© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2009.00990.x
RESEARCH ARTICLE

Provider Monitoring and Pay-for-Performance When Multiple Providers Affect Outcomes: An Application to Renal Dialysis

Richard A. Hirth, Marc N. Turenne, John R.C. Wheeler, Qing Pan, Yu Ma, and Joseph M. Messana

Objective. To characterize the influence of dialysis facilities and nephrologists on resource use and patient outcomes in the dialysis population and to illustrate how such information can be used to inform payment system design.

Data Sources. Medicare claims for all hemodialysis patients for whom Medicare was the primary payer in 2004, combined with the Medicare Enrollment Database and the CMS Medical Evidence Form (CMS Form 2728), which is completed at onset of renal replacement therapy.

Study Design. Resource use (mainly drugs and laboratory tests) per dialysis session and two clinical outcomes (achieving targets for anemia management and dose of dialysis) were modeled at the patient level with random effects for nephrologist and dialysis facility, controlling for patient characteristics.

Results. For each measure, both the physician and the facility had significant effects. However, facilities were more influential than physicians, as measured by the standard deviation of the random effects.

Conclusions. The success of tools such as P4P and provider profiling relies upon the identification of providers most able to enhance efficiency and quality. This paper demonstrates a method for determining the extent to which variation in health care costs and quality of care can be attributed to physicians and institutional providers. Because variation in quality and cost attributable to facilities is consistently larger than that attributable to physicians, if provider profiling or financial incentives are targeted to only one type of provider, the facility appears to be the appropriate locus.

Key Words. Pay-for-performance, monitoring, dialysis

Private and public payers are focusing on measuring and rewarding quality and efficiency in health care (Milgate and Cheng 2006; Rosenthal et al. 2006). Such efforts include "pay-for-performance" (P4P) systems that reward measured performance, capitation systems that put providers at financial risk for high utilization, and tiered networks in which insurers use measured performance to assign providers to "preferred" status levels.

A key component of the design of such systems is the determination of whose performance to measure and reward. Typically, patients have contact with multiple physicians and institutions (Pham et al. 2007). For example, surgical outcomes could be affected by the surgeon, the surgical team (surgeons, nurses, and anesthetists), and the institution where the surgery was performed. Therefore, determining rules regarding attribution of outcomes to providers creates major challenges in payment system design. Failure to accurately identify the provider or providers with the greatest influence on outcomes could adversely affect the credibility and impact of the payment system, and it could make providers accountable for decisions outside their control. Similarly, the validity of provider profiles, which are being developed for quality assessment and improvement (Bodenheimer 1999) using more rigorous methods (Huang et al. 2005; Shahian et al. 2005; Zheng et al. 2006), also depends on whether they identify the providers with the greatest ability to affect the outcome the policy maker is trying to influence.

In principle, outcomes could be measured and rewarded for any or all of the providers or types of providers involved in a patient's care. However, doing so would present a variety of challenges. Data may not be available across all providers and only limited case mix adjusters may be available to control for differences in the patient populations. Even when data are available, statistical power to differentiate outcome variations associated with different physicians and facilities is often limited by sample size. Likewise, providers treating atypical patient populations could face substantial financial risks under a prospective payment system (PPS).

Owing in part to these difficulties, decisions about whom to measure and reward have generally not been based on empirical analyses of the relative impact of particular providers (or of different types of providers) on outcomes. Rather, decisions about which provider (or type of provider) to attribute

Address correspondence to Richard A. Hirth, Ph.D., Department of Health Management and Policy, University of Michigan School of Public Health, 109 S. Observatory, Ann Arbor, MI 48109-2029; e-mail: rhirth@umich.edu. Marc N. Turenne, Ph.D., is with Kidney Epidemiology and Cost Center, Ann Arbor, MI. John R.C. Wheeler, Ph.D., is with Department of Health Management and Policy, University of Michigan School of Public Health, 109 S. Observatory, Ann Arbor, MI. Qing Pan, Ph.D., is with George Washington University, Department of Statistics, Washington, DC. Yu Ma, M.S., Joseph M. Messana, M.D., are with Kidney Epidemiology and Cost Center, Ann Arbor, MI.

responsibility have been based on factors such as convenience (e.g., measurement at the institutional level due to easier availability of data or large sample sizes), prospective assignment of patients to a designated "gatekeeper" physician (Rosenthal et al. 2006), or arbitrary retrospective rules such as attributing responsibility to all providers with a minimum level of patient contact or to the single provider with the most patient contact during the year (Dudley and Rosenthal 2006; Milgate and Cheng 2006). The Medicare Payment Advisory Commission (MedPAC) as well as physicians and their professional societies have expressed concern regarding the attribution methods in use today (American College of Cardiology 2006; Milgate and Cheng 2006; Sinsky 2007; American Academy of Family Physicians 2008).

