CHAPTER 1

Using episode groupers to assess physician resource use

CHAPTER 1

Using episode groupers to assess physician resource use

Chapter summary

Physicians are central to the delivery of all types of health care. Research on variation in the use of services in Medicare implies that physicians and other providers may not always be directing resources efficiently. Some service use may not lead to higher quality care and could be unnecessary. As both expenditures and volume of services continue their steep climb, physicians are also central to efforts to use Medicare resources as efficiently as possible.

The Commission recommended in March 2005 that CMS use Medicare claims data to measure fee-for-service physicians' resource use and to provide individual physicians with confidential information on their resource use relative to their peers. Many private health plans already measure and compare physicians' resource use using episode groupers, which group claims into clinically distinct episodes adjusted for patient severity. Plans also share this information with physicians in their networks. The Commission is exploring the use of episode groupers on Medicare claims to better understand how these tools might work for Medicare. Because efficiency is defined both by resource use and

In this chapter

quality, we are also examining quality indicators in our analyses. In this chapter, we describe our findings from applying two episode grouper tools to a nationally representative, randomly selected 5 percent sample of Medicare claims. In general, the analysis shows that it is possible to use these types of episode groupers as one tool to measure physician resource use at the aggregate metropolitan statistical area (MSA) level, but some technical and analytic issues will need to be addressed as Medicare considers using these groupers to understand physician resource use.

Differences between the groupers. The two groupers we used—Episode Treatment Groups (ETGs) and Medstat Episode Groups (MEGs)—differ in their logic but also have some similarities. Both groupers use diagnosis codes to begin episodes and to assign claims to those episodes. In addition, both acknowledge disease severity and complexity when creating episodes. ETGs use the presence of specific procedures or comorbidities to further classify episodes. MEGs allow a single episode type to be broken into three stages based on the progression of the condition. Regardless of their differences, the groupers agree on the number of episodes created for most conditions. The ETG grouper was able to assign 90 percent of claims to 24 million episodes. These claims accounted for 94 percent of total dollars. The MEG grouper was able to assign 80 percent of the same sample of claims to 30 million episodes. These claims accounted for 96 percent of total dollars.

Risk adjustment. We discuss how risk-adjustment techniques can be used to further adjust episodes for overall patient severity. Risk adjustment can be used to try to avoid characterizing physicians who predominantly treat sicker patients as high resource use physicians compared with those who treat healthier patients.

Attribution. To apply relative resource use tools at the individual physician level, one must identify a method for attributing episodes and performance on quality indicators to individual physicians. This is not necessarily the same as identifying the physician who was actually directing care, but

is instead a statistical analysis to identify the physician responsible for providing most of the services furnished to a given beneficiary. We found that the vast majority of episodes can be attributed to a physician using either evaluation and management (E&M) spending or visits. About 90 percent of our selected episodes could be assigned to one physician who billed Medicare for 30 percent of E&M spending in an episode. We also tested multiple attribution rules. We found that only 11 percent of episodes had multiple physicians providing at least 35 percent of care (measured by spending on E&M services).

The vast majority of quality indicators can also be attributed to one physician using attribution rules based on either E&M visits or dollars. About 93 percent of our quality indicators could be assigned to one physician who billed Medicare for 35 percent or more of E&M visits associated with an indicator. We also tested multiple attribution rules. We found that only 10 percent of quality indicators had multiple physicians providing at least 35 percent of care, as measured by E&M visits.

Variation across MSAs. To better understand variation among different units of analysis, we applied the groupers and the quality indicators to 13 MSAs. We found that beneficiaries' use of resources and the quality of their care vary across MSAs, but the variance is greater for resource use. Among our selected conditions, we saw MSAs where the care provided was 35 percent less costly than the national average and MSAs where it was 41 percent more costly than the national average on a per episode basis. Quality scores ranged from 16 percent lower than the national average to 18 percent higher than the national average. Interestingly, we found that certain MSAs known for high resource use have low per episode costs for certain types of episodes. However, these MSAs have higher resource use when calculated on a per capita basis, partly because they tend to have more episodes per patient.

Performance on the quality indicators shows room for improvement nationally. The scores generally clustered around the national average or a

little above in the MSAs we studied. Performance on different conditions also varies within MSAs. In addition, our quality analysis shows that some indicators may not be useful because of limited sample size or occurrence. Further, how indicators are weighted affects relative MSA scores and rankings. In summary, we find that using claims-based quality indicators is possible, but a broader set less related to process measures is desirable.

The second step of our analysis (using a 100 percent sample of Medicare claims in several geographic regions) will provide information on the feasibility of applying episode groupers and quality indicators at the individual physician level. Among the issues we will address in the upcoming research are the minimum number of episodes or quality indicators needed to evaluate a physician's performance and the application of risk-adjustment techniques. ■

Background

Physicians are central to the delivery of all types of health care. Medicare beneficiaries rely on them to diagnose their health conditions and to recommend the timing and type of services they need. As both expenditures and volume of services continue their steep climb, physicians are also central to efforts to use Medicare resources as efficiently as possible.

Research on variation in the use of services in Medicare implies that physicians and other providers may not always be directing resources efficiently. Fisher and colleagues (2003) found significant regional variation in the amount and type of services beneficiaries receive with no discernible difference in quality of care. Some service use may not lead to higher quality care and could be unnecessary.

Over the last few years, the Commission has sought a greater understanding of the interaction between resource use and quality of care, which together define efficiency. We have identified quality measures for various providers and have recommended their use to distinguish among providers for payment purposes (MedPAC 2005). We have also explored resource use and found that the private sector is using a variety of tools to assess resource use by physicians and other providers. Plans use this type of information to provide confidential feedback, build tiered networks (with lower copayments for patients who see more efficient physicians), and create payment incentives (with higher payments for more efficient physicians, and vice versa). At times, information on resource use is used along with information on quality. The goal is generally to decrease the costs of care, while maintaining or improving quality.

The Commission recommended in March 2005 that CMS use Medicare claims data to measure physician resource use and share the results with physicians confidentially to educate them about how they compare with their peers.

Episodes of care emerged as a concept in the mid-1980s in studies that observed that health care is typically provided in a series of separate but related services, and that all of these services should be included in a comprehensive analysis of health care delivery. An episode of care comprises a series of clinically related health care claims over a defined time period, such as all claims related to a patient's diabetes (Hornbrook et al. 1985). Episodes can comprise all types of health care claims: inpatient admissions, physician visits, other outpatient services, and prescription drugs. Patients can have multiple episodes at any given time, such as concurrent diabetes and pneumonia episodes. In recent years, commercial software packages have emerged that comb through administrative claims data using clinical algorithms to create episodes of care.

We describe our findings from applying two of these grouper tools—Episode Treatment Groups (ETGs), developed by Symmetry Health Data Systems, and Medstat Episode Groups (MEGs), developed by Thomson Medstat—to a nationally representative, randomly selected 5 percent sample of Medicare claims.[1] We also examine a set of claims-based indicators developed by the Commission—the Medicare Ambulatory Care Indicators for the Elderly (MACIEs)—using the same 5 percent sample of claims (MedPAC 2006).

We applied the ETG and MEG groupers throughout the analysis to a 5 percent sample of Medicare claims. However, for simplicity some of the findings will be presented for MEGs only. We also describe our findings from the application of the MACIEs. Our discussion focuses on:

- ***Broad findings***—The number of episodes created, the total and average resource use by episode, and variation in resource use for each of our selected episodes.
- ***Addressing episode severity and complexity***—The grouping logic, risk-adjustment mechanisms that sort patients into episodes, and relative costs and types of services (e.g., inpatient, evaluation and management, and post-acute services).
- ***Attribution***—The ability to attribute episodes and quality indicators to individual physicians using different assumptions and the type of specialists to which episodes and quality indicators are attributed.
- ***Resource use and quality performance by region***—The variation in resource use and quality across selected MSAs and the drivers of that variation.

Future research will use the lessons learned from these analyses to measure and evaluate resource use for individual physicians using a 100 percent sample of Medicare claims in six selected metropolitan statistical areas (MSAs). Using all claims for an area will allow us to construct physician caseloads and to determine the

Construction of episodes

Not all episodes are suitable for further analysis. Both of the groupers we used—Episode Treatment Groups and Medstat Episode Groups—rely on the use of a clean period to determine when an episode is started and finished. If an individual does not have a claim related to a specific episode for a specific period of time, then the episode is closed and considered to have a clean finish. These clean periods can vary in length from 30 days for certain acute episodes to 365 days for chronic conditions.

