician was a true outlier. The dark blue circles represent physicians who were declared as an outlier in one year, but whose *p*-value was greater than .05 in the other year. We regard these physicians as potentially "false outliers," although not necessarily. It is possible that they were truly an outlier in one year and truly not an outlier in the other year.

While Figure 9 clearly shows the high year-to-year correlation in the estimated physician residuals, it is difficult to judge the percentage of outliers because the point cloud is much more concentrated in the center than it is at the edges. Figure 10 and Figure 11 do a better job of summarizing the frequency and stability of the outliers.

Figure 10 is the “look forward” for 2002 outliers. Physicians are grouped by whether they were an outlier in 2002 and the bars in the plot show the percentage of physicians who had *p*-values < .05 in 2003. The red bars correspond to physicians who were outliers in 2002. There were 918 outliers (4.4 %) out of a total of 20,902 physicians with at least 20 episodes in both years. Of the 918 outliers, 833 (90.7 %) of them had a *p*-value under .05 in 2003.

Figure 11 is the “look backward” for 2003 outliers. Physicians are grouped by whether they were an outlier in 2003 and the bars in the plot show the percentage of physicians who had *p*-values < .05 in 2002. The red bars correspond to physicians who were outliers in 2003. There were 972 outliers (4.7 %) out of a total of 20,902 physicians with at least 20 episodes in both years. Of the 972 outliers in 2003, 861 (88.6 %) of them had a *p*-value under .05 in 2002.

In both years, about 6.4 percent of non-outlier physicians have *p*-values under .05. Nominally, we would expect about 5 percent of the non-outlier physicians to have *p*-values under .05. However, it’s possible that we failed to identify some outlier physicians because of the very low significance level that was required to attain outlier status, leading to somewhat more than 5 percent of physicians with *p*-values under .05 in the adjacent year.   Figure 10 and Figure 11 indicate that approximately 90 percent of the outliers identified in one year had low *p*-values in the other year, demonstrating fairly strong consistency between years.

Figure 9: Physician Efficiency Scores (Residuals), 2002 versus 2003, Multilevel Models.

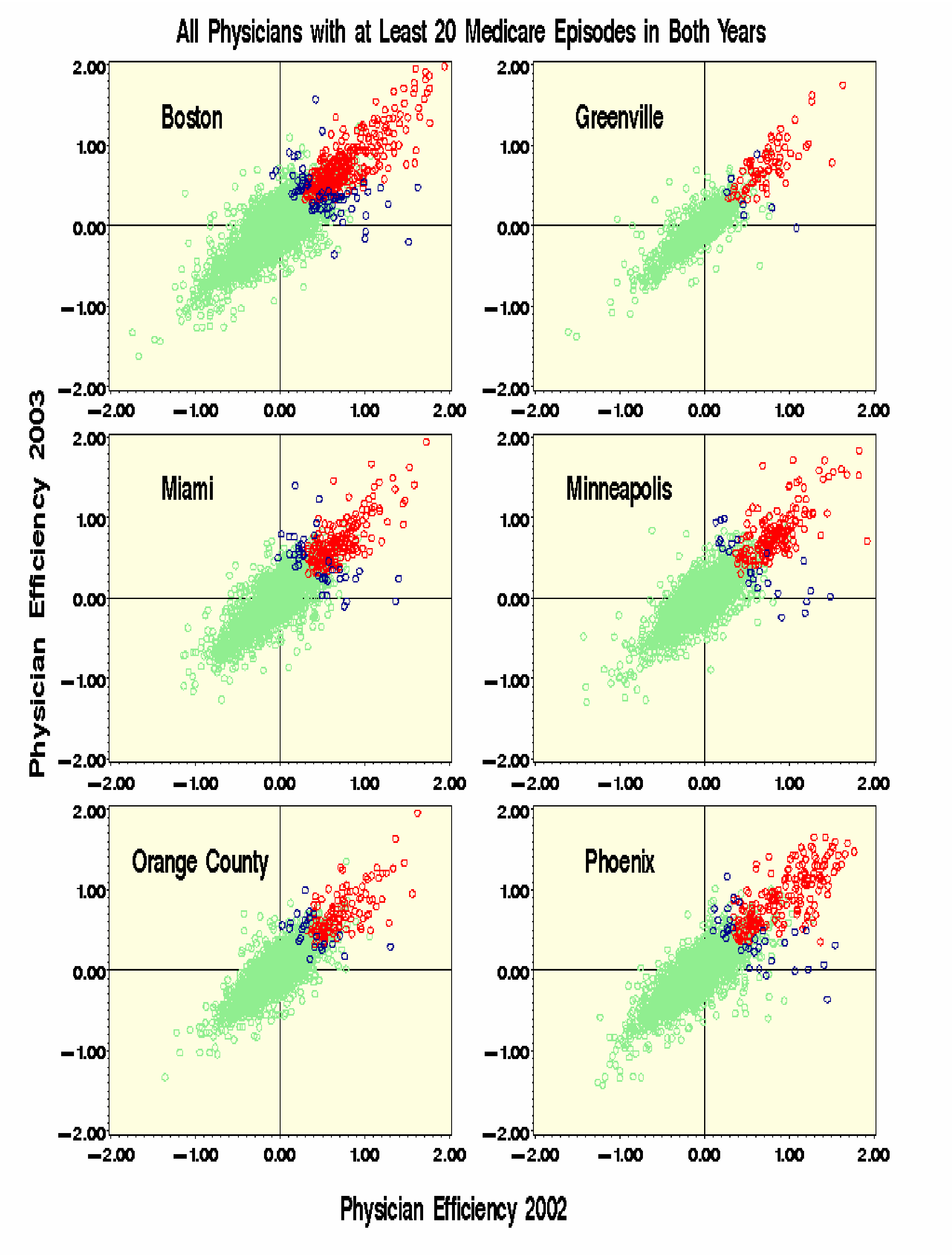


Figure 10: Look Forward: 2002 Outliers and 2003 *p*-values, Multilevel Models.