This paper uses renal dialysis services to demonstrate a method for identifying the extent to which different types of providers influence variation in resource use and patient outcomes. Dialysis provides an excellent context for this study. Patients have ongoing relationships with both an institutional provider (the dialysis facility) and a physician (the nephrologist who manages dialysis-related services). Multiple nephrologists practice within most dialysis facilities, and most nephrologists practice in multiple facilities. This double "cross-over" facilitates the statistical identification of physician and facility effects on outcomes. Because the vast majority of dialysis patients are insured by Medicare, available data include a large number of patients. Further, detailed clinical data are available to adjust for case mix. Finally, several clinical performance measures are well established and relatively well accepted.

Dialysis facilities have a financial incentive to increase the use of the services, primarily injectable medications, which are paid on a fee-for-service basis by Medicare. However, physicians, who generally do not profit from these services, are ultimately responsible for prescribing care. Given the recent controversy about appropriate anemia management in dialysis facilities, these issues are particularly salient. Researchers have presumed that the organization is the decision making locus (e.g., Thamer et al. 2007), while others have argued that institutional protocols are physician driven and modified by individual physicians in response to patient condition (Lazarus and Hakim 2007).

Although outcomes and resource utilization may depend on both the dialysis facility and the nephrologist, public reporting of performance measures (Dialysis Facility Compare; http://www.cms.hhs.gov/DialysisFacilityCompare/), quality improvement initiatives (e.g., http://www.esrdnetworks.org), P4P proposals (Milgate and Cheng 2006), and the development of an expanded case mix–adjusted dialysis PPS as required by the Medicare Prescription Drug, Improvement and Modernization Act of 2003 (Pub. L. 108-173) all use the dialysis facility as the locus of measurement and/or reward. Not only does this implicitly attribute responsibility to the facility for the practices of nonemployee physicians but also failure to report at the physician level provides no guidance to patients regarding choice of physician, and failure to provide physician incentives may forego opportunities to improve care.

Two prior studies are particularly relevant. Krein et al. (2002) developed an empirical basis for deciding which provider level to profile (facility, professional group, or physician) in the context of diabetes care in the Veterans Administration (VA) system. They found that for outcome and resource use measures, variation at the facility level is dramatically higher than that at the physician level. Physician variation was substantial only for narrow process measures (ordering of specific laboratory tests), and the provider group explained relatively little variation in any measure. However, their study was limited to 13 facilities in one VA region.

A second prior study investigated the relative variation of resource use in U.S. dialysis facilities across four levels: facilities, nephrologists, patients, and time (different months for a given patient) (Turenne et al. 2008). The analysis of four levels of variation created computational limitations which required a sampling strategy that limited the analysis to a 4 percent random sample of facilities and distinguished provider-level effects only through multiple physicians practicing within a facility (and not from physicians practicing in multiple facilities). Although this study also found that the variation across facilities exceeded across physicians, the physician-level variation was relatively more important than that found by Krein, with financially significant variation in resource use across both facilities and physicians.

The current study extends this previous research in several significant ways. First, by aggregating data across multiple months for each patient, this study uses data from almost all physician-facility pairs and both types of "cross-over" between physicians and facilities. Second, the prior dialysis study only examined resource utilization (costs per dialysis session for a set of services, primarily injectable medications and laboratory tests). The current study uses the same utilization measure but adds two outcome measures (achieving treatment targets for dose of dialysis and anemia management). Third, this study uses slightly more recent data (2004) than the prior dialysis study (2003).

Research Objectives

Our primary objectives are to characterize the influence of dialysis facilities and nephrologists on resource use and patient outcomes and to suggest how such information could inform payment system design. Therefore, we wish to build a model that controls for case mix factors that may vary across providers. We expect that resource use and outcomes are influenced by observed and unobserved patient characteristics. Based on prior research, we identified a set of potential case mix adjusters (Hirth et al. 2003, 2007; Wheeler et al. 2006). Given that each dialysis patient has an ongoing relationship with both a dialysis facility and a nephrologist, we also expect that utilization of services and clinical outcomes could be independently influenced by facility and physician. Because of the focus on payment policy, we do not consider provider characteristics such as nonprofit status or membership in a dialysis chain because the Centers for Medicare & Medicaid Services (CMS) would be unlikely to implement payment levels or incentives specific to these provider subgroups. Conversely, distinguishing between facilities and physicians is useful because separate performance measures, payment rates, and incentives could be designed for each type of provider.