Strictly speaking, chronic conditions by their very nature have no clean period, but for analytic purposes, both groupers close chronic episodes after 365 days.

Complete episodes are identified by looking both backwards and forwards around the window of the claims under analysis. In this analysis, we look at claims six months prior and subsequent to our analysis window to identify complete episodes. If an episode did not have a clean start or a clean finish, it was deleted from our analysis.

It is also important to note that episodes for a specific beneficiary can overlap. For example, a beneficiary with a 365-day chronic episode of coronary artery disease can also have a shorter episode of sinusitis within that same time frame. However, the costs from any given claim can only be assigned to one episode. Clinical algorithms decide to which episode claims will be assigned, particularly for patients with multiple chronic conditions. ■

feasibility of developing overall physician-level indicators of resource use and quality of care.

Data and methods

We analyzed a 5 percent sample of Medicare claims data for calendar years 2001, 2002, and 2003 with both groupers. We used claims from the hospital inpatient, hospital outpatient, skilled nursing facility (SNF), home health, and physician/supplier (including laboratory claims) sectors.[2]

We do not have any prescription drug claims because this analysis is focused on a period prior to the 2006 implementation of Medicare's prescription drug benefit. Prescription drug costs are an important component of total costs for many episodes, and the lack of these data may limit our ability to see the whole picture of physician performance. However, recent research found that even without pharmacy claims, analysis can draw valid conclusions about performance for some conditions (Thomas 2006a).

Of the 204 million claims processed, the ETG grouper was able to assign 184 million (90 percent) of the claims to 24 million episodes. These claims accounted for 94 percent of Medicare payments for the sampled claims. By contrast, the MEG grouper was able to assign 163 million (80 percent) of the same claims to 30 million episodes. These claims accounted for 96 percent of Medicare payments for the sampled claims. We saw no single identifiable pattern among the claims that were not grouped to episodes, although many of them were made up of ancillary services from the physician and outpatient files. Home health records were less likely to be grouped into episodes by either grouper.

We selected only clean episodes (the text box describes our methods in more detail) and deleted any outlier episodes with unusually high or low values to minimize any potential bias in our results. We tested a variety of approaches for trimming outliers. For this analysis we chose to delete the top and bottom percentile of each episode based on total payments—any episode for which total payments were greater than the 99th percentile or less than the 1st percentile. We also deleted any episode for which total payments were less than $30.[3] We chose this method because it removed extremely high and low outliers while not reducing sample size excessively.

Standardized payments

We standardized payments to help compare the resources in each episode. Standardizing excludes variation in resource costs due to geographic differences in input costs or policy considerations (e.g., teaching payments).

For example, Medicare will pay a community hospital in a rural area that is discharging a patient treated for a stroke less than a major teaching hospital in an urban area because of differences in the wage index, disproportionate share, and indirect and direct graduate medical education (IME/GME) payments. For this analysis, we want a hospital admission for stroke in all areas to have the same payment rate. We then can focus on the underlying differences in resource use due to utilization rates and practice patterns.

Payment rates in some settings were easier to standardize than in others. For inpatient prospective payment system (PPS) hospitals, it was a relatively straightforward task to link each diagnosis related group (DRG) to the appropriate standardized base payment and then multiply by the weight for that DRG. For physician claims, we matched the line item on the physician claim to the physician fee schedule relative value file and multiplied by a conversion factor.

We also developed ways of addressing differences in payment formulas that are more difficult to standardize. Among these are laboratory payments and payments under the SNF PPS. For a more detailed description of how payments were standardized, see the text box on page 25.

Selected conditions

Because each grouper classifies claims into more than 500 discrete episodes, assessing resource use using all episodes from both groupers would be overwhelming. For this exploratory analysis, we focused on a subset of conditions particularly relevant to the Medicare population, including both acute and chronic conditions. Using both groupers, we chose the subset of conditions based on their prevalence, total and per beneficiary resource use, variation in resource use within episodes, and the availability of MACIE quality indicators.

There is considerable agreement among ETGs and MEGs on the most prevalent and costly conditions among the Medicare population: coronary artery disease (CAD), hypertension, diabetes, chronic obstructive pulmonary disease, stroke, and congestive heart failure (CHF). Both groupers assign most of these conditions to the top decile for total resource use and prevalence. The results of the two groupers overlap less for conditions that occur less frequently. Though we have selected a mix of chronic and acute conditions, the labels chronic and acute can be misleading. Many patients with chronic conditions can have acute events. Patients with CAD have heart attacks and patients with cerebral vascular disease have strokes. There is also overlap among episodes. Patients with year-long CHF episodes can have episodes for other conditions in the same time period.

We also determined whether the MACIEs include quality measures for these episode types. We often have quality measures for chronic conditions. The text box on p. 10 describes our quality measures in greater detail. Some episodes have no corresponding quality measures, and other episodes need to be combined to be compared to quality indicators for a specific condition. For example, the MACIEs make no distinction between type 1 and type 2 diabetes, yet the ETG and MEG groupers sort patients into different episodes on this basis.

Results

In this section, we describe the results of applying the groupers to the selected conditions for several analytic purposes. To provide a context for the results of the analysis, we first describe how the groupers differ in how they assign claims to episodes. We then describe how the groupers address patient severity and complexity. For simplicity we describe most findings solely from MEGs, but we did conduct parallel analyses using ETGs.

As we used two different groupers on the same set of claims, we could see whether they grouped claims in similar or different ways. This comparison allows us to begin to identify differences and similarities in grouper logic. Both groupers use diagnosis codes to begin episodes and assign claims to those episodes. Both groupers distinguish clinical conditions in terms of patient severity and complexity. ETGs use the presence of specific procedures or comorbidities to create distinct episodes. MEGs allow selected episodes to be broken into three stages based on the progression of the condition. In some cases, the types of episodes created by each grouper for the same disease differ. For example, MEGs have two diabetes episodes, one each for type 1 and type 2 diabetes. However, ETGs have six episodes for diabetes: types 1 and 2 diabetes with and without comorbidities and diabetic retinopathy with and without comorbidities. ETGs have up to 15 different episodes for CAD, depending on the progression of the disease and the procedures performed, whereas MEGs have a single episode. MEGs, however, address severity by stage of disease, such as breaking a

Medicare Ambulatory Care Indicators for the Elderly

For the physician resource use analysis, we used the Medicare Ambulatory Care Indicators for the Elderly (MACIEs) to measure quality. The MACIEs were developed to assist in analyzing Medicare quality and access using measures that were clinically meaningful and could feasibly be analyzed from claims data. In May 2004, the Commission convened an expert panel of physicians, clinicians, and researchers to review and update the original set. The experts reviewed clinical evidence from existing guidelines, other organizations' efforts using ambulatory indicators, and the limits of claims data. The indicators were revised to reflect this review.

The MACIEs are designed to reflect basic clinical standards of care for common medical diagnoses. They focus on two types of measures: the percentage of beneficiaries who 1) receive necessary services for their diagnoses and 2) have potentially avoidable hospitalizations directly related to their diagnoses.

Necessary services are defined as routine care that has benefits that outweigh risk, that has benefits that are likely and substantial, and that physicians have judged improper not to recommend. Measures of potentially avoidable hospitalizations include use of emergency department services and inpatient hospitalizations that might have been averted had patients received better ambulatory care.

For the MACIEs, we selected medical conditions that are prevalent among the elderly population, have effective medical treatment available, and are readily identifiable from diagnoses codes on Medicare claims. Our physician resource use analysis used the MACIE set to measure the quality of care for beneficiaries with breast cancer, colon cancer, coronary artery disease, diabetes, congestive heart failure, depression, hypertension, chronic obstructive pulmonary disease, and stroke.

The MACIEs reflect minimum standards of acceptable care for certain diagnoses. For example, they include lipid testing for people with coronary artery disease. The MACIEs are not intended to show optimal care and can not account for reasons why patients do not receive necessary care. Needed services may not be provided for a number of reasons, including problems accessing the health care system, failure of providers to perform or recommend services, or failure of beneficiaries to follow provider recommendations.