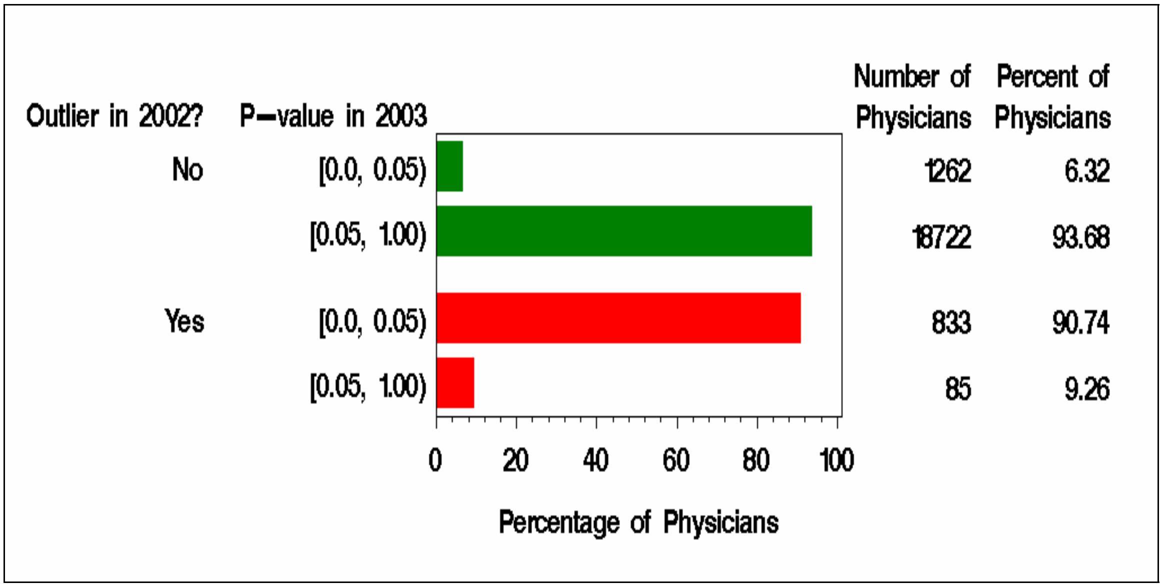
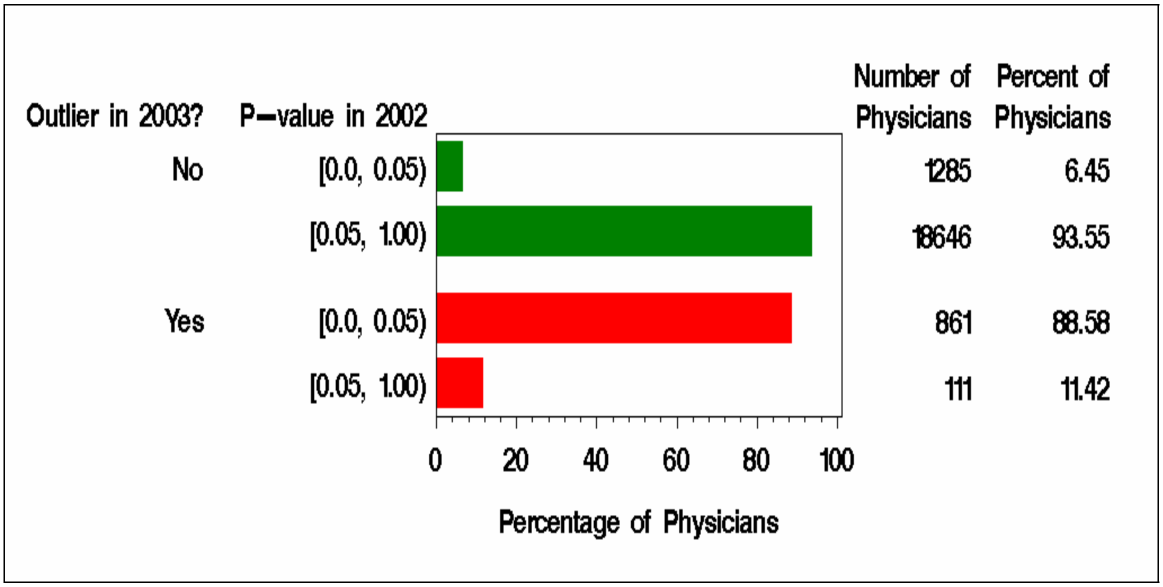


Figure 11: Look Backward: 2003 Outliers and 2002 *p*-values, Multilevel Models.



Results for Approximate Randomization Tests

For the sake of comparisons, we calculated residuals for the randomization tests that were on the same scale as residuals for the multilevel models. For each physician we calculated an "efficiency" score:

efficiency score = Observed mean log(payment) – Expected mean log(payment).

The observed mean log(payment) is the physician's average episode log(payment). The expected mean log(payment) is the average of the 10,000 Monte Carlo sample log(payment) means. This is the scale on which the multilevel models were based and on which the randomization tests were conducted. *P*-values were calculated as described in the methods section.

The 2002 and 2003 randomization test efficiency scores are highly correlated, as shown in Table 8. These correlations are nearly all within a couple of percentage points of the corresponding multilevel correlations shown earlier in Table 7. In these tables, the largest difference is between the multilevel correlations and the Monte Carlo correlations for Minneapolis cardiologists: 0.79 versus 0.73, respectively. In fact, of the year-to-year correlations shown, Minneapolis cardiologists have the lowest for both methods.

Table 8: Correlation Between 2002 and 2003 Scores, Randomization Tests.

MSA	All MDs	Urologists	Cardiologists
Boston	0.87	0.87	0.84
Greenville	0.89	0.93	0.88
Miami	0.86	0.93	0.87
Minneapolis	0.84	0.81	0.73
Orange County	0.84	0.88	0.82
Phoenix	0.88	0.84	0.85
Total	0.87	0.88	0.84

Figure 12 plots 2002 versus 2003 efficiency scores for all physicians. The symbols in this plot are defined the same was as they were for the multilevel regression results in Figure 9. The green circles represent physicians who were not declared outliers in either year. The red circles represent physicians who were declared as an outlier in at least one year (*p*-value < .0001) and whose *p*-value was less than .05 in the other year. The blue circles represent physicians who were declared as an outlier in one year, but whose *p*-value was greater than .05 in the other year.

Figure 13 and Figure 14 summarize the frequency and stability of the randomization test outliers.

Figure 13 is the "look forward" for 2002 outliers. Physicians are grouped by whether they were an outlier in 2002 and the bars in the plot show the percentage of physicians who had *p*-values < .05 in 2003. The red bars correspond to physicians who were outliers in 2002. There were 611 outliers (2.9 %) out of a total of 20,911 physicians with at least 20 episodes in both years. Of the 611 outliers, 572 (93.6 %) of them had a *p*-value under .05 in 2003.

Figure 14 is the "look backward" for 2003 outliers. Physicians are grouped by whether they were an outlier in 2003 and the bars in the plot show the percentage of physicians who had *p*-values < .05 in 2002. The red bars correspond to physicians who were outliers in 2003. There were 712 outliers (3.4 %) out of a total of 20,911 physicians with at least 20 episodes in both years. Of the 712 outliers in 2003, 641 (90.0 %) of them had a *p*-value under .05 in 2002.

In both years, about 7.9 percent of non-outlier physicians have *p*-values under .05. Nominally, we would expect about 5 percent of the non-outlier physicians to have *p*-values under .05. However, it is possible that we failed to identify some outlier physicians because of the very low significance level that was required to attain outlier status, leading to somewhat more than 5 percent of physicians with *p*-values under .05 in the adjacent year.   Figure 13 and Figure 14 indicate that more than 90 percent of the outliers identified in one year had low *p*-values in the other year, demonstrating appreciable year-to-year stability.

Figure 12: Physician Efficiency Scores (Residuals), 2002 versus 2003, Randomization Tests.