Methods

Data Sources

Data for this study come from several CMS sources. Medicare claims for all renal dialysis patients for whom Medicare was the primary payer in 2004 (307,805 patients) were used to identify resource utilization, dose of dialysis, and anemia management. Demographic information was obtained from the Medicare Enrollment Database and the CMS Medical Evidence Form (CMS Form 2728), which is completed at onset of renal replacement therapy (RRT). Height, weight, and several patient comorbidities present at the start of RRT are also reported on CMS Form 2728. Diagnosis codes reported on Medicare claims between 1999 and 2004 were used to identify comorbidities that were not included on CMS Form 2728 and to capture changes in patient condition since start of RRT.

Assignment of Patients to Facilities and Physicians

Medicare provider identification numbers and unique physician identification numbers (UPINs) as reported on monthly dialysis claims were used to identify the treating facilities and physicians. For 75.8 percent of the analysis sample, this process identified a unique physician/facility pair that delivered all care during 2004; the remaining patients switched facilities and/or physicians, resulting in more than one record for that patient.

Dependent Variables

Resource use was measured based on Medicare allowable payments (MAP) from Medicare claims. MAP includes both Medicare payments and patient co-pay obligations, and therefore reflects a societal perspective on resource use. Secondary analyses take Medicare's perspective as a payer by excluding patient obligations. Three types of services were included in this utilization measure. The first is injectable medications (primarily erythropoietin [EPO], iron, and vitamin D products) billed by dialysis facilities. The second is laboratory tests that were either billed by dialysis facilities or ordered by physicians receiving monthly capitation payments for treating ESRD patients and billed by freestanding laboratory suppliers on Medicare carrier claims. This includes a broad spectrum of tests, including those used to monitor patient response to medications. The third and smallest category includes other services billed by dialysis facilities, such as syringes and other supplies that may be used with medications or laboratory tests. Taken together, these are the services that are currently billed on a fee-for-service basis, but they are expected to become part of an expanded, prospectively paid bundle of services subsequent to Medicare legislation passed in 2008. The costs of the dialysis treatment itself are not included because they have been paid as a prospective bundle since 1983. Therefore, Medicare claims reflect only the number of treatments received, and not patient-level utilization of resources. Restricting the measure of resource use to those services billed separately from the dialysis treatment addresses the current policy context of expanding the prospective bundle and the ongoing controversy about the intensity of anemia management, which accounts for the majority of these services.

These MAPs were summed across all treatments delivered to each patient during the year while under the care of a given physician/facility dyad, and they were then standardized to the number of outpatient dialysis sessions by calculating the average MAP per session. This standardization is appropriate because dialysis units have little discretion over the number of treatments delivered per unit of time. Medicare pays for only three treatments per week and almost all patients are on that schedule. Medicare grants medical exceptions allowing a fourth weekly treatment for <1 percent of patients, and less than 1 percent of patients receive two weekly treatments (these are individuals who retain residual renal function during the first months of dialysis).

We capped outlier values for average EPO MAP/session ($>$\$300 or $>$30,000 units/session, which represented 1.2 percent of claims and may reflect clinically implausible or inappropriate doses). CMS currently places a similar limit on EPO reimbursement (500,000 units/month). Extreme values for MAP/session for all other services were capped at the greater of the upper outer fence (75th percentile + [3 × interquartile range]) or the 99th percentile of the distribution. These caps were employed to prevent a few extreme values from contributing a disproportionate share of variance.

The other dependent variables were clinical measures reported on claims submitted by dialysis facilities. The first measure is the urea reduction ratio (URR), which indicates the percent of urea removed from the patient's blood during the dialysis treatment. Therefore, this variable represents the "dose" of dialysis delivered. Clinical guidelines call for a URR of at least 65 percent (http://www.kidney.org/professionals/kdoqi/guidelines_updates/doqi_uptoc.html#hd, accessed on February 6, 2008) and CMS reports the percentage of patients achieving this target at the facility level. Because renal failure patients often have some residual renal function when starting dialysis, URR is not a pure measure of dose. Therefore, we excluded months occurring during a patient's first year of dialysis from the URR analysis. The second measure is hematocrit (HCT), which indicates how well anemia is being managed.[1] Clinical guidelines call for maintaining a HCT of at least 33 percent (http://www.kidney.org/professionals/KDOQI/guidelines_anemia/index.htm, accessed on February 6, 2008), and CMS also reports the percentage of patients achieving this target at the facility level. Potential adverse effects of HCT levels substantially above the target value became controversial after the period of this study (2004), but revised guidelines have continued to call for a minimum value of 33 percent.