The MACIE data analysis requires two years of claims data for each beneficiary cohort in order to check for service use within a specified amount of time (e.g., eye exam within a two-year period for diabetics). Therefore, the data set is restricted to the population of beneficiaries who were continuously in Medicare fee-for-service during the two-year study period. Beneficiaries were excluded from the data set if they died, were newly enrolled in Medicare, used hospice care, or were in managed care plans during the study period. Beneficiaries younger than age 65 were also excluded from the sample. For purposes of this chapter, we tracked these quality indicators at the national and the metropolitan statistical area levels. ■

CAD episode into three stages depending on the severity of the condition.[4]

For certain conditions the two groupers agree fairly well on the number of episodes created. For example, ETGs create 74,045 CHF episodes and MEGs create 78,124 CHF episodes with the sample of claims in our analysis (Table 1-1). Similarly, ETGs create 458,212 hypertension episodes compared with 415,151 created by MEGs. The largest difference among our selected conditions is for urinary tract infections in which ETGs create 137,684 episodes, 29 percent more than the 106,900 created by MEGs.

The two groupers also vary on the number of dollars they allocate to these episodes. In general, ETGs have higher (sometimes substantially higher) average total payments per episode. Average per episode payments for CHF in ETGs ($3,161) are more than twice those for MEGs ($1,394). Similarly, average per episode payments for

TABLE 1–1 **Medicare claims classified into MEGs compared with ETGs**

Condition	MEGs			ETGs		
	Number of episodes	Total spending (in millions)	Average spending	Number of episodes	Total spending (in millions)	Average spending
Coronary artery disease	201,936	$622	$3,079	233,673	$934	$3,998
Bacterial pneumonia	74,890	332	4,427	68,704	210	3,054
Cerebrovascular disease	107,561	295	2,743	90,630	255	2,811
Essential hypertension	415,151	175	423	458,212	215	469
Congestive heart failure	78,124	109	1,394	74,045	234	3,161
Urinary tract infections	106,900	89	830	137,684	72	521
Diabetes:						
Type 1	34,196	28	833	54,348	98	1,798
Type 2	157,337	83	526	142,106	90	636
Cholecystitis and cholelithiasis	16,959	77	4,549	19,970	85	4,261
Cancer:						
Prostate	47,211	67	1,410	34,484	80	2,330
Breast	23,421	55	2,341	28,753	69	2,400
Peptic ulcer disease	19,896	34	1,724	23,870	53	2,216
Sinusitis	78,520	12	158	90,827	21	230

Note: MEG (Medstat Episode Group), ETG (Episode Treatment Group).

Source: MedPAC analysis of 5 percent sample of Medicare claims using ETG and MEG episode groupers.

type 1 diabetes in ETGs ($1,798) are more than two times that for MEGs ($833).

Other conditions have more similar average per episode payments. For example, there is relatively little difference in average per episode payments for breast cancer ($2,400 for ETGs vs. $2,341 for MEGs) and for cerebrovascular disease ($2,811 for ETGs vs. $2,743 for MEGs).

These differences warrant further investigation. We should expect some differences between the groupers as each uses proprietary clinical algorithms to assign claims to episodes. The clinical logic of the groupers often drives the differences. For example, the MEG grouper treats CHF not as a disease but as a symptom of many different diseases. The ETG grouper retains CHF in a distinct episode and does not distribute CHF services to other episodes. In the MEG grouper, CHF is classified as a comorbidity in 40 other episodes; about 20 percent of all patients with CHF were found in these other episodes. The fact that these costs are distributed among a variety of episode types by MEGs could be one of the reasons that payments for MEG CHF episodes are lower on average than those for ETG CHF episodes. Medstat has found that average costs for CHF patients in non-CHF episodes are higher than for those in CHF episodes, and that these episodes tend to have higher proportions of hospital inpatient dollars than CHF episodes. We found a similar result when we examined costs by type of service within CHF episodes in the ETG and MEG groupers: ETG CHF episodes had a higher proportion of dollars attributable to hospital inpatient stays than MEG CHF episodes.

We will continue to examine differences between the two groupers in more detail for the 100 percent analysis. In addition, CMS is analyzing at least two of the available groupers to determine their characteristics when applied to the Medicare population. The focus of its analysis is on how the groupers work on this population, recognizing that many beneficiaries have multiple conditions that overlap in time. CMS will be analyzing, in detail, differences among the groupers in both output and clinical logic for selected important conditions.[5]

Addressing episode severity and overall patient complexity

An important question about episode groupers is whether they account for the underlying health status of beneficiaries. Some researchers and physicians are

TABLE 1–2

Variation in number and distribution of selected episodes, 2002

		Episodes		Payments			
Episode group	**Stage**	**Number**	**Percent**	**Total (in millions)**	**Percent of total**	**Average**	**CV**
Coronary artery disease	1	134,501	67%	$140	22%	$1,037	262
	2	28,354	14	152	24	5,361	143
	3	39,081	19	330	53	8,450	109
	Total	201,936	100	622	100	3,079	202
Essential hypertension	1	306,789	74	84	48	273	173
	2	55,757	13	24	14	426	154
	3	52,605	13	68	39	1,292	146
	Total	415,151	100	176	100	423	210
Colon cancer	0	1,764	13	1	1	648	247
	1	248	2	0	0	1,441	228
	2	9,714	70	47	57	4,874	155
	3	2,192	16	35	41	15,750	62
	Total	13,918	100	83	100	5,990	145
Type 1 diabetes	1	20,015	59	8	27	377	275
	2	11,883	35	14	48	1,140	219
	3	2,298	7	7	26	3,213	127
	Total	34,196	100	29	100	833	253
Type 2 diabetes	1	101,751	65	38	46	371	156
	2	48,374	31	33	40	691	150
	3	7,212	5	12	14	1,612	129
	Total	157,337	100	83	100	526	173
Peptic ulcer disease	1	10,859	55	6	17	532	229
	2	8,621	43	25	74	2,946	129
	3	416	2	3	9	7,541	74
	Total	19,896	100	34	100	1,724	182

Note: CV (coefficient of variation). Outlier episodes have been removed. All episodes in the table started in 2002. Values may not sum to total due to rounding.

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

concerned that differences in health status among patients may influence treatment costs within episodes, and that the average health status of patients may differ among physicians (Thomas 2006b). Without adjusting for risk, physicians who care for less severely ill patients may look more efficient than those who care for more severely ill patients.

The groupers we used have two separate mechanisms for addressing patient severity. First, they use severity in the logic used to group claims. Second, each has developed a method to risk adjust the episode information based on the overall health status of the beneficiary.

Grouping logic

The logic embedded in ETGs and MEGs is designed to create episodes that make it possible to compare similar patients. ETGs create separate types of episodes for patients with the same underlying condition by distinguishing among patients with procedures or comorbidities that might affect the costs of their care

for that specific episode. For example, prostate cancer episodes are divided into those with and without surgery (Symmetry 2005). MEGs generally use a single episode per condition but incorporate a concept known as disease staging that addresses the severity of that condition. The disease staging logic employed by the MEG grouper subdivides most episodes into three different disease severity categories.[6] This allows us to examine the amount of resources directed toward treating a condition as it progresses. Not surprisingly, as a disease becomes more severe the treatment costs increase (except for the late stages of some diseases when reduced treatment options may result in decreased costs). This is consistent with other research showing that a small percentage of seriously ill patients account for a disproportionate percentage of Medicare spending (MedPAC 2004). For example, while stage 3 CAD episodes account for only 19 percent of all such episodes, they account for 53 percent of all CAD episode dollars (Table 1-2). Similarly, stage 3 colon cancer episodes account for only 16 percent of episodes, but 41 percent of episode dollars.

Examining the amount of variation within each stage of an episode can provide an indication of variation in practice patterns. Stage 3 episodes, while being more expensive on average, exhibit less variation than stage 1 and stage 2 episodes. The coefficient of variation (CV) for stage 3 CAD is 109 percent compared with 262 percent for stage 1 CAD. Similarly, the CV for stage 3 peptic ulcer disease is less than one-third of the variation of stage 1 (74 percent vs. 229 percent). This could be because care for sicker patients, while more resource intensive, is potentially more clearly defined than care for individuals with less severe manifestations of the condition. However, the difference in variation among episode stages is less pronounced for some chronic conditions, such as type 2 diabetes and hypertension.