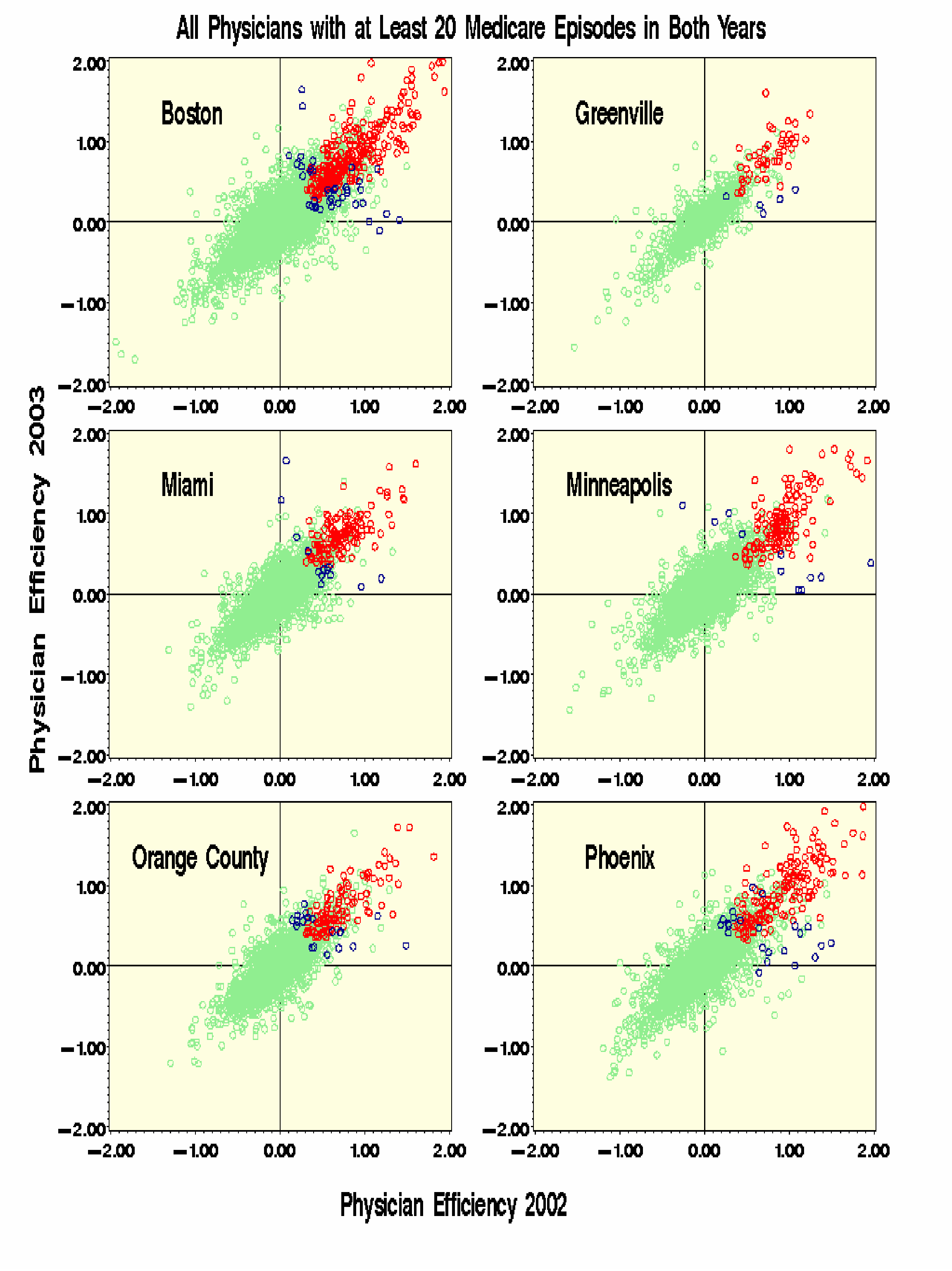


Figure 13: Look Forward: 2002 Outliers and 2003 *p*-values, Randomization Tests.

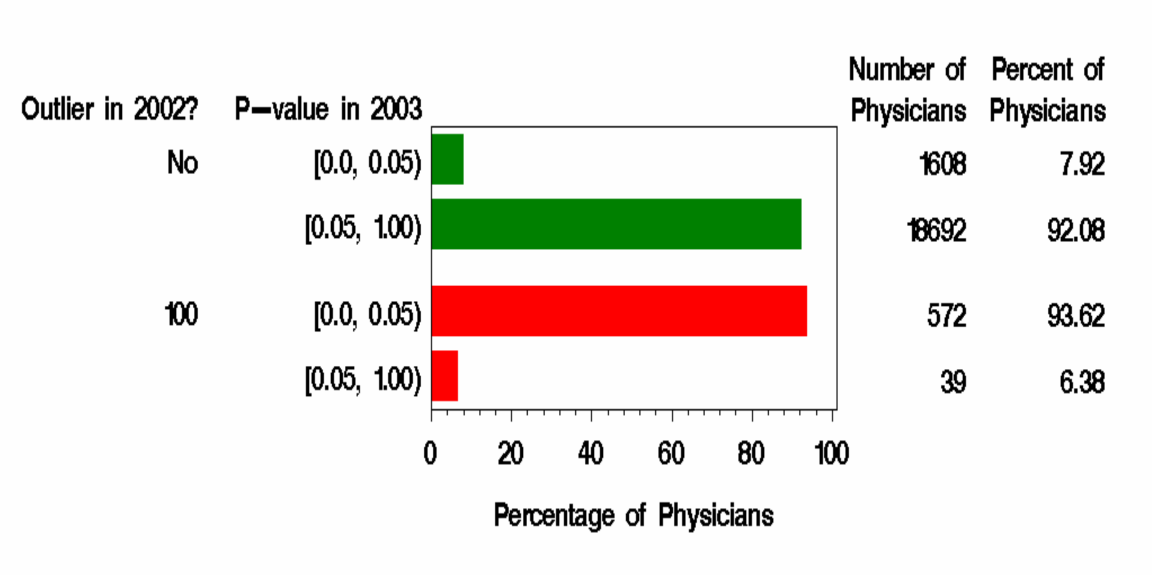
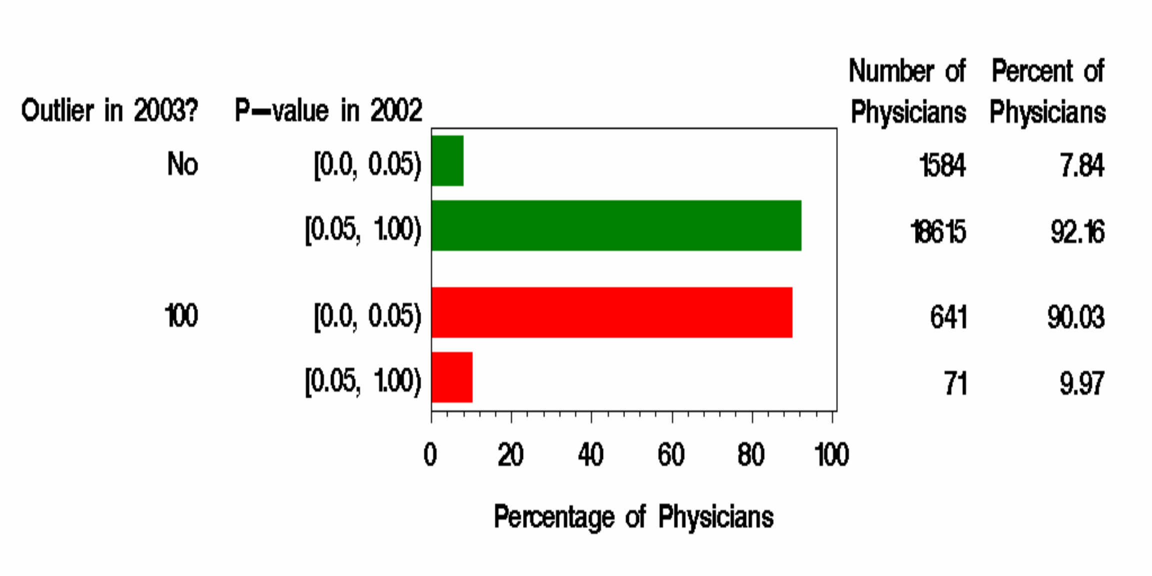


Figure 14: Look Backward: 2002 Outliers and 2003 *p*-values, Randomization Tests.



Comparison of Multilevel Model Results to Randomization Test Results

The correlation between the residuals (estimated efficiencies) for the two methods is quite high, at 92.6 % for both 2002 and 2003. Figure 15 and Figure 16 show plots of the estimated efficiencies for 2002 and 2003, respectively. The two methods are in substantial agreement: residuals that are high (low) for one method tend to be high (low) for the other method.

Table 9 and Table 10 compare the outlier status of physicians between the two methods for 2002 and 2003, respectively. In both years, the randomization test identified a lower percentage of outliers (2.9 % and 3.5 %) compared with the multilevel model (4.3 % and 4.9 %). In 2002, 550 physicians were identified as outliers by both methods, representing 86 % of the randomization test outliers and representing 57 % of the multilevel outliers. In 2003, 679 physicians were identified as outliers by both methods, representing 84 % of the randomization test outliers and representing 59 % of the multilevel outliers.

Therefore, for outlier identification the multilevel model approach supports the randomization test approach more than the reverse. A further strategy, which we did not explore, would be to identify as outliers only those physicians who where identified as outliers using both approaches in a given year.

Table 9: Comparison of Outlier Results: All Physicians, 2002.

Randomization Test Result	Multilevel Model Results		
	Not outlier	**Outlier**	**Total**
Not outlier	21,206 (95.3 %)	407 (1.8 %)	21,613 (97.1 %)
Outlier	90 (0.4 %)	550 (2.5 %)	640 (2.9 %)
Total	21,296 (95.7 %)	957 (4.3 %)	22,253 (100.0 %)

Table 10: Comparison of Outlier Results: All Physicians, 2003.