Analysis Sample

307,805 chronic renal failure patients were identified in the Medicare claims as having received outpatient dialysis during 2004, over 90 percent of whom received in-center hemodialysis. To create a more homogeneous study population, we excluded patients receiving other dialysis modalities (primarily peritoneal dialysis) because they tend to use fewer injectable medications than hemodialysis patients and because the URR target of 65 percent applies only to hemodialysis patients. In addition, we excluded patients who received less than 1 month of outpatient hemodialysis (<13 sessions), and those treated by physician/facility dyads with fewer than five patients. Finally, patients with missing values for the dependent variables, physician identifiers, or several key case mix adjusters (e.g., body size measures) were excluded. The analysis sample contains 196,670 unique patients, treated by 4,166 facilities, 4,820 physicians, and 10,737 facility/physician pairs. There were 8,714 facility/physician pairs treating at least five patients for the URR analysis sample due to exclusion of patient-months during first year of dialysis. The average physician/facility pair cared for 18.3 patients. Due to computational limitations, the MAP/session and HCT models were estimated using an 80 percent random sample of the data. The URR model was estimated using 100 percent of the data.

Variance Components Analyses

Multilevel mixed effects models were used to estimate the variation in resource use associated with dialysis facilities, physicians, and patients. The following linear mixed model was estimated (Searle, Casella, and McCulloch 1992; Verbeke and Molenberghs 2000; Goldstein 2003):

$$Y_{f,d,p} = X_p\beta + \varepsilon_f + \varepsilon_d + \varepsilon_{f,d,p}, \tag{1}$$

where $Y_{f,d,p}$ is the average MAP/session for facility f, physician d, and patient p; X_p is a vector of patient characteristics that are included as risk adjusters, and β is the corresponding vector of coefficients. There are random effects for facility (ε_f), physician (ε_d), and the residual error for each patient ($\varepsilon_{f,d,p}$). To determine the impact of case mix adjustment on the relative and absolute contributions of physicians and facilities to observed variation, unadjusted models (without covariates X_p) were also estimated.

A linear cost model was estimated for ease of interpretation and because preliminary analyses revealed only mild skewness. Some use of the laboratory tests and drugs was nearly universal (less than .1 percent of patients had zero costs), and the skewness of the tail was limited (99th percentile of spending was less than four times the mean). A log-transformed model was estimated as a sensitivity analysis.

For the clinical performance measures, the dependent variables represented the proportion of months in which the patient achieved the URR and HCT targets. In addition to the linear specifications, alternative specifications were estimated to ensure that conclusions regarding the absolute and relative magnitudes of variation at the physician and facility level were not sensitive to functional form. In particular, the data-generating process underlying the proportions can be thought of as a series of binomial trials (meet the target or not in a given month). Therefore, we estimated a generalized linear model (GLM) with a binomial error distribution with a logit link function for a 20 percent random sample of the data. However, because some patients were consistently more or less likely than average to meet the targets, the actual distribution of the proportions was bimodal (reflecting overdispersion relative to the binomial error distribution). Therefore, we also estimated GLM models with dichotomized outcomes based on whether the target was met in >50 percent of months to determine if the conclusions are robust to the functional form of the models.

The prediction error that results from the difference between actual $Y_{f,d,p}$ and predicted $\hat{Y}_{f,d,p} = X_p\hat{\beta}$ outcomes has a separate component for each of the three levels of variation:

$$Y_{f,d,p} - \hat{Y}_{f,d,p} = \hat{\varepsilon}_f + \hat{\varepsilon}_d + \hat{\varepsilon}_{f,d,p}. \tag{2}$$

The estimated facility component ($\hat{\varepsilon}_f$) reflects consistently higher or lower outcomes than predicted for individual facilities compared with an average facility. Similarly, the estimated physician component ($\hat{\varepsilon}_d$) captures consistently higher or lower outcomes than predicted for individual physicians compared with the average physician, given the facility in which the physician practices. The estimated residuals ($\hat{\varepsilon}_{f,d,p}$) reflect the remaining unexplained variation from patient to patient.

The magnitudes of the prediction error components were compared using standard deviations (SD) (e.g., SD of $\hat{\varepsilon}_p$). These models were estimated using the lmer procedure in *R* (version 2.2.1; Faraway 2006). To assess model fit, we checked the residual plots at patient level, physician level, and facility level, finding only mild deviations from normality.

Case Mix Measures

Patient characteristics identified as potential case mix adjusters included age, sex, race, Hispanic ethnicity, and Medicaid eligibility at the onset of renal failure, rural versus urban location, duration of RRT, HCT, and body size at the start of RRT, plus 39 comorbid conditions. Comorbidities included heart disease, cancer, infections, anemia, and bleeding conditions that were expected to affect resource use and clinical outcomes. Only recent claims diagnoses were used for acute conditions (e.g., within 3 months for gastrointestinal bleeding) or in cases where preliminary bivariate analyses for chronic conditions revealed that recent diagnoses were more highly predictive (based on the comparisons of diagnoses reported within 1, 2, or 5 years). These characteristics represent a more inclusive list of factors that would likely be used to risk adjust an expanded PPS.