Risk adjustment

Both the ETG grouper and the MEG grouper have additional capabilities that can be used to risk adjust episodes.[7] Risk adjustment can account for differences in health status that go beyond a particular disease that the episode grouper is trying to capture. The ETG software uses a companion product known as episode risk groups, which employs the same underlying methodology as ETGs. The software classifies a patient by episode and then looks at a person's age, gender, and mix of episodes to create a clinical and demographic risk profile. Using this risk profile, the software computes both a retrospective and prospective risk score for each person.

The MEG grouper employs the diagnostic cost group (DCG) method, which uses the conditions and diseases for which a person receives treatment over a specified period of time (usually one year), and the person's age and gender. The model estimates the level of expected cost in a given year as a function of medical problems treated in that year and creates a relative risk score (RRS) (Thomson Medstat 2005). Combined with the disease staging approach, researchers can segment episodes according to both episode severity and patient complexity (sample size permitting).

The DCG/RRS approach can further refine comparisons within and across episodes. Ultimately, it also allows for the construction of an overall risk score for a physician's caseload of patients. As described earlier, CAD episodes have overall average resource use of $3,079 (Table 1-2). However, using the DCG/RRS risk-adjustment technique, each CAD episode stage can be further subdivided into five categories of overall patient complexity, ranging from 1 (low complexity) to 5 (high complexity). Average resource use increases as patient complexity increases. Average resource use for stage 1 CAD episodes with a relative risk score of 1 is $564, while average resource use for stage 3 CAD episodes with a relative risk score of 5 is $11,509 (Table 1-3, p. 14).

Types of services

To learn more about what drives the overall resource use in these episodes and how resource use may vary by the severity of the condition, we divided all payments associated with an episode into types of services (Table 1-4, p. 15).[8] For example, 64 percent of total CAD episode dollars are spent in an inpatient setting compared with only 4 percent for sinusitis episodes and 35 percent for CHF episodes. For those with stage 1 CAD, dollars are far more concentrated in imaging than they are in the two higher stages of severity. Table 1-4 (p. 15) also shows how important evaluation and management (E&M) services are to beneficiaries with chronic conditions such as diabetes, hypertension, and sinusitis. For example, E&M services for beneficiaries with stage 1 diabetes (types 1 and 2) represent 52 and 62 percent of the spending for their episodes, respectively.[9]

TABLE 1–3

Average episode costs rise with disease severity and overall patient complexity, 2002

		Complexity level				
	Severity stage	**1**	**2**	**3**	**4**	**5**
Congestive heart failure	3	$1,002	$1,351	$1,611	$1,895	$2,070
Coronary artery disease	1	564	1,312	1,833	1,944	1,944
	2	3,051	7,028	8,072	8,484	8,484
	3	5,067	7,860	10,596	11,509	11,509

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

Attributing episodes to physicians

One of the main goals of grouping claims into episodes is to attribute episodes to physicians and to ultimately identify efficient physicians.[10] In this analysis we do not have a full picture of any one physician's patient panel; therefore, we can not assess any one individual physician's resource use. Nonetheless, in preparation for future analyses that will measure individual physicians, we applied a variety of attribution rules to this database. We looked at how many episodes could be linked to physicians using various attribution methods. The goal is to identify a single physician who provides a significant portion of a beneficiary's care for a given episode.

Some plans, such as health maintenance organizations (HMOs), formally assign patients to a provider so attribution is relatively straightforward. However, in other private plan arrangements and the Medicare fee-for-service (FFS) program, patients have greater freedom to see any physician. This makes attribution less straightforward. Identifying an individual physician who sees the patient for a significant portion of his or her care has to be determined by patterns in claims data. For a discussion of attribution issues, see text box on page 16.

Attribution to physicians is also necessary to determine performance on quality indicators. We determined quality scores using an attribution logic independent of the logic used to associate episodes to physicians. It is not always possible to attribute quality indicators with specific episodes of care. Separate attribution is also consistent with private sector efforts to measure resource use and quality. Typically, resource use episodes are created and attributed to physicians, then physicians are compared to a peer group. At the same time, quality scores are created and attributed to physicians, then the scores are compared to a peer group. While we would expect overlap between the patients attributed to a physician for resource use and quality, the quality indicators are not generally tied to specific episodes.

We attributed episodes to physicians using E&M codes to avoid giving too much weight to procedures or hospitalizations. Through claims, we can identify the number of visits beneficiaries had with a physician and the amount of dollars associated with those visits. We also looked at how our results shifted if episodes could be assigned to more than one physician.

Our main focus was to evaluate how many episodes we could assign to physicians using different attribution methods and the type of specialty to which they were assigned. A broader question is the appropriateness of using the same attribution method across episodes and quality indicators that may differ significantly from each other (either clinically or in terms of resource use). A uniform attribution approach may not fit all episodes.

Our analysis found that the key factor in attributing episodes to physicians is the threshold for attribution. Less important was the choice of dollars versus visits or whether to use all E&M services or just those furnished outside of a hospital.

While this analysis is primarily focused on technical approaches to attribution, further discussion is needed on

TABLE 1–4

Episode costs are driven by different mixes of services, 2002

Episode group	Stage	Average payment	Percentage of episode costs, by type of service						
			Inpatient	E&M	Post-acute care	Procedures	Imaging	Tests	Other
Coronary artery disease	1	$1,037	32%	20%	5%	8%	19%	7%	9%
	2	5,361	67	8	1	10	6	2	5
	3	8,450	76	7	3	7	3	1	2
	Total	3,079	64	10	3	8	8	3	4
Essential hypertension	1	273	8	68	4	2	5	13	1
	2	426	12	60	3	2	10	11	2
	3	1,292	60	24	4	2	4	4	3
	Total	423	28	50	4	2	5	9	2
Congestive heart failure	3	1,394	35	31	13	3	10	5	2
	Total	1,394	35	31	13	3	10	5	2
Type 1 diabetes	1	377	12	52	23	3	1	9	1
	2	1,140	55	21	8	9	1	4	3
	3	3,213	73	12	7	4	0	2	2
	Total	833	48	27	12	6	1	5	2
Type 2 diabetes	1	371	9	62	4	2	2	20	2
	2	691	21	48	3	8	1	15	5
	3	1,612	53	29	3	5	1	7	3
	Total	526	20	52	4	5	2	16	3
Sinusitis	1	153	4	64	0	15	9	5	3
	2	402	5	34	0	34	13	5	9
	3	428	15	36	0	26	15	4	4
	Total	158	4	62	0	16	10	5	3

Note: E&M (evaluation and management). Outlier episodes have been removed.

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

the extent to which accountability should be derived from these types of attribution rules. In a payment system as fragmented as Medicare FFS, a physician with 30 percent of the E&M visits in a given episode may not necessarily be aware of the kind of care being provided in the other E&M visits. For some episodes, meaningful accountability might rest with a single physician, while accountability for other episodes might rest with a team of physicians. In some instances, the cooperation of a hospital and its physicians may be important for efficiency.

Resource use attribution

We found that more episodes can be attributed to a physician as the thresholds for attribution are lowered. We also found that as the thresholds are lowered, any differences between the use of E&M dollars versus E&M visits to determine attribution largely disappear. Seventy-five percent of the episodes we selected could be attributed to a physician when using a threshold of 50 percent of E&M visits to identify the physician versus 82 percent of the episodes when we use a threshold of 50 percent of

Attribution issues

Some issues involved in determining the appropriate approach for assigning episodes include:

- *Is the attribution method conceptually valid?* It must be defensible and accepted by payers, providers, and other users of the information.
- *What is the appropriate unit of measurement?* Using dollars may more accurately reflect the intensity of services provided, whereas the number of visits might better identify the physician who had the greatest involvement in managing a patient's care.
- *What type of dollars or visits should be counted in determining attribution?* Should all dollars be counted in determining attribution? An approach like this might result in certain specialists and surgeons being assigned episodes for which they had relatively little control, but were attributed a majority of dollars due to the need for surgery or some other resource intensive intervention. Dollars associated with evaluation and management (E&M) visits might be a better indicator of physician involvement in an episode. Even if E&M dollars are used, should all physician E&M claims be used, including E&M claims that occur while a beneficiary is hospitalized?
- *What is the appropriate threshold of dollars or visits to use?* The higher a threshold is set the less likely it is to be assigned to a physician. Using a lower threshold might result in an episode being attributed to a provider who had less involvement in a patient's care.
- *What is the best attribution method?* Multiple attribution recognizes that more than one physician may have been involved in managing the resources and quality of care. However, such an approach may work more effectively when physicians work within structured networks. Different attribution methods may be needed for different types of episodes. ■

E&M dollars to identify the physician. Using a threshold of 30 percent of E&M dollars or E&M visits allows 90 percent of our selected episodes to be assigned to a provider. The share of episodes that can be attributed using a 30 percent threshold varies by condition—it ranges from a low of 71 percent for cerebrovascular disease episodes to a high of 96 percent for sinusitis episodes (Table 1-5).