Randomization Test Result	Multilevel Model Results		
	Not outlier	**Outlier**	**Total**
Not outlier	22,088 (94.6 %)	466 (2.0 %)	22,554 (96.6 %)
Outlier	127 (0.5 %)	679 (2.9 %)	806 (3.5 %)
Total	22,215 (95.1 %)	1,145 (4.9 %)	23,360 (100.0 %)

Figure 15: Randomization Test Efficiency vs. Multilevel Model Efficiency, 2002.

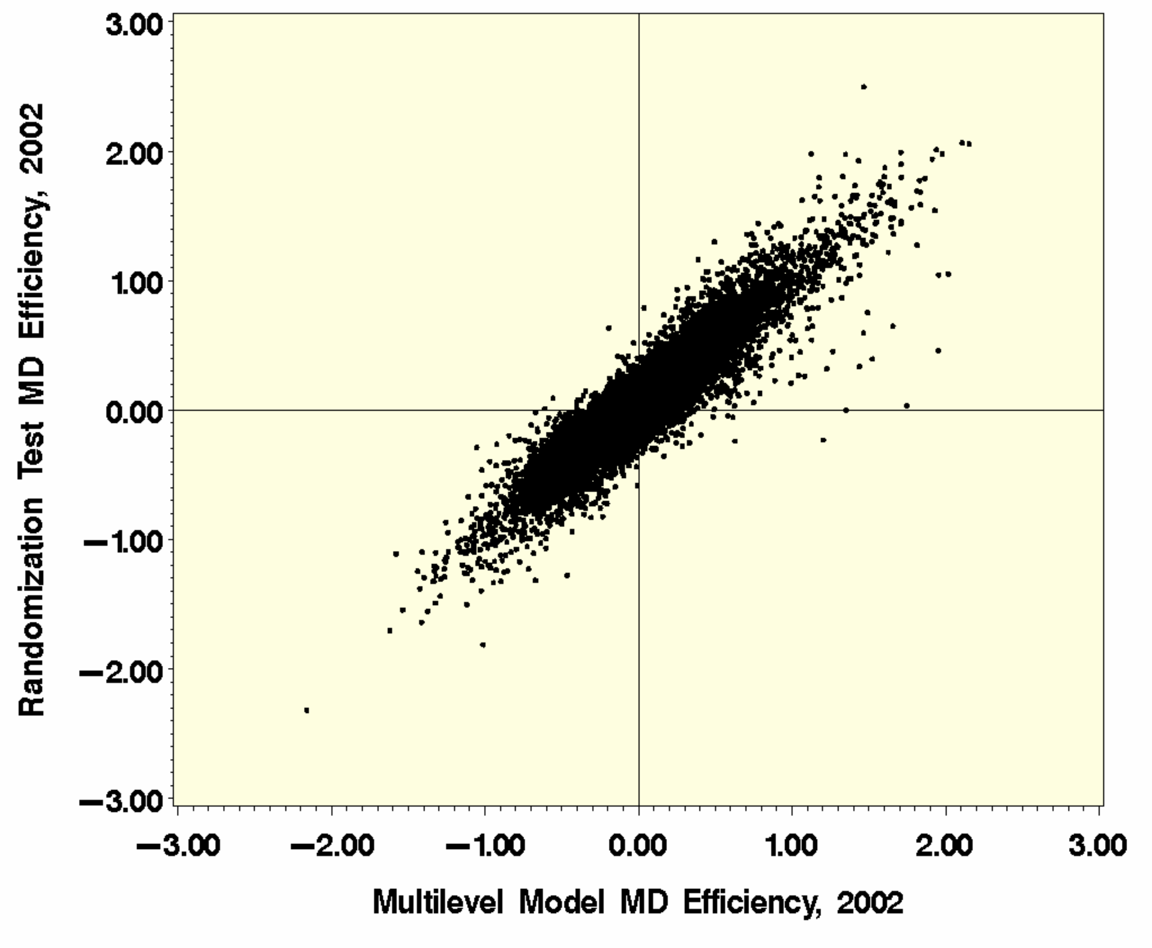
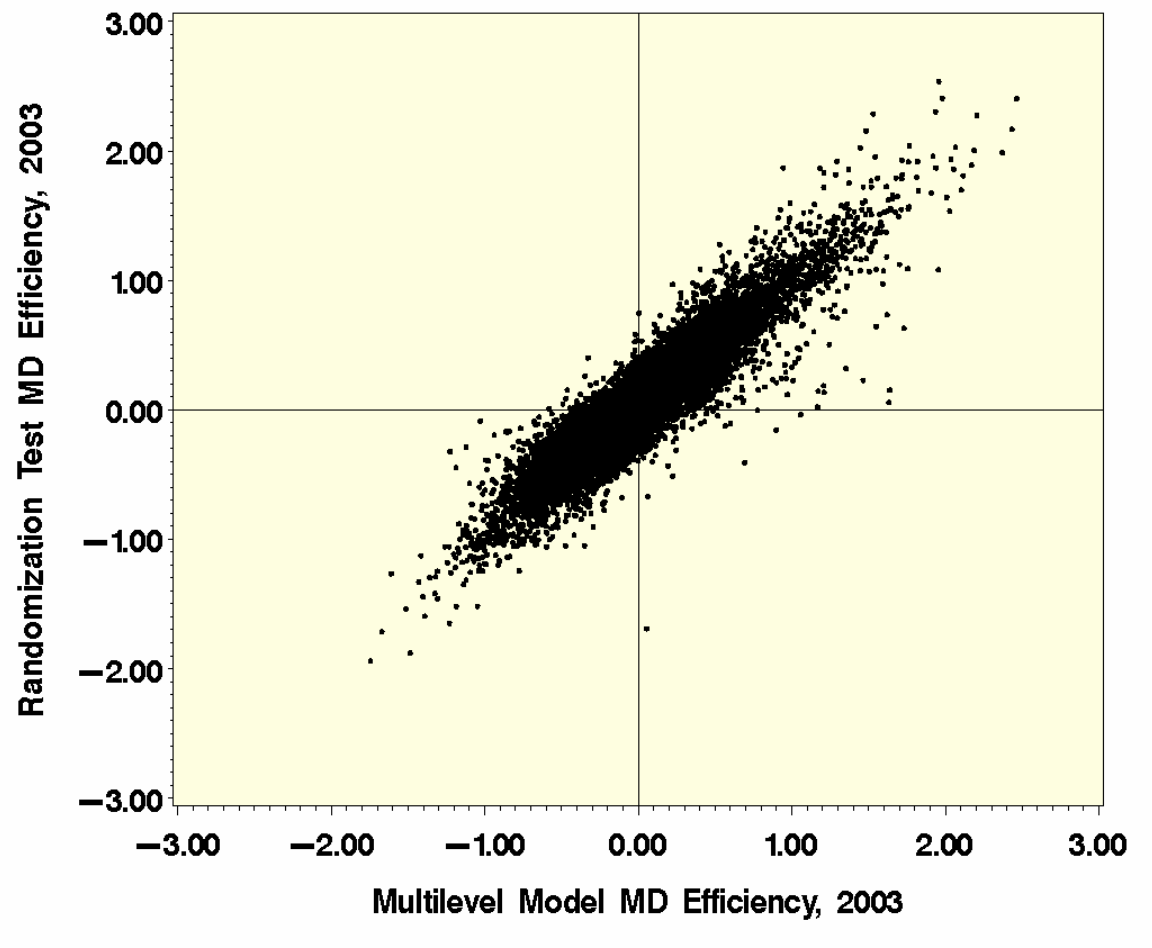


Figure 16: Randomization Test Efficiency vs. Multilevel Model Efficiency, 2003.



CONCLUSIONS AND RECOMMENDATIONS

These analyses indicate that episodes can be useful for identifying physicians with extraordinarily high average standardized payments for episodes of Medicare patient treatments. We employed two statistical approaches to identify outliers: multilevel models and randomization tests. Both approaches produced consistent year-to-year results based on 2002 and 2003 Medicare data. About 90 percent of the outliers identified in one year had significantly high average payments in the other year. Both methods are practical for wider application to the Medicare physician population. All analyses were performed with a personal computer using SAS Version 9.1[7] on a Microsoft Windows XP Professional operating system.