RESULTS

To statistically distinguish variations at the facility and physician levels, it is necessary that some facilities have multiple physicians and/or some physicians treat patients at multiple facilities. Both types of crossover occurred frequently in the dialysis setting. In nearly two-thirds of facilities, more than one physician cared for at least five patients (frequency distribution shown in Figure 1). Similarly, more than half of physicians cared for at least five patients in multiple facilities (Figure 2). Descriptive statistics appear in Table 1.

Variation across facilities, physicians, and patients for the entire sample is described in Table 2. For purposes of illustration, variation at the physician, facility, and patient levels are characterized as the mean for the outcome variable ± 1 SD. A consistent result is that variation across facilities exceeded that across physicians, with SD at the facility level more than double those observed at the physician level. In addition, the results show that variation across providers was considerably lower than variation across individual patients treated by a given physician at a given facility.

Figure 1: Frequencies of Physicians by Facility

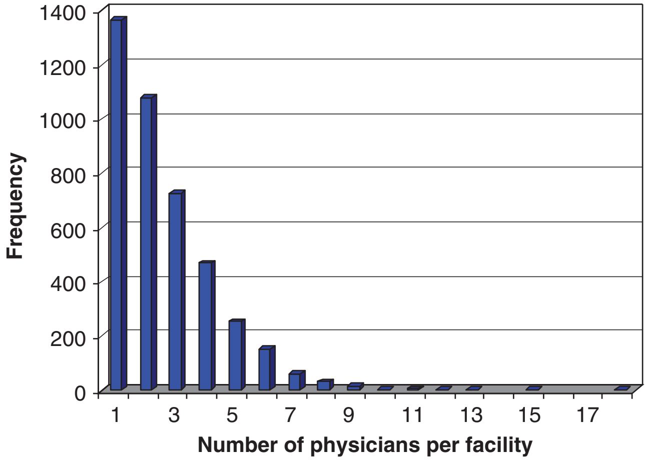
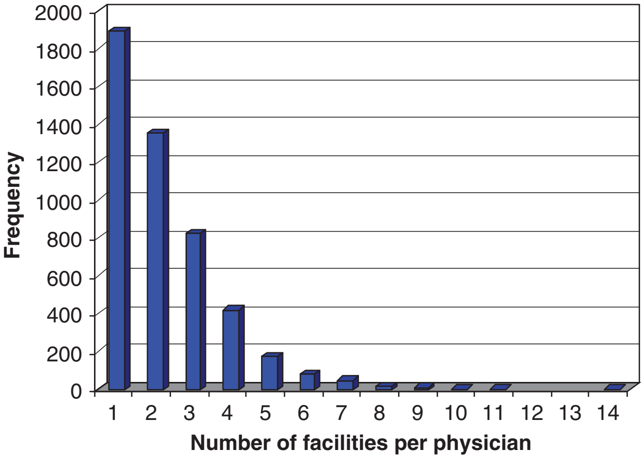


Figure 2: Frequencies of Facilities by Physician



To determine the effect of case mix adjustment on the variation attributable to providers, Table 3 reports adjusted and unadjusted SDs across facilities, physicians, and patients. Case mix adjustment generally decreased the SDs, but only to a modest extent. Both the absolute magnitudes of the SDs across providers and the relative magnitudes of the facility and physician SDs remained similar after case mix adjustment. Finally, the alternative functional forms did not affect the qualitative conclusions. For SB MAP/session, variances in the log model were transformed to the dollar scale using the delta method (Casella and Berger 2002). The transformed SDs at the facility, physician, and patient levels were about 10 percent lower than those from the linear specification, but the relative magnitudes were virtually unchanged. Likewise, for the clinical measures the absolute and relative variation at the three different levels was very similar in the alternative, GLM specifications. Using only actual Medicare payments per dialysis session (that is, excluding patient obligations) as a measure of resource use reduced mean spending to \$93.01 per session and resulted in SDs of \$6.51 across physicians, \$18.38 across facilities, and \$49.74 across patients.