To understand how hospitalizations affect episode attribution, we also restricted the analysis to dollars and visits associated with noninpatient settings. Using this approach, 86 percent of all episodes could be attributed to a physician who had 30 percent or more of the noninpatient E&M visits or noninpatient E&M dollars in the episode, compared to 90 percent of episodes if dollars and visits were not restricted to those that occurred in a noninpatient setting. The decrease in the number of attributed episodes is most pronounced in episodes that have a high proportion of inpatient dollars, such as peptic ulcer disease, depression, and choleocystitis (data not shown). The decrease is less pronounced for episodes that use a lower proportion of inpatient dollars, such as type 2 diabetes and essential hypertension (Table 1-5).

We also explored multiple attribution, or assigning an episode to more than one physician. We found that for most episodes a single physician tends to be dominant in the provision of care (Table 1-5). Using single attribution and a 35 percent threshold of all E&M visits, 88 percent of our selected episodes could be attributed to a physician. Using the same threshold but permitting an episode to be assigned to any physician meeting the threshold results in 78 percent of selected episodes still being assigned to a single physician. The remaining 10 percent of episodes could be assigned to more than one physician using a multiple attribution approach.[11] The proportion of episodes assigned to more than one physician ranged from a low of 7 percent for several conditions to a high of 11 percent for CAD episodes.

TABLE 1–5

Most episodes can be attributed to one physician, 2002

Attribution rule	Percentage of episodes						
	All selected episodes	CAD	Essential hypertension	Type 1 diabetes	Type 2 diabetes	Sinusitis	Cerebrovascular disease
Percentage of E&M visits:							
30%	90%	86%	94%	91%	95%	96%	71%
35%	88	83	93	90	93	95	68
40%	86	81	92	89	92	94	65
50%	75	69	85	81	81	88	53
Percentage of noninpatient E&M visits:							
30%	86	81	92	85	93	95	65
Percentage of E&M dollars:							
30%	90	86	94	91	96	96	71
35%	89	85	94	91	95	95	69
40%	87	82	93	90	93	95	66
50%	82	77	90	87	89	93	60
Multiple attribution based on visits:							
35%							
1 doctor	78	73	86	82	84	88	60
2 doctors	10	11	7	8	10	7	9
Total	88	84	93	90	94	95	69
Number of episodes	1,671,638	201,936	415,151	34,196	157,337	78,520	107,561

Note: E&M (evaluation and management), CAD (coronary artery disease).

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

We examined the types of specialties to which the episodes were attributed. In general, episodes were attributed to specialties one would expect to be associated with care for that condition. For example, 38 percent of CAD and 20 percent of CHF episodes are attributed to cardiologists, 64 percent of prostate cancer episodes are attributed to urologists, 21 percent of peptic ulcer disease episodes are attributed to gastroenterologists, and 39 percent of breast cancer episodes are assigned to oncologists using a threshold of 35 percent of all E&M dollars (data not shown).

Quality indicator attribution

We attribute responsibility for performance on condition-specific quality indicators to the physician most likely to be managing the beneficiary's care for that condition. We used a year of claims for the condition for which the beneficiary was eligible for the indicator (the denominator). For example, attributing care for beneficiaries with diabetes using a 35 percent threshold of E&M dollars means that we identified a single physician who billed the Medicare program for 35 percent or more of the E&M dollars for care of that beneficiary's diabetes.[12]

We find a similar ability to attribute quality indicators using these various methods as we do for resource use (Table 1-6, p. 18). The care on these quality indicators for the vast majority of beneficiaries can be attributed to a single physician. The lower the threshold, the more care can be attributed to one physician. We also find that using E&M visits instead of dollars results in a small increase in the ability to attribute care to a single physician.

TABLE 1–6

Most beneficiaries eligible for quality indicators can be attributed to one physician

	Percentage of beneficiaries with condition					
Attribution rule	**All indicators**	**CHF**	**CAD**	**Diabetes**	**COPD**	**Stroke**
Percentage of:						
E&M visits						
35%	93%	92%	92%	93%	95%	90%
50%	74	74	70	74	80	64
E&M dollars						
35%	89	89	88	89	92	85
50%	72	73	68	72	78	62

Note: CHF (congestive heart failure), CAD (coronary artery disease), COPD (chronic obstructive pulmonary disease), E&M (evaluation and management). The percentage of beneficiaries whose care was attributed for each condition is based on an average across all condition-specific indicators. The all indicators column is an average across all indicators, including several conditions not shown in the chart.

Source: MedPAC analysis using the Medicare Ambulatory Care Indicators for the Elderly on a 5 percent sample of 2002 Medicare claims.

As with the resource use analysis, shifting the percentage thresholds does shift the number of beneficiaries able to be attributed for each indicator. For example, we find that with a 35 percent threshold of visits, care for 93 percent of beneficiaries across the conditions can be attributed to a single physician. This percent decreases to 74 percent of beneficiaries when the threshold is raised to 50 percent of visits. The ability to identify a single physician who manages beneficiary care also varies within a set of condition-specific indicators. For example, across all eight CHF indicators the average percentage of beneficiaries able to be attributed to a single physician using 35 percent of E&M dollars is 89 percent. We are able to attribute six of eight indicators for CHF care to a single physician for 90 percent to 93 percent of the beneficiaries using 35 percent of E&M dollars (data not shown). However, we can only attribute 78 percent of the beneficiaries on the other two CHF indicators.

We examined whether using only noninpatient E&M visits would affect attribution. As was the case in our resource use analysis, we found that slightly fewer beneficiaries could be attributed to a single physician using noninpatient E&M claims (data not shown). We also explored multiple attribution and again found results similar to our resource use analysis. When we applied a multiple attribution rule of 35 percent of E&M visits across all quality indicators, only 10 percent of beneficiaries were attributed to more than one physician across our selected conditions.

As was the case for resource use, we found that care for the quality indicators tends to be attributed to specialties one would associate with the specific condition. We also found that the same type of specialties tend to be responsible for both resource use and quality for the condition. For example, the top three types of specialties to which quality of care for beneficiaries with CAD is attributed are cardiology, internal medicine, and family practice. These are the same as the top three types of specialties to which CAD episodes in the resource use analysis are attributed (data not shown).

Comparison of resource use and quality across MSAs

To better understand how relative resource use and quality vary across different units of analysis, we applied the groupers and the quality indicators to our 5 percent sample in 13 MSAs. We chose 13 large MSAs to achieve the widest geographic distribution possible. We observed variations in both resource use and quality by MSA and identified technical issues that may need to be addressed before we apply episode grouping approaches or quality indicators to individual physicians.

TABLE 1–7

Relative resource use ratios for selected MSAs, 2002

	CHF	CAD	Type 1 Diabetes	Type 2 Diabetes	Breast cancer	Hypertension	Pneumonia
National average episode costs	$1,394	$3,079	$833	$526	$2,341	$423	$4,427
MSA							
Boston	1.00	0.86	0.99	0.95	0.86	0.96	0.96
Chicago	1.05	1.04	1.13	1.17	1.06	1.14	1.11
Denver	0.91	1.03	1.04	0.96	1.24	0.92	1.04
Detroit	0.90	0.79	1.07	1.08	0.94	1.15	0.91
Greenville	0.91	1.24	1.21	0.91	1.02	0.87	0.80
Houston	1.16	1.04	1.13	1.16	1.02	1.20	1.11
Kansas	1.31	0.98	0.96	0.99	0.79	1.13	0.96
Miami	0.99	0.66	1.06	1.28	0.82	1.20	1.16
Minneapolis	1.00	1.28	0.72	0.88	1.14	0.87	0.76
New York	0.86	0.65	1.41	1.11	1.01	1.13	1.12
Orange County	1.01	0.76	1.17	1.31	0.98	1.00	1.03
Philadelphia	1.11	0.78	1.09	1.07	0.81	1.05	1.08
Phoenix	0.81	0.91	0.95	1.05	1.08	0.94	0.78

Note: MSA (metropolitan statistical area), CHF (congestive heart failure), CAD (coronary artery disease). Relative resource use scores for individual MSAs are calculated by dividing the MSA's average for a given episode by the national average for that episode. A score of more than 1.0 indicates higher-than-average episode costs and a score of less than 1.0 indicates lower-than-average episode costs.