For both approaches we tested whether each physician's average episode payment exceeded an average physician's mean payment by at least 25 percent at the .0001 level of significance. Other thresholds and significance levels could be used. These thresholds produced a somewhat high percentage of physician outliers—between 3 and 5 percent—but it is a reasonable place to start. A higher threshold, for example 150 percent of an average physician's mean payment, would produce fewer outliers.

About 85 percent of the randomization test outliers were also identified as outliers by the multilevel models. Conversely, about 58 percent of the multilevel outliers were also identified as outliers by the randomization tests. This disparity is largely explained by the higher percentage of outliers generated by the multilevel model compared with the randomization test. A slightly higher threshold for the multilevel model might have produced a similar percentage of outliers. In any case, the correlation was high—about 90 percent—between the underlying physician efficiency measures estimated by the two methods. On balance, we slightly favor the randomization approach for a few reasons.

First, the randomization test requires fewer assumptions concerning statistical distributions than the multilevel approach does. For example, it does not assume that distributions be normal. It assumes that the episodes attributed to other physicians are representative of episodes treated by the physician under test. In reality, episodes treated by other physicians are probably correlated to some extent within physicians. However, we believe that this would work in favor of the physician under test because the episode payments sampled across physicians would tend to increase variability in the "null" distribution of means.

Second, it is easier to explain randomization tests than multilevel models. The histogram of 10,000 sample means provides an intuitive benchmark against which to compare the physician's mean.

Third, the randomization approach more easily allows for a drill-down on the payment distribution. This is illustrated by Figure 17. Dr. Smith is an outlier with a mean episode payment of $1,301. This is well into the tail of the expected mean payment distribution as shown in the histogram for total episode payments. Upon disaggregating the payments into service categories. It can be seen that Dr. Smith's payments are especially excessive for procedures. Therefore, Dr. Smith could start by analyzing his or her utilization of procedures compared to other physicians' utilization of procedures for similar cases.

[7] SAS Institute Inc., Cary, NC, USA.

This study has the following limitations:

1. Standardized payments were the basis for measuring episode resource intensity and physician "efficiency." For example, hospital payments were the same for every patient hospitalized with a given diagnosis related group. This standardization no doubt masked some true episode cost variation.
2. Each episode was attributed to the single physician that billed the highest percentage of E&M dollars (at least 35 %) for that episode. For episodes involving multiple physicians, it is possible that less than full responsibility should have been accorded to that physician.
3. Risk adjustment was based on episode severity as measured by the episode's principal disease, the stage of the principal disease, and the relative risk score. Although these factors incorporated patient diagnoses and demographics, other factors might have provided further risk adjustment.
4. Physician comparisons were based only on episodes attributed to physicians within the same specialty group and within the same MSA. There might be an argument for comparing performance across a broader spectrum of specialties and geographic areas.
5. These analyses were strictly episode-based. They only compared physicians on their average episode-level resource intensity. They did not account for the frequency of episodes. It is possible that some physicians broke up the treatment for a condition into several low-intensity episodes, while other physicians combined the treatment for a condition into a few high-intensity episodes. However, the several low-intensity episodes would have had to have been widely spaced to create separate episodes using the MEG algorithms.
6. The episodes in this analysis were based on the MSA of the patient, not on the MSA of the physician. For example, all episodes for Boston physicians were based solely on patients residing in the Boston MSA. However, this excluded episodes for patients outside the Boston MSA that were treated by Boston physicians.

To partially address the second point, MedPAC has commissioned a study currently under way to test multiple physician attribution in place of single physician attribution. We also recommend that MedPAC should repeat the analyses in the present study to address the sixth limitation. If some physicians treated a large number of patients outside their own MSA, then their estimated mean episode payment could have been biased to the extent that those patients had different treatment patterns compared with patients in the physician's own MSA. At the least, the larger sample of episodes could produce a more reliable estimate of their mean episode payments.

Figure 17: Distribution of Mean Payments from 10,000 Monte Carlo Samples for Dr. Smith

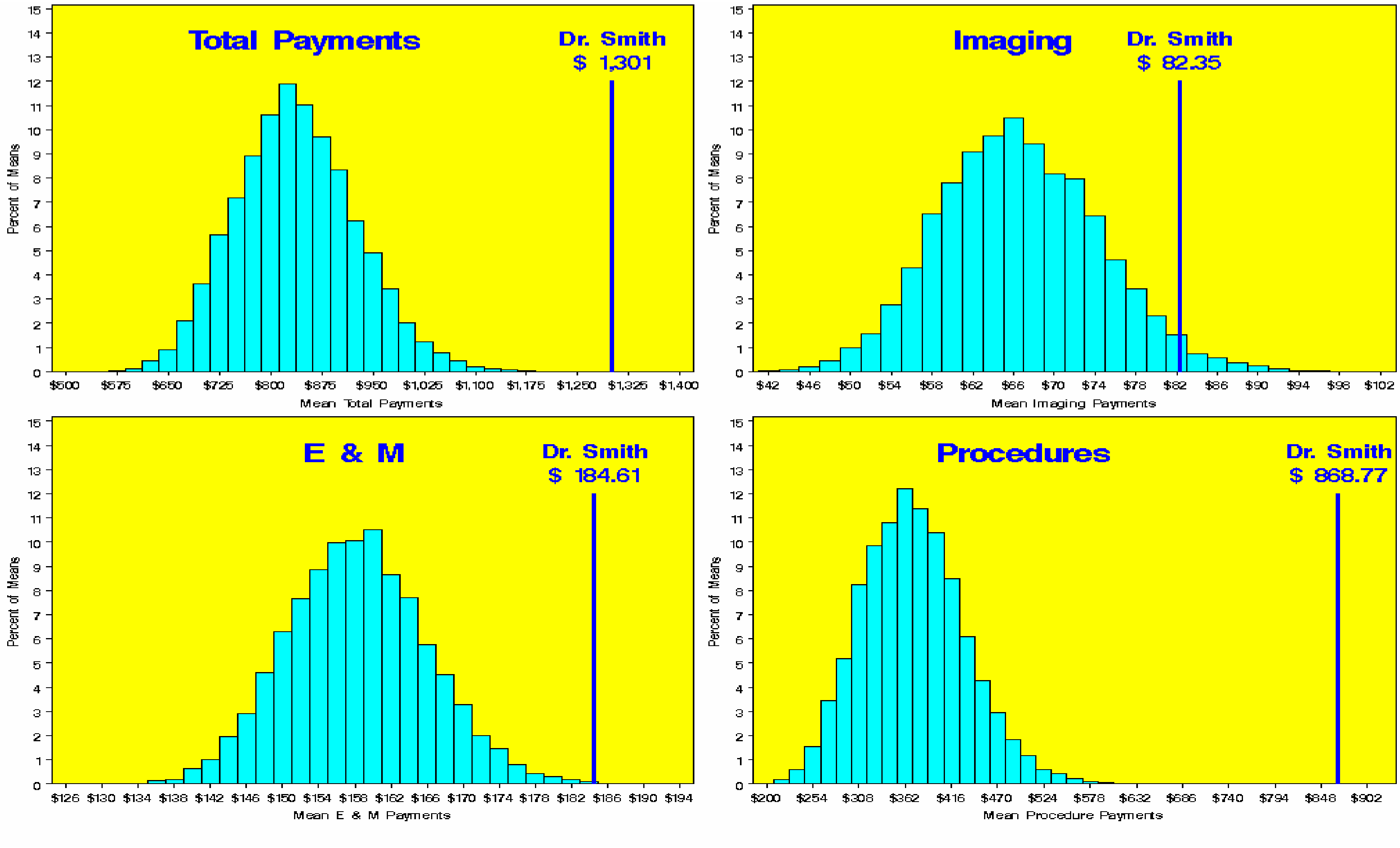
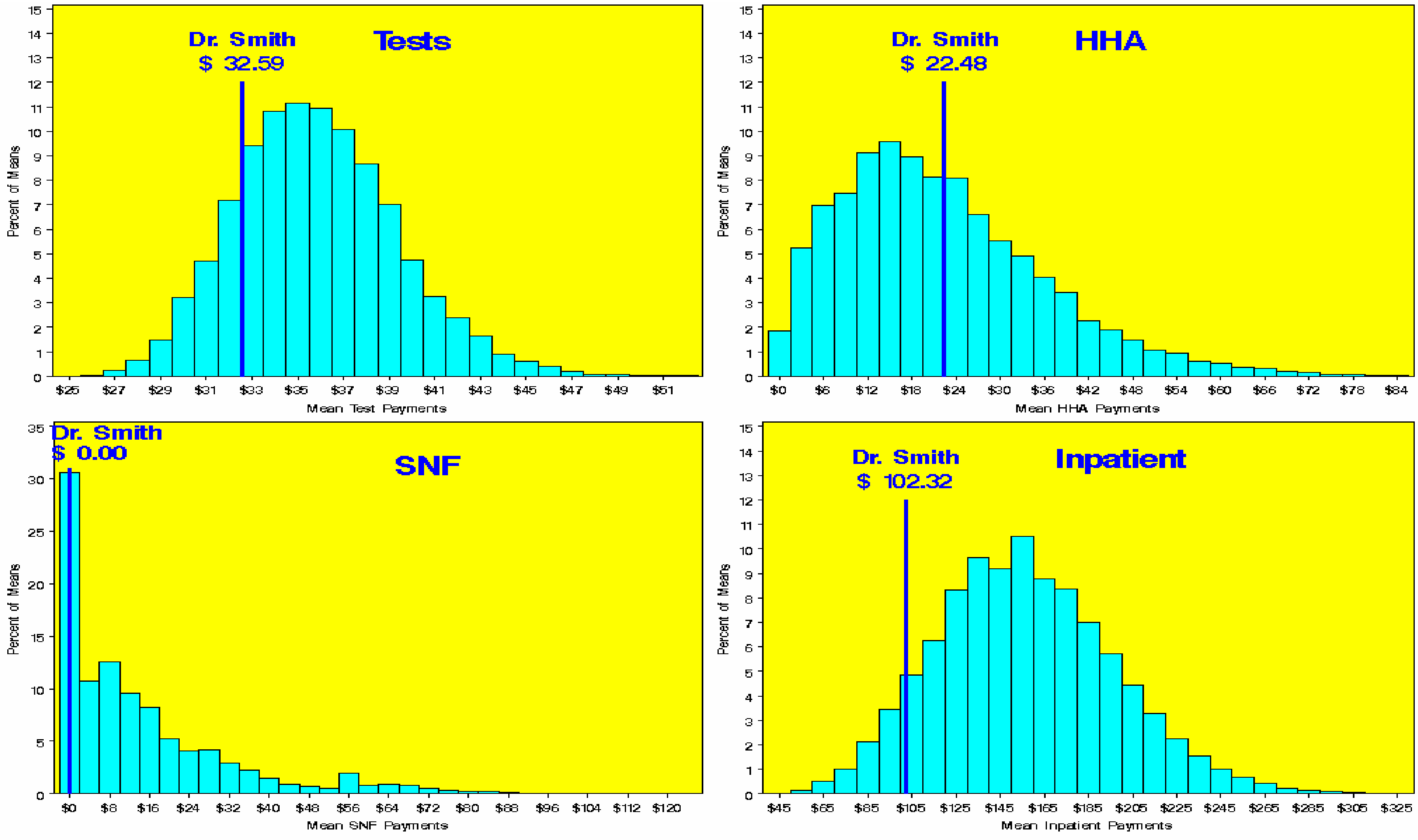