Discussion

The success of attempts to improve care through P4P and provider profiling relies upon the identification of providers most able to enhance efficiency and

Table 1: Characteristics of Medicare Hemodialysis Patients, 2004

	*Study Sample**	
Variable	*N or Mean*	*SD*
Levels of variation		
Number of dialysis facilities	4,166	
Number of physicians	4,820	
Number of patients	196,670	
Patients per facility/physician pair	18.3	
Dependent variables		
Average MAP per dialysis session including patient obligation	$113.06	$69.67
Average MAP per dialysis session excluding patient obligation	$93.08	$57.46
Urea reduction ratio $\geq$ 65%	90.20%	
Hematocrit $\geq$ 33%	77.10%	
Selected patient characteristics		
Age (years)		
<18	0.1%	
18–44	10.3%	
45–54	22.1%	
60–69	22.8%	
70–79	27.4%	
80+	17.3%	
Hematocrit at start of RRT	29.4	5.4
Cardiac dysrhythmia within 1 year	35.7%	
Cardiac arrest within 5 years	2.7%	
Metastatic cancer within 5 years	2.8%	
Acquired hemolytic anemias within 1 year	1.5%	
Unable to ambulate	2.8%	
AIDS diagnosis within 5 years	1.9%	
Rural	16.7%	
Medicaid eligibility	27.3%	
Race (Native American)	1.5%	
Race (Asian or Pacific Islander)	3.3%	
Race (Black)	36.0%	
Race (White)	57.0%	
Race (other or unknown)	2.1%	
Hispanic	13.0%	

*Includes all Medicare hemodialysis patients with available case-mix measures (including but not limited to those shown above) and physician identifiers (for physician counts and ratios), who received at least 13 hemodialysis treatments and were treated by a facility/physician pair with at least five patients.

quality. Using renal dialysis services, we demonstrate a method for determining the extent to which variation in health care costs and quality can be attributed to physicians and institutional providers. Essentially, variation at the provider level serves as a proxy for the degree of control over clinical resource

Table 2: Variation in Resource Use and Quality Measures across Physicians, Facilities, and Patients, Adjusted for Case Mix

	Physician (n = 4,280)	*Dialysis Facility (n = 4,166)*	*Patient (n = 165,465)*
Resource use for separately billable services ($/session)*			
Mean+SD	$120.30	$135.35	$173.08
Mean	$112.96	$112.96	$112.96
Mean – SD	$105.62	$90.57	$52.84
Percent of months with hematocrit ≥ 33%			
Mean+SD	79.58%	83.48%	100.00%
Mean	77.03%	77.03%	77.03%
Mean – SD	74.48%	70.58%	49.67%
	Physician (n = 4,456)	*Dialysis Facility (n = 3,978)*	*Patient (n = 141,312)*
Percent of months with URR ≥ 65%†			
Mean+SD	92.72%	96.50%	100.00%
Mean	90.20%	90.20%	90.20%
Mean – SD	87.68%	83.90%	70.52%

*Number of physician/facility pairs: 10,737. Results are based on an 80% random sample. Resource use included patient obligations.

†Number of physician/facility pairs: 8,714.

utilization or outcomes (Young 2008). This analysis allowed us to use the two-way crossover between physicians and facilities to identify the sources of variation. Earlier studies relied primarily on crossover arising from multiple

Table 3: Effect of Case-Mix Adjustment on Variation across Providers and Patients

	Physician	*Dialysis Facility*	*Patient*
Resource use for separately billable services ($/session)*			
SD (unadjusted)	$7.55	$24.52	$64.64
SD (adjusted)	$7.34	$22.39	$60.12
Percent of months with hematocrit ≥ 33%			
SD (unadjusted)	2.64%	6.90%	28.25%
SD (adjusted)	2.55%	6.45%	27.36%
Percent of months with URR ≥ 65%†			
SD (unadjusted)	2.73%	6.54%	20.75%
SD (adjusted)	2.52%	6.30%	19.68%

*Number of physician/facility pairs: 10,737. Results are based on an 80% random sample. Resource use included patient obligations.

†Number of physician/facility pairs: 8,714.

physicians practicing in relatively small samples of facilities (Krein et al. 2002; Turenne et al. 2008).

Variation in quality and cost attributable to facilities was larger than that attributable to physicians for all three measures. Therefore, the results confirmed those from the limited sample used in the earlier analysis of resource utilization (Turenne et al. 2008) and generalize them to two important clinical performance measures. They were also broadly consistent with Krein et al.'s (2002) finding that facility mattered more than physician. However, the share of variation attributable to physicians appears higher in dialysis care than for similar measures in the VA diabetes context.

Given the greater variation found at the institutional level, if provider profiling or financial incentives such as P4P or bundled payments are targeted to only one type of provider, the dialysis facility appears to be the more appropriate locus. This suggests that incentives for quality and efficiency can be directed toward organized providers, consistent with conclusions drawn by Sautter et al. (2007) based on their evaluation of hospital P4P. They conclude that institutional providers are able to bring organizational resources to bear in response to incentives and thereby improve care processes. Nonetheless, the existence of clinically meaningful variation across physicians implies that quality reports, bundled payments, and P4P may place facilities at risk for outcomes they only partially control. Further, physicians may be relatively more influential in a subset of facilities, potentially making them a better target for incentives. Likewise, physicians may influence protocols at the facility level, so it is possible that the average effects reported here understate physicians' contributions to the facility's clinical practices.