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

Variation in resource use by MSA

Research shows that the amount and type of services beneficiaries receive vary significantly from region to region (Fisher et al. 2003). Looking at resource use from the perspective of an episode of care, will areas that are high resource use on a per capita basis also be high resource use on a per episode basis? It is important to note that episode groupers have rarely, if ever, been used to measure relative resource use across different regions. Because of the fragmented nature of health insurance coverage for the nonelderly population, many different health insurance companies can insure the population of any given area. In using episode groupers, most plans try to assess the relative resource use of physicians in their network, not resource use relative to other physicians in the same market area or different markets.

We examined resource use for a subset of episodes by calculating average per episode costs for the providers in each MSA and nationally.[13] We then calculated ratios of each MSA's average episode cost to the national average (Table 1-7). Resource use across episodes varied markedly with relative resource ratios ranging from 0.65 (lower than average) to 1.41 (higher than average), depending on the type of episode. At first glance the results seem similar to previous studies that have focused on per capita spending. MSAs that other research has shown to use more resources than average (e.g., Miami, Detroit, and Houston) have resource use ratios of more than 1 for many episodes. Others that use fewer resources than average (e.g., Minneapolis and Denver) have resource use ratios of less than 1 on many episodes.

Miami has higher-than-average resource use ratios for type 1 and type 2 diabetes (1.06 and 1.28, respectively), while Minneapolis has lower-than-average ratios (0.72 and 0.88, respectively). Similarly, Miami's relative resource use ratio for hypertension is 1.20 compared to 0.87 for Minneapolis.

However, for certain conditions the results are more surprising. Miami has a relative resource use ratio of 0.66 for CAD, while Minneapolis has a ratio of 1.28.

We conducted several additional analyses to better understand what was driving such a large difference in per episode resource use costs for CAD between the

TABLE 1–8 **Comparison of resource use in Miami and Minneapolis MSAs, 2002**

	Miami	Minneapolis
CAD per episode ratio	0.66	1.28
Total episodes per beneficiary with a CAD episode	11	8
Total episode dollars per beneficiary with a CAD episode	$11,700	$11,900
Per episode ratio (all episodes, all beneficiaries)	0.98	1.03
Per capita ratio (all episodes, all beneficiaries)	1.32	0.88

Note: MSA (metropolitan statistical area), CAD (coronary artery disease). Per episode ratios for individual MSAs are calculated by dividing the MSA's average for a given episode by the national average for that episode. Per episode ratios are calculated by dividing each MSA's average per episode amount by the national average per episode amount. Per capita ratios are calculated by dividing each MSA's average per capita amount by the national average per capita amount. A score of more than 1.0 indicates higher-than-average costs and a score of less than 1.0 indicates lower-than-average costs.

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

Minneapolis and Miami MSAs. The results of the analyses indicate that it may be helpful to combine per episode approaches to resource use with per capita approaches.

While beneficiaries in Miami have significantly lower per episode costs than beneficiaries in Minneapolis for CAD, they are also more likely to have more total episodes (including non-CAD episodes) than beneficiaries in Minneapolis (11 vs. 8) (Table 1-8). In particular, beneficiaries in Miami were more likely to be diagnosed and coded with other cardiovascular episodes such as varicose veins or tibial, iliac, femoral, or popliteal artery disease. Some of the costs of care for these beneficiaries may be captured in these other episodes in Miami, therefore lowering the costs of their concurrent CAD episodes.

Additionally, when we look at total spending across all episodes for beneficiaries with a CAD episode in each MSA, the difference between the two areas all but disappears. Per beneficiary Medicare spending across all episodes for beneficiaries with CAD is $11,700 in Miami compared to $11,900 in Minneapolis. This suggests that regional differences in coding and utilization may lead to situations where a beneficiary who might stay in the same CAD episode group in Minneapolis could be shifted to another heart-related episode group in Miami. Because of this, per episode relative resource use scores for CAD in Miami appear to show that physicians' care is very efficient, but care may not be as efficient when we look across all care for beneficiaries with CAD episodes.

Finally, when we compare the Miami and Minneapolis MSAs across all episodes both from a per episode and a per capita perspective, the results are dramatically different. On a per episode basis, Miami has a relative resource use score of 0.98 compared to 1.03 for Minneapolis across all episodes. However, on a per capita basis, Miami has a relative resource use score of 1.32 compared to 0.88 for Minneapolis (Table 1-8).

These results suggest that beneficiaries in Miami are being classified into more episodes of care than beneficiaries in Minneapolis. These variations in resource use scores also highlight that episode groupers can not judge the clinical appropriateness of any given service, just the efficiency of that service relative to similar services. Put another way, it is possible that in some areas with low resource use scores, beneficiaries are more likely to go to the doctor or receive certain services, whereas in other areas they might not go to the doctor at all. In addition, when a beneficiary sees a physician, the physician may be more likely to order more tests or treatment or to identify an additional diagnosis. This could lead to an MSA having a lower-than-average resource use score for a given condition if it has a large number of low-cost, low-intensity episodes that other areas might not have. Alternatively, physician coding practices may differ by region. To the extent that physicians code claims more extensively, patients may be classified into additional episodes.

Variation in type of service by MSA

The types of services that are used within episodes differ across MSAs. It is important to note that the overall composition of costs in any given episode will strongly influence any differences in relative resource use by type of service. For example, inpatient stays account for 64 percent of total CAD episode costs. Therefore any differences in inpatient resource use for CAD episodes will have a large impact on an MSA's overall relative resource use ratio (Table 1-4, p. 15). There are significant differences in the type of services used within each CAD episode in Miami and Minneapolis MSAs. Beneficiaries in Miami are significantly less likely to have hospital

inpatient costs, with a relative hospital inpatient resource use ratio of 0.51 compared to 1.46 for beneficiaries in Minneapolis. Conversely, Miami has an imaging relative resource use ratio of 1.62 compared to 0.76 for Minneapolis (Table 1-9).

Other episodes, such as hypertension and type 1 and type 2 diabetes, have a stronger E&M focus. Therefore, Miami, which has high relative resource use ratios for these episodes (1.20 for hypertension, 1.06 for type 1 diabetes, and 1.28 for type 2 diabetes) is above average for E&M resource use in all three episodes (1.37, 1.19, and 1.42, respectively).

MSA quality analysis

The Commission finds that MSA scores on quality tend to cluster around the national average or slightly above. However, the national averages are relatively low. The variation in the ratio of MSA quality scores to national quality scores across the MSAs is less than the variation on resource use. Further, we find that some MSAs did well on some indicators of necessary care but not others. No MSA is above average on all the conditions (Table 1-10, p. 23).

We also identify several important technical issues. Small sample size or low incidence rates make it difficult to use some indicators. We created composite scores in order to compare MSAs. The method for weighting each individual indicator when creating a composite condition score affects the relative rankings of the MSAs. The nature of the indicators, which rely on claims for services delivered, may limit their utility when trying to determine whether the resources used produce high-quality care.

Methods

Our quality measures consist of indicators of necessary care and potentially avoidable hospitalizations for certain conditions—the MACIE set. The text box on page 10 describes these indicators. For example, one indicator of necessary care for diabetics is whether the beneficiary received a hemoglobin A1C test. An indicator of a potentially avoidable hospitalization is whether the diabetic was admitted to a hospital for long-term complications related to diabetes.