Figure 17 (continued)



APPENDIX

This appendix contains more details concerning the multilevel models, the Medical Episode Grouper (MEG), the physician attribution rules, and the application of MEG to the claims data.

Further Adjustments to the Multilevel Models

We modified the multilevel model (equation 1) as follows. Using the property of logarithms that ln(a/b) = ln(a) – ln(b), the regression Model 1 can be rewritten as:

$$\ln\left(O_{ij}\right)-\ln\left(E_{ij}\right)=\beta_{0j}+e_{ij}$$

and adding $\ln(E_{ij})$ to both sides yields the equation:

$$\ln\left(O_{ij}\right)=\beta_{0j}+\ln\left(E_{ij}\right)+e_{ij}$$

which we generalize by adding a regression coefficient β_1 to the term $\ln(E_{ij})$, resulting in the equation:

$$\ln\left(O_{ij}\right)=\beta_{0j}+\beta_1\ln\left(E_{ij}\right)+e_{ij}$$

This is a multilevel model with episodes at the first level, grouped within physicians at the second level. It is called a “random intercepts” model because the physician-specific intercepts are assumed to be random variables (distributed according to a normal distribution).

We also added terms to the right side for the top K high-frequency diseases (MEGs) by including a binary (0, 1) indicator D_k for disease k, k=1, 2, 3,…, K:

$$\ln\left(O_{ij}\right)=\beta_{0j}+\beta_1\ln\left(E_{ij}\right)+\sum_{k=1}^{K}\delta_k D_k+e_{ij}$$

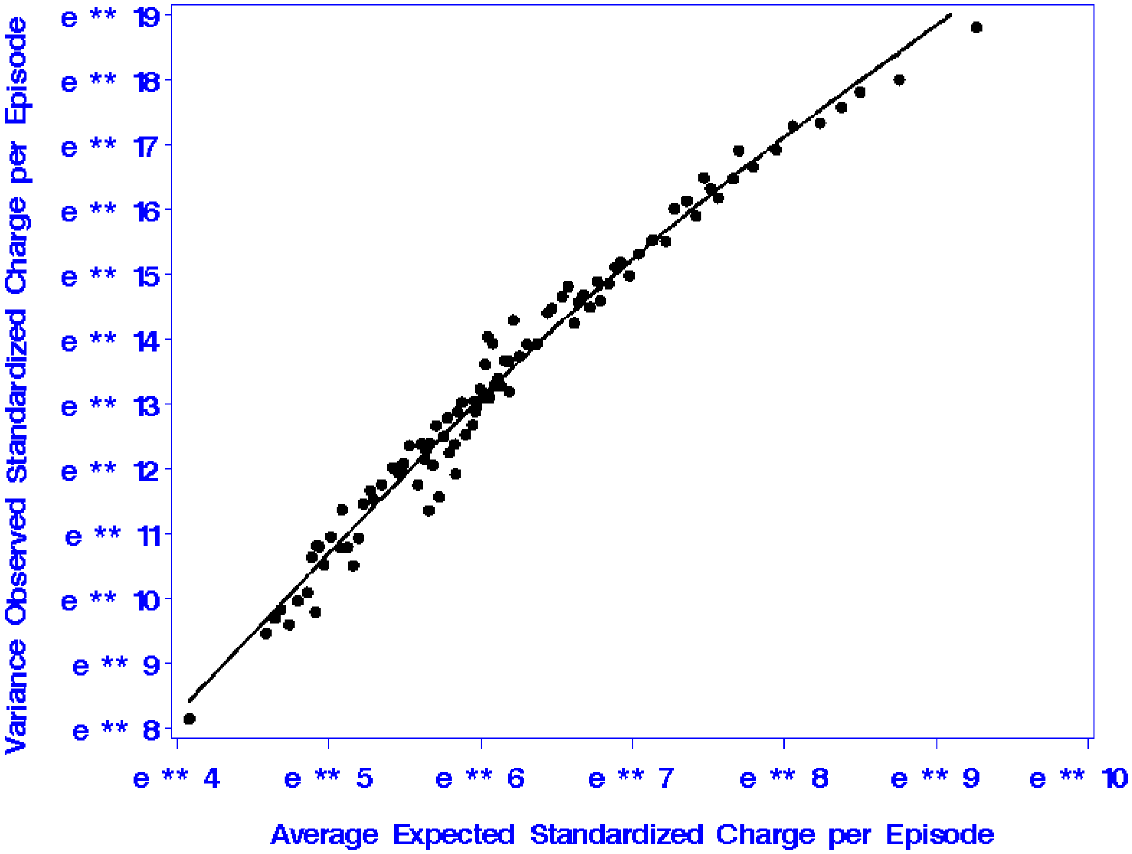
The corresponding regression coefficient for disease k, δ_k, ensures that the residuals average zero for disease k. In other words, the average predicted response will equal the average observed response for each of the K diseases included in the model. Thus, physician efficiency will be measured relative to the average efficiency for each disease.