The financial impact of the observed variation in resource use was large, with the facility-level SD of $22.39 per session translating into to $179,120 for a moderately sized facility performing 8,000 hemodialysis treatments annually. Similarly, clinical variations occurring at the facility and physician levels were meaningful. Relative to the percentage of patients failing to attain treatment targets (10 percent for URR and 23 percent for HCT), the magnitude of the outcome variations observed across providers (SDs of 3 percent across physicians and 6–7 percent across facilities) was substantial. Therefore, cooperation between managers and physicians in the development and adoption of protocols to optimize clinical outcomes and resource utilization is likely to become increasingly important under P4P programs and proposed reforms to pay prospectively for drugs and lab tests. Further, methods to align incentives of dialysis facilities and nephrologists should be developed. These findings support MedPAC's position that facilities and physicians should both be included in dialysis P4P programs in order to encourage collaboration (Milgate and Cheng 2006).

This line of research may also help inform which interventions are most likely to improve performance. As noted by Young (2008), identifying performance variation across facilities suggests that managerial intervention may be successful (e.g., improvements in the organization's health information technology), while the existence of variation across physicians points to the use of measures such as evidence-based guidelines. Young also noted that the type of provider to which incentives are targeted will itself influence the types of investments made to improve performance. For example, programs targeting individual physicians are unlikely to induce organizational infrastructure investments. Indeed, the substantial residual variation at the patient level also suggests that efforts such as encouraging patient compliance with therapy could also be valuable.

Several limitations should be noted. First, the random effects identified the statistical contribution of providers to observed outcomes, but they cannot distinguish differences arising from discretionary practices from those arising from unobserved case-mix differences. However, our conclusions were robust to controlling for a much broader set of comorbidities than is used to case mix adjust the current, publicly reported dialysis facility outcomes data. Second, MAP/session is a cost measure based on utilization and payment rates. Actual input costs were not available and may affect provider practices. Third, other levels of variation may exist beyond those explored here (e.g., physician groups and chains of facilities) and could be explored in future research. Fourth, it is likely that the physician UPINs reported on dialysis claims misidentify the primary physician for some patients. This would bias estimates of physician's contributions downward. Finally, physicians might have greater influence in other domains such as hospitalization.

Future research could apply similar methods to determine the extent to which patterns observed in dialysis generalize to other settings. For example, the dialysis setting has similarities to the choice of a primary care physician (PCP) in a managed care plan. In both settings, the assignment of a patient to a responsible physician is seemingly straightforward, but other types of providers may still contribute to measured performance. Similar analytic methods could also be used to assess the validity of common, but arbitrary, attribution rules (e.g., attributing outcomes to the physician who accounted for the plurality of visits during a year). For example, the relationship between percentage of visits accounted for by the primary and secondary providers and their respective effects on observed outcomes could be estimated.

To the extent that these analyses can help policy makers and insurers understand the sources of cost and outcome variation, they will be more able to develop appropriate and effective monitoring, reporting, and incentive systems. Likewise, such information can be used by providers to identify opportunities for improvement and to anticipate and manage financial risks and opportunities under such systems.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The research for this article was supported by the Centers for Medicare & Medicaid Services (CMS) under CMS contract numbers 500-2006-0048C and 500-00-0028.

Disclosures: Richard Hirth has served as a consultant for Affymax and Cormedix. These relationships began after the submission of the manuscript.

Disclaimers: None.

NOTE

1. Erythropoietin stimulates the production of red blood cells. Almost all patients with chronic kidney failure experience anemia due to the lack of natural erythropoietin production by their kidneys. CMS requires reporting of the patient's HCT level in order for a facility to receive payment for any of the erythropoietin-stimulating agents used to treat anemia. Therefore, in the small percentage of patient-months in which none of these drugs are billed, HCT values are usually unreported.

REFERENCES

American Academy of Family Physicians. 2008. "Pay-for-Performance" [accessed on February 6, 2008]. Available at http://www.aafp.org/online/en/home/policy/policies/p/payforperformance.html

American College of Cardiology. 2006. "American College of Cardiology 2006 Principles to Guide Physician Pay-for-Performance Programs." *Journal of the American College of Cardiology* 48: 2603–9.

Bodenheimer, T. 1999. "The American Health Care System: The Movement for Improved Quality in Health Care." *New England Journal of Medicine* 340 (6): 488–92.

Casella, G., and R. L. Berger. 2002. *Statistical Inference.* 2nd Edition. Belmont, CA: Duxbury Press.