We deleted from the analysis any indicators with a small sample size either across all MSAs or for a specific MSA. This step eliminated all the potentially avoidable hospitalizations from the composite condition scores. As

TABLE 1–9

Relative resource use, by type of service, Miami and Minneapolis MSAs

Episode group	National average	Miami	Minneapolis
CAD			
E&M	$307	1.05	0.93
Imaging	234	1.62	0.76
Inpatient	1,968	0.51	1.46
Other	3	1.25	0.92
Post-acute care	97	0.99	0.82
Procedures	248	0.39	1.24
Tests	86	1.14	0.90
Unclassified	136	0.33	1.07
Total	3,079	0.66	1.28
Hypertension			
E&M	210	1.37	0.80
Imaging	21	1.01	0.47
Inpatient	120	0.99	0.88
Other	1	2.78	0.29
Post-acute care	17	1.14	1.56
Procedures	7	0.91	0.73
Tests	39	1.17	1.01
Unclassified	8	0.71	1.38
Total	423	1.20	0.87
Type 1 diabetes			
E&M	223	1.19	0.85
Imaging	6	0.54	2.25
Inpatient	401	1.13	0.64
Other	2	2.85	1.23
Post-acute care	96	0.70	0.45
Procedures	49	0.52	0.72
Tests	38	0.92	0.86
Unclassified	17	1.55	1.45
Total	833	1.06	0.72
Type 2 diabetes			
E&M	271	1.42	0.91
Imaging	8	0.84	0.99
Inpatient	105	1.33	0.49
Other	3	1.69	3.36
Post-acute care	19	1.16	1.31
Procedures	25	0.74	0.80
Tests	84	1.06	1.00
Unclassified	12	0.96	1.46
Total	526	1.28	0.88

Note: MSA (metropolitan statistical area), CAD (coronary artery disease), E&M (evaluation and management). Relative resource use scores for individual MSAs are calculated by dividing the MSA's average for a given episode by the national average for that episode. A score of more than 1.0 indicates higher-than-average episode costs and a score of less than 1.0 indicates lower-than-average episode costs.

Source: MedPAC analysis of 5 percent sample of Medicare claims using Medstat Episode Group grouper.

a result, the composite score for each condition reflects the average of all the necessary care indicators for each condition for each MSA. To create the ratios found in Table 1-10, we then compared each MSA condition composite with the national average.

In general, the national averages show room for improvement across all conditions. With the exception of chronic obstructive pulmonary disease, 20 percent to 40 percent of beneficiaries are not receiving necessary care on our selected conditions. Relative to these national benchmarks by condition, quality appears to vary less across the MSAs than resource use. Table 1-10 shows that no MSA is more than 16 percent lower (breast cancer composite in New York) than the national breast cancer average or more than 18 percent higher (stroke composite in Orange County) than the national average on stroke quality. In contrast, resource use ratios range from 35 percent to 41 percent higher.

Some MSAs did well on some indicators and poorly on others. No MSA was above average on all the conditions. For example, we found that beneficiaries in Phoenix received necessary care for CHF at a rate 8 percent higher than the national average, but 5 percent lower than the national average for stroke patients (Table 1-10).

Sample size and incidence

Before we grouped the indicators by condition, we considered whether the individual MSA scores were stable enough to be included in our analysis. We found two issues.

First, because of their low incidence, we did not use potentially avoidable hospitalizations in our composite scores for the MSAs. Potentially avoidable hospitalizations specific to each condition may provide useful information at the national level and for population-based analyses, but occur too rarely to be useful with a 5 percent sample of claims. We will continue to analyze these indicators to consider ways they could be grouped to increase their incidence and will also revisit the issue when we are using a 100 percent sample. Another issue with potentially avoidable hospitalizations is that other factors besides physician management may affect whether a patient is hospitalized, and that multiple physicians may also be involved.

Second, we found that for some indicators the number of beneficiaries eligible for specific quality indicators was too low to be used in our composite MSA scores. We used a threshold of 30 eligible patients in any one MSA. That is, if fewer than 30 beneficiaries in any one MSA were eligible for the necessary service, we did not use the indicator in calculating the MSA score.

Composites and weighting

To use this list of quality indicators to compare MSAs, we grouped them into composite scores for each condition. Indicators of necessary ambulatory care are usually grouped by condition because they often apply to the same type of beneficiary. For some of the conditions, the denominators (those eligible for the service) are even the same.

To create composites, it is necessary to choose a method for combining the various indicators for each condition and to determine what weight to place on each. Different weighting methods place importance on different dimensions. Relative importance can be determined by the number of beneficiaries affected, or how important the indicator might be to the beneficiary. In addition, the level of evidence supporting the indicator and the precision of the measurement of the indicators are also factors to be considered. The cost-effectiveness of the intervention could also be considered (See Chapter 10 for discussion.)

We used two weighting methods when calculating composites. First, we created a straight average for each condition by adding the scores for each indicator in each condition and then dividing the sum by the number of indicators. This method, because it weights each indicator the same, does not account for the fact that some indicators affect more beneficiaries than others. For this reason, we also calculated the composites by weighting the indicators by the number of beneficiaries each affects. This "opportunity" model is based on the idea of measuring the number of opportunities that physicians have to provide necessary care; it is often used for quality indicators in the same condition (Nolan and Berwick 2006). For example, for beneficiaries with diabetes, we summed the number of beneficiaries that should receive certain necessary services (the denominators for all of the diabetes indicators) and the number of beneficiaries who actually received the necessary services (the numerator for all of the diabetes indicators) and then divided the sum of the numerators by the sum of the denominators.

When we use the opportunity approach to weighting, the national average scores for the conditions shift the most for breast cancer and stroke. This is because those

TABLE 1–10 **Ratio of MSA to national quality scores on six conditions, 2002**

	Breast cancer	CHF	CAD	Diabetes	COPD	Stroke
National average quality score	57%	73%	78%	71%	93%	68%
MSA						
Boston	0.97	1.04	0.97	1.01	1.02	0.93
Chicago	0.95	1.00	1.01	0.95	1.01	1.02
Denver	0.85	0.97	0.99	1.08	0.97	N/A
Detroit	1.01	1.05	1.06	0.98	1.02	1.01
Greenville	1.10	1.00	0.99	0.98	1.03	1.06
Houston	1.03	1.01	1.03	1.00	1.04	1.05
Kansas	0.93	1.07	1.01	1.02	1.02	0.99
Miami	0.93	1.07	1.10	1.05	1.05	1.00
Minneapolis	1.01	1.02	0.97	1.02	0.99	1.05
New York	0.84	1.01	1.08	1.02	1.03	1.07
Orange County	0.99	1.04	1.06	1.09	1.00	1.18
Philadelphia	0.97	1.07	1.01	1.02	1.00	0.96
Phoenix	1.14	1.08	1.05	1.06	1.01	0.95

Note: MSA (metropolitan statistical area), CHF (congestive heart failure), CAD (coronary artery disease), COPD (chronic obstructive pulmonary disease). National scores indicate the percentage of Medicare beneficiaries nationwide who are receiving necessary care for their conditions. MSA ratios are calculated by dividing the average MSA quality score for each condition by the national average for that condition. If the MSA ratio is above 1.0, the MSA score is above the national average. If the MSA ratio is below 1.0, the MSA score is below the national average.

Source: MedPAC analysis using the Medicare Ambulatory Care Indicators for the Elderly on a 5 percent sample of Medicare claims.

two conditions have at least one indicator with a very large denominator, indicating that a large number of beneficiaries should receive that service. Thus, the scores on those indicators dominate the analysis.[14] For the other four conditions, the numbers of eligible beneficiaries are similar so the national scores are less sensitive to the weighting method used.

However, using different weighting strategies changes some of the relative rankings of the various MSAs for all conditions to some degree. As would be expected, the greatest shift was for the two that moved the most at the national level. However, relative rankings of the MSAs also shifted for the other conditions, even when little change occurred at the national level.

Using a broader set of quality measures important

When looking at these MSA quality scores together with resource use scores, it is important to consider the types of quality measures used. Because both our analyses (resource use and quality) measure the services provided, it is difficult to assess quality independent of resource use. Further, as discussed in our March 2005 report, claims-based process measures, especially without lab values or prescription drug data, represent a limited picture of quality (MedPAC 2005).

Because our scores clustered so closely around the national average, it was hard to assess whether MSAs that used more resources also had higher quality scores—the relationship we would expect if our resource use and quality measures moved together. The vast majority of quality scores were very close to the national average, and MSA scores on resource use and quality varied based on the condition measured. We found that quality scores could be slightly better than the national average both when resource use was lower than the national average and also when it was higher, depending on the condition.

However, we did test another measure of quality to determine whether MSA scores would change if another type of quality measure were introduced. We found that MSA quality rankings by condition did vary depending on the measure type. The other quality measures we used were the potentially avoidable hospitalizations. We did

not include these in our initial MSA composite scores because of their low incidence. However, for this analysis we combined all potentially avoidable hospitalization indicators into one score for each MSA.