Finally, as shown in Figure A1, episode payments tend to vary more for high-payment episodes than for low-payment episodes. Therefore, we adjusted the model so that the episode-level residual variance—denoted by σ_e^2 in the above regression Model 1—varies with the expected episode payment, E_{ij}. Notice that the axes in Figure 4 are log scaled. The variance in our regression is with respect to the ln(payment). Therefore, we specify the variance function:

$$Var(e_{ij})=\sigma_e^2\left[\ln(E_{ij})\right]^{\theta}$$

In words, the residual variance of the logarithm of the *observed* episode payment was assumed to be proportional to a power, θ, of the logarithm of the *expected* episode payment. The variance function shown in Figure A1 is based on all episodes across the six MSAs. The variance function for subsets of the data we analyze are always an increasing function, but the shape of the curve varies across subsets.

Figure A1: Episode Payment Variance versus Average Expected Payment.



Data source: All episodes for Medicare patients from six MSAs during 2002.

It is important to model this variance function properly because means for physicians who tend to treat high-payment episodes will have higher standard errors than those for physicians who tend to treat low-payment episodes. Therefore, physicians who tend to treat high-payment episodes should given greater latitude when they are compared to the overall average.

With these modifications, the final multilevel regression model was of the form:

$$\begin{aligned}
&\ln\left(O_{ij}\right) = \beta_{0j} + \beta_1 \ln\left(E_{ij}\right) + \sum_{k=1}^{K} \delta_k D_k + e_{ij} \\
&\beta_{0j} = \beta_0 + u_j \\
&u_j \sim N\left(0, \sigma_u^2\right) \\
&e_{ij} \sim N\left(0, \sigma_e^2 \left[\ln(E_{ij})\right]^{\theta}\right)
\end{aligned} \qquad \text{(Model 2)}$$

The residual for physician j, u_j, can be taken as a measure of efficiency for physician j. In fact, small (absolute) values of u_j can be interpreted as percentages above and below the "average" physician. For example, a value of $u_j = 0.10$ indicates that the mean episode payment for physician j is about 10 percent above the average.

This remainder of this appendix contains information regarding the Medical Episode Grouper^{TM} (MEG), the physician attribution methodology, and the results of applying these methods to the MedPAR claims described in the Data section of this report.

MEG – Medical Episode Group™

An episode of care describes a series of related health care services for the treatment of a given spell of illness. Episodes can be comprised of inpatient admissions, outpatient services, and prescription drugs. The Medical Episode Grouper (MEG) was commercially released in 1998.

All episode grouping methods are built on two central concepts; a disease classification system and an episode grouping logic.

Disease Staging is the disease classification system that forms the foundation of MEG episode groups. Disease Staging defines levels of biological severity for specific diseases – episodes of care – where illness severity is defined as the risk of organ failure or death. The severity levels include:

Stage 0:	History of or exposure to a disease.
Stage 1:	The disease involves no complications
Stage 2:	The disease involves local complications
Stage 3:	The disease involves multiple sites, or has systemic complications
Stage 4:	Death

In the definition of the Disease Staging criteria, most diseases begin at Stage 1 and continue through Stage 4. There are several exceptions to this rule. Some self-limiting diseases, such as cataracts, do not include a Stage 3 or 4. Other criteria begin at either Stage 2 or 3 since they are often complications of other diseases (e.g., bacterial meningitis, which can be a complication of sinusitis, otitis media, or bacterial pneumonia). Stage 0 has also been included in the classification of diseases for patients with a history of a significant predisposing risk factor for the disease, but for whom there is currently no pathology (e.g., history of carcinoma).

The MEG episode grouping logic dictates the accumulation of claims into episode groupings, and allocates claims into discrete episodes of care. The logic employed by MEG includes:

Starting Points - An episode of care is initiated by a contact with the health delivery system and is generally the first claim received for a given disease. The MEG methodology allows physician office visits and hospitalizations to initiate patient episodes. The coding of claims for imaging services and laboratory tests are not always reliable. Therefore, such claims can join existing episodes but they cannot create new episodes.

Clean Periods - Episode Duration – The MEG episode logic is designed to capture all relevant treatments related to a given episode. The end of an episode cannot be directly determined from medical claims. Therefore, episodes are deemed complete when a specified “clean period” has passed without claims related to an episode that has been initiated.

Episode Severity – Episode severity is defined as the highest Disease Staging severity stage observed during the episode.

Multiple Diagnosis Codes – It is often the case that a professional claim (or claim lines) will have two or more diagnosis codes associated with a single procedure code. The episode grouper determines which diagnosis code is most related to the procedure, ensuring the accurate allocation of claims to episode groups.

Lookback – Frequently, tests are ordered and performed before a patient has been diagnosed. Since lab tests and imaging studies cannot initiate an episode, a “lookback” logic links an established episode to these claims if they are clinically related to the episode’s disease. If such an episode is found within 15 days of a lab or imaging claim, it is added to the episode.

Inclusion of Non-specific Coding - Non-specific, initial diagnoses are relatively common in the billing of treatments for patients. The inclusion logic is a process that examines each episode after the initial grouping to determine whether a non-specific episode (e.g., Episode Group 179, "Other Gastrointestinal or Abdominal Symptoms") can be included with a clinically related specific episode (e.g., Episode Group 138, "Appendicitis").

Drug Claims - MEG has been designed to ‘group’ drug claims into episodes. National Drug Code (NDC) information is reviewed by clinical and coding experts and mapped to each episode group. This mapping is then used to assign drug claims to episodes.

Complete Episodes –Episodes of care are created from claims datasets that span a given period of time and are used to profile and evaluate the economic efficiency of physicians. Since it cannot be known whether an episode that was initiated by a claim near the beginning of the dataset is the true beginning of the episode or would join an existing episode created earlier if the data had been available, the true payment for the episode could be understated. Analogously, it cannot be known whether episodes near the end of the dataset would extend beyond this date if more data were available. Episodes which may understate the true payments for treatment are removed from the dataset prior to analysis. The remaining are considered to be complete.

A complete episode is defined to be an episode that begins later than the beginning date of the claims data set plus the number of days of the episode clean period. For example, if a given dataset is comprised of claims occurring on or after January 1 and an episode with a clean period of 30 days begins on January 15, it cannot be known whether this is the true beginning of the episode or if it would have been created in December. In this case, the episode would be considered incomplete and would not used when profiling the physician responsible for the episode.

Based on the episodes created from the four years of medical claims, two study periods – 2002 and 2003 – were established. Episodes with a beginning date in 2002 were assigned to that year. Similarly, the 2003 episode data set was determined. The episodes falling into 2002 and 2003 are complete episodes because a full year of claims data preceded the 2002 data and followed the 2003 data. This ensured that episodes bridging two years would not be eliminated from the study data. Consequently, some 2002 episodes extended into 2003 and some 2003 episodes extended into 2004.

Risk Adjusted Expected Episode Payments

Risk adjusted expected episode payments were calculated for the 2002 and 2003 episodes. *Episode severity* was measured by the stage of disease for that episode. *Patient complexity* was measured by the DCG relative risk scores (RRS) for the patient treated in the episode. The RRS is an estimate of the expected medical payments for a patient based on the patient’s age, gender, and the medical conditions for which the patient was treated over a specified period of time, usually one year (Ellis and Ash, 1995; Ash, et al., 2000).