Dudley, R. A., and M. B. Rosenthal. 2006. "Pay for Performance: A Decision Guide for Purchasers." Agency for Healthcare Research and Quality publication 06-0047, Rockville, MD.

Faraway, J. J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Raton, FL: Chapman & Hall/CRC Press.

Goldstein, H. 2003. *Multilevel Statistical Models.* 3rd Edition. London: Edward Arnold.

Hirth, R. A., M. N. Turenne, J. R. Wheeler, A. S. Pozniak, P. Tedeschi, C. C. Chuang, Q. Pan, K. Slosh, and J. M. Messana. 2007. "Case-Mix Adjustment for an Expanded Renal Prospective Payment System." *Journal of American Society of Nephrology* 18 (9): 2525–33.

Hirth, R. A., R. A. Wolfe, J. R. Wheeler, E. C. Roys, R. J. Tedeschi, A. S. Pozniak, and G. T. Wright. 2003. "Is Case-Mix Adjustment Necessary for an Expanded Dialysis Bundle?" *Health Care Financing Review* 24 (4): 77–88.

Huang, I. C., C. Frangakis, F. Dominici, G. B. Diette, and A. W. Wu. 2005. "Application of a Propensity Score Approach for Risk Adjustment in Profiling Multiple Physician Groups on Asthma Care." *Health Service Research* 40 (1): 253–78.

Krein, S. L., T. P. Hofer, E. A. Kerr, and R. A. Hayward. 2002. "Whom Should We Profile? Examining Diabetes Care Practice Variation among Primary Care Providers, Provider Groups, and Health Care Facilities." *Health Services Research* 37 (5): 1159–77.

Lazarus, J. M., and R. M. Hakim. 2007. "Dialysis Facility Ownership and Epoetin Dosing in Hemodialysis Patient: A Provider's Perspective." *American Journal of Kidney Disease* 50 (3): 366–70.

Milgate, K., and S. B. Cheng. 2006. "Pay-for-Performance: The MedPAC Perspective." *Health Affairs* 25 (2): 413–9.

Pham, H. H., D. Schrag, A. S. O'Malley, B. Wu, and P. B. Bach. 2007. "Care Patterns in Medicare and Their Implications for Pay for Performance." *New England Journal of Medicine* 356 (11): 1130–9.

Rosenthal, M. B., B. E. Landon, S. T. Normand, R. G. Frank, and A. M. Epstein. 2006. "Pay for Performance in Commercial HMOs." *New England Journal of Medicine* 355 (18): 1895–902.

Sautter, K. M., B. G. Bokhour, B. White, C. J. Young, J. F. Burgess, D. Berlowitz, J. R. C. Wheeler, and S. R. Grossbart. 2007. "The Early Experience of a Hospital-Based Pay-for-Performance Programme." *Journal of Healthcare Management* 52 (2): 95–107.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance Components.* New York: John Wiley & Sons Inc.

Shahian, D. M., D. F. Torchiana, R. J. Shemin, J. D. Rawn, and S. T. Normand. 2005. "Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methodology." *Annals of Thoracic Surgery* 80 (6): 2106–13.

Sinsky, C. A. 2007. "Letter." *New England Journal of Medicine* 356 (8): 872.

Thamer, M., Y. Zhang, J. Kaufman, D. Cotter, F. Dong, and M. A. Hernan. 2007. "Dialysis Facility Ownership and Epoetin Dosing in Patients Receiving Hemodialysis." *Journal of American Medical Association* 297 (15): 1667–74.

Turenne, M. N., R. A. Hirth, Q. Pan, R. A. Wolfe, J. M. Messana, and J. R. C. Wheeler. 2008. "Using Knowledge of Multiple Levels of Variation in Care to Target Performance Incentives to Providers." *Medical Care* 46 (2): 120–6.

Verbeke, G., and G. Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Wheeler, J. R., J. M. Messana, M. N. Turenne, R. A. Hirth, A. S. Pozniak, Q. Pan, C. C. Chuang, K. Stish, P. Tedeschl, E. C. Roys, and R. A. Wolfe. 2006. "Understanding the Basic Case-Mix Adjustment for the Composite Rate." *American Journal of Kidney Disease* 47 (4): 666–71.

Young, G. J. 2008. "Can Multi-Level Research Help Us Design Pay-for-Performance Programs?" *Medical Care* 46 (2): 109–11.

Zheng, H., R. Yucel, J. Z. Ayanian, and A. M. Zaslavsky. 2006. "Profiling Providers on Use of Adjuvant Chemotherapy by Combining Cancer Registry and Medical Record Data." *Medical Care* 44 (1): 1–7.

Supporting Information

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

Appendix SA1: Author Matrix.

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