Potentially avoidable admissions are also related to the amount of resources used. In this case if more resources were used (hospitalizations), the MSA would look worse on quality and worse on resource use relative to other MSAs. We found that using a measure that could have a different relationship with resource use did shift MSA quality rankings.[15] For example, Miami's overall score relative to the national average on necessary care indicators was 1.02, or 2 percent higher than the national average. However, this shifted to 19 percent lower than the national average when quality was measured using potentially avoidable hospitalizations.

To put this in context, in a region where beneficiaries generally use more resources, if quality is assessed solely on claims-based indicators of the provision of clinically necessary services, then an area with higher resource use could look better on quality. However, if measured on whether potentially avoidable hospitalizations were avoided, they would look worse. Both types of indicators are linked with measuring resource use, but provide different pictures of quality. If Medicare measures physician quality along with resource use, it would be important to have a broader set of clinical indicators less linked with resource use. Measures of the use of quality-enhancing tools, such as information systems to track patient care and outcomes, would also be important, as the Commission recommended in its pay-for-performance recommendations.

Future work

The analysis shows that it is possible to use these types of episode groupers as one tool to measure physician resource use at the aggregate MSA level. We found that both groupers assign a large proportion of claims and dollars to episodes, and that episodes can be analyzed for variation in resource use and the types of services within an episode. We also found that the vast majority of episodes could be attributed to a provider, and that even when we allowed episodes to be assigned to more than one provider, most episodes were still assigned to a single provider.

However, the analysis also raised several technical and analytic issues that will need to be addressed as Medicare considers using these groupers to understand physician resource use. Additional research is needed to better understand how the clinical logic underpinning the groupers affects the construction of episodes and whether differences in coding patterns can affect per episode comparisons of resource use both within and across regions. We also found that using claims-based quality indicators is possible, but that a broader set of indicators less related to service delivery will be important.

The second step of our analysis (with a 100 percent sample in several geographic regions) will provide more information on the feasibility of applying these tools at the individual physician level. Among the issues we will address in the upcoming research are:

- What is the minimum number of individual episodes or quality indicators a physician must have in order to qualify for those episodes to be included in a resource use or quality analysis?
- What is the minimum number of total episodes physicians must have in order to be compared to their peers?
- What proportion of a physician's practice is captured after removing episodes without a clean start or finish, outlier episodes, and episodes that do not meet the minimum threshold for inclusion in resource use comparisons?
- How large is the distribution of physician resource use and quality scores?
- How should risk-adjustment techniques be implemented?
- Are there differences in practice patterns within more defined geographic areas that could lead to bias in per episode comparisons? ■

Standardized payments

We used the following methods to standardize payments.

Hospital inpatient services—We applied the standardized amount for each diagnosis related group (DRG) for each year to all records uniformly. Cases involving transfers were adjusted according to the payment rules laid out in regulation.

Skilled nursing facility (SNF) services—We merged the SNF Medicare Provider Analysis and Review records to the DataPro SNF Stay file, which contains linked claims, Minimum Data Set data, and Online Survey, Certification, and Reporting system data for SNF stays nationwide. This information was combined with specific standardized amounts of resource utilization groups from CMS to create standardized payment amounts.

Long-term care hospital services—For discharges that occurred on or after October 1, 2002, we applied the standardized amount for each DRG. For discharges prior to this date, we backed out local area wage-index adjustments from each hospital's payment, 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. We then created a DRG-level per diem amount, which was multiplied by the length of stay for each record.

Home health—We identified the home health case-mix weight on each claim and multiplied the weight by the base payment rate for the appropriate fiscal year.

Physician services—We identified the relative value unit (RVU) for each record by matching the Healthcare Common Procedure Coding System (HCPCS) and modifier on the record to the physician fee schedule RVU file. We then multiplied the RVU 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—We used the HCPCS code on the ASC facility records to match records to ASC payment rate files. We then assigned the ASC payment rate to each record based on the HCPCS and reduced the payment rate for multiple surgical procedures on the same claim (the payment for second and subsequent procedures was reduced by 50 percent, consistent with Medicare payment rules).

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 on the clinical lab fee schedule. Each service on the lab fee schedule has a separate payment for each carrier, as well as the national limitation amount (NLA). The NLA is based on the median of the carrier rates and represents an upper payment limit for each service. In practice, most lab services are paid the NLA rate. The standardized payment rate for each lab record is the NLA for the service.

Anesthesia services—We summed the base units and the time units for each anesthesia record, and multiplied the sum 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—We used the HCPCS code to match outpatient records to an outpatient prospective payment system payment rate file. We then assigned a standardized payment amount to each record based on that payment rate. ■

Endnotes

1 Three groupers are primarily used by the private sector—ETGs, MEGs, and the Cave Grouper. Our choice of ETGs and MEGs was not based on any analysis of the utility of applying them to Medicare. Several large private plans use the Cave Grouper and it is being considered by CMS for testing, alongside the other two, as part of the agency's study on episode groupers.

2 Information on inpatient hospital and SNF stays was taken from a 5 percent sample of the Medicare Provider Analysis and Review file.

3 We will continue to evaluate different outlier approaches in our analysis of 100 percent of Medicare claims in selected geographic areas. We will also evaluate whether deleting outlier episodes versus truncating outlier episodes has a measurable impact on physician rankings.

4 The MEG grouper employs two different approaches to staging, integer staging and substages, which further categorize within each integer stage.

5 Additionally, the Ambulatory Quality Alliance is developing standard methodologies that would apply to all three major groupers.

6 Not all episodes have three stages, and some episodes have four stages. For example, all CHF episodes are assumed to be stage 3 episodes and prostate and colon cancers have four stages running from stage 0 (significant predisposing risk factor for the disease, but no current pathology) to stage 3.

7 We intend to use both types of risk adjusters in the second step of our physician resource use analysis.

8 Using Berenson-Eggers Type of Service codes, we divided total episode payments into seven categories: inpatient, E&M services performed in both physician offices and hospital outpatient departments, post-acute care (including SNF, long-term care hospital, and home health services), procedures, imaging, tests, and other.

9 Analyzing ETG episodes by type of service produced broadly similar results.

10 In this context, efficiency does not necessarily mean cost of care only. To the extent that quality of care can be measured, it should be incorporated into any analysis of physician efficiency.

11 Using E&M dollars, 11 percent of episodes were attributed to more than one physician.

12 We used 35 percent instead of 30 percent, which we used for resource use.

13 These results have not been risk adjusted.

14 One condition where this is well illustrated is breast cancer. Women who need mammographies in our sample number 237,081. The other indicators for breast cancer have 30,000 or fewer beneficiaries eligible because they require that women have a diagnosis of breast cancer.

15 In this case, the quality would be lower if resource use were higher (more hospitalizations).

References

Fisher, E. S., D. E. Wennberg, T. A. Stukel, et al. 2003. The implications of regional variations in Medicare spending. Part 1: The content, quality, and accessibility of care. *Annals of Internal Medicine* 138, no. 4 (February 18): 273–287.

Hornbrook, M. C., A. V. Hurtado, and R. E. Johnson. 1985. Health care episodes: Definition, measurement, and use. *Medical Care Review* 42, no. 2 (Fall): 163–218.

Medicare Payment Advisory Commission. 2006. *Medicare Ambulatory Care Indicators for the Elderly: Refinement of the Access to Care for the Elderly Project indicators.* Contractor report by Magnacare Health Services Improvement, Inc. www.medpac.gov.

Medicare Payment Advisory Commission. 2005. *Report to the Congress: Medicare payment policy.* Washington, DC: MedPAC.

Medicare Payment Advisory Commission. 2004. *Report to the Congress: New approaches in Medicare.* Washington, DC: MedPAC.

Thomson Medstat. 2005. *Medstat Episode Grouper white paper.* Ann Arbor, MI: Thomson Medstat.

Nolan, T., and D. Berwick. 2006. All-or-none measurement raises the bar on performance. *Journal of the American Medical Association* 295, no. 10: 1168–1170.

Symmetry Health Data Systems, Inc. 2005. Users guide, Episode Treatment Groups™, version 5.2.

Thomas, J. W. 2006a. Economic profiling of physicians: Does omission of pharmacy claims bias cost efficiency measurement? *American Journal of Managed Care.* Forthcoming.

Thomas, J. W. 2006b. Should episode-based economic profiles be risk-adjusted to account for differences in patients' health risks? *Health Services Research* 41, no. 2 (April): 581–598.