For each MEG, a table was constructed with rows for each integer stage of disease, and with up to 5 columns corresponding to five RRS categories corresponding to consecutive ranges of relative risk score values. For each MEG, the ranges for the five RRS categories were determined by maximizing the variance explained over the entire episode file.

Each cell in the MEG table was then populated with an “expected payment” calculated as the average payment taken over all episodes within the table cell (excluding outlier payments). Finally, these expected episode payments were assigned to each episode based on the episode’s MEG disease, the stage of disease, and the patient’s RRS. These expected payments were used in the multilevel models.

Summary of Episode Grouping Results

Since lab and imaging claims can only join an existing episode and not create one, there will be claims that cannot be grouped to episodes. Table 11 summarizes the percent of claims and payments that comprised disease-specific episodes – ‘grouped’ - and those that could not be associated with an episode – ‘ungrouped’. In 2002, 96.5 percent of the payments and 84.5 percent of the claims were assigned to episodes. In 2003, 96.6 percent of payments and 85.6 percent of claims were grouped. These results are consistent with earlier episode studies conducted by MedPAC.

Table 11: Summary of Grouped and Ungrouped Claims, 2002 and 2003.

2002	**Payments**	**Percent**	**Claims**	**Percent**
Ungrouped	$276,264,425	3.5	11,544,787	15.5
Grouped	$7,662,107,877	96.5	62,709,123	84.5
Total 2002	$7,938,372,302	100.0	74,253,910	
2003				
Ungrouped	$301,534,386	3.4	11,391,248	14.4
Grouped	$8,649,500,603	96.6	67,587,729	85.6
Total 2003	$8,951,034,989	100.0	78,978,977	100.0

Outliers were removed from the data set to mitigate the likelihood that a single extreme episode would unduly influence the analysis of a physician’s performance. Outliers were defined as episodes with total payments falling below the 1st percentile (low outliers) and above the 99th percentile (high outliers) of episode payments within each integer stage of a MEG. Standardized payments were missing on about 0.7 percent of the claims in the data, and episodes created from these claims were also excluded from this study.

Table 12 and Table 13 display the effects of removing outliers from the 2002 and 2003 episode datasets. In 2002, 7.8 percent of the episodes were eliminated representing 6.1 percent of the claims and 16.0 percent of the payments. In 2003, 8.2 percent of episodes representing 6.1 percent of claims and 16.6 percent of payments were outliers.

Table 12: Episode Exclusions, 2002.

	Episodes		Claim Records		Payments	
	Number	Percent	Number	Percent	Total	Percent
Episodes 1-560	8,455,433	100.0	62,709,123	100.0	$7,662,107,877	100.0
Exclusions	663,010	7.8	3,817,305	6.1	1,228,408,980	16.0
Low Outliers	570,505	6.7	1,112,733	1.8	$9,444,155	0.1
High Outliers	83,956	1.0	2,687,917	4.3	$1,218,964,825	15.9
Missing Pmt.	8,549	0.1	16,655	0.0	Unknown	0.0
Study Episodes	7,792,423	92.2	58,891,818	93.9	$6,433,698,897	84.0

Table 13: Episode Exclusions, 2003.

	Episodes		Claim Records		Payments	
	Number	Percent	Number	Percent	Total	Percent
Episodes 1-560	9,011,921	100.0	67,587,729	100.0	$8,649,500,603	100.0
Exclusions	735,523	8.2	4,118,416	6.1	1,438,949,909	16.6
Low Outliers	630,502	7.0	1,181,317	1.7	$10,796,370	0.1
High Outliers	89,441	1.0	2,907,054	4.3	$1,428,153,539	16.5
Missing Pmt.	15,580	0.2	30,045	0.0	Unknown	0.0
Study Episodes	8,276,398	91.8	63,469,313	93.9	$7,210,550,694	83.4

Physician Attribution Methodology

Episodes of care methodologies present a challenge unique to medical claims-based case mix and risk adjustment methodologies. In contrast to other methodologies, episodes of care may include claims for more than one provider. Physician attribution strategies have been developed to link physicians to episodes in order to profile their relative cost efficiencies. In this study, the physician assigned responsibility was the one who billed the plurality and at least 35 percent of payments for the Evaluation and Management (E&M) claims included in the episode.

In using this methodology, not all episodes can be attributed to a physician. This may result from episodes lacking E&M claims or from episodes for which all physicians failed to meet the 35 percent threshold. For 2002 and 2003, respectively, Table 14 and Table 15 present the number and percent of episodes, claims and payments compared to 1) the original complete set of episodes, and 2) the set trimmed of outlier episodes.

Table 14: Episodes Attributed to Physicians, 2002

	Episodes		Claim Records		Payments	
	Number	Percent	Number	Percent	Total	Percent
Complete Episodes	8,455,433	100.0	62,709,123	100.0	$7,662,107,877	100.0
Trimmed of Outliers	7,792,423	92.2	58,891,818	93.9	$6,433,698,897	84.0
Attributed Episodes	6,089,729	72.0	47,587,517	75.9	$5,016,903,464	65.5

Table 15: Episodes Attributed to Physicians, 2003

	Episodes		Claim Records		Payments	
	Number	Percent	Number	Percent	Total	Percent
Complete Episodes	9,011,921	100.0	67,587,729	100.0	$8,649,500,603	100.0
Trimmed of Outliers	8,276,398	91.8	63,469,313	93.9	$7,210,550,694	83.4
Attributed Episodes	6,440,590	77.8	50,651,067	79.8	$5,496,843,387	76.2

REFERENCES

Ash AS, Ellis RP, Pope GC, Ayanian JZ, Bates DW, Burstin H, Iezzoni LI, MacKay E and Yu w. Using diagnoses to describe populations and predict costs, *Health Care Financing Review*, 21(3), 7-27, 2000.

Dubois RW, Brook RH, Rogers, WH. Adjusted hospital death rates: Potential screen for quality of medical care. *American Journal of Public Health,* 77: 1162-1167, 1987.

Ellis, RP and Ash A, Refinements to Diagnostic Cost Group (DCG) Model, *Inquiry*, 32(4) 418-29, 1995.

Epstein, A. Performance reports on quality-prototypes, problems and prospects. *New England Journal of Medicine,* 333: 57-61, 1995.

Goldstein H, Spiegelhalter D. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*, Series A, 159: 385-444, 1996.

Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J. Interpreting hospital mortality data, *Journal of the American Medical Association*, 260: 3611-3616, 1988.

Leyland AH, Boddy FA. League tables and acute myocardial infarction, *Lancet*, 351: 555-558, 1998.

Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. NY, NY: Chapman & Hall, 2007.

Marshall EC, Spiegelhalter DJ. “Institutional Performance.” In Goldstein H, Leyland AH (eds.). *Multilevel Modelling of Health Statistics*, NY: Wiley, 2001.

Medicare Payment Advisory Commission. Report to the Congress: Increasing the value of Medicare. Washington, DC: MedPAC, 2006.

Morris CN, Christiansen CL. “Hierarchical Models for Ranking and for Identifying Extremes, with Application.” In Bernardo JO, Berger Dawid AP (eds.), *Bayesian Statistics 5*, Oxford University Press, 1996: 277-297.

Noreen EW. *Computer Intensive Methods for Testing Hypotheses.* NY, NY: John Wiley & Sons, 1989.

Normand SL, Glickman ME, Ryan TJ. “Modelling Mortality Rates for Elderly Heart Attack Patients: Profiling Hospitals.” In Gatsonis C, et al. (eds.). *The Cooperative Cardiovascular Project. Case Studies in Bayesian Statistics.* NY: Springer-Verlag, 1995: 435-456.

Normand SL, Glickman ME, Gatsonis CA. Statistical models for profiling physicians of medical care: Issues and applications. *Journal of the American Statistical Association*, 92: 803-814, 1997.

Rice N, Leyland A. Multilevel models: Applications to health data. *Journal of Health Services Research and Policy,* 1(3): 154-164, 1996.

Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital specific mortality rates. *Statistics in Medicine*, 13: 889-903, 1994.