HSR Health Services Research

DOI: 10.1111/1475-6773.12467
RESEARCH ARTICLE

An Evaluation of Performance Thresholds in Nursing Home Pay-for-Performance

Rachel M. Werner, Meghan Skira, and R. Tamara Konetzka

Objective. Performance thresholds are commonly used in pay-for-performance (P4P) incentives, where providers receive a bonus payment for achieving a prespecified target threshold but may produce discontinuous incentives, with providers just below the threshold having the strongest incentive to improve and providers either far below or above the threshold having little incentive. We investigate the effect of performance thresholds on provider response in the setting of nursing home P4P.

Data Sources. The Minimum Data Set (MDS) and Online Survey, Certification, and Reporting (OSCAR) datasets.

Study Setting and Design. Difference-in-differences design to test for changes in nursing home performance in three states that implemented threshold-based P4P (Colorado, Georgia, and Oklahoma) versus three comparator states (Arizona, Tennessee, and Arkansas) between 2006 and 2009.

Principal Findings. We find that those farthest below the threshold (i.e., the worst-performing nursing homes) had the largest improvements under threshold-based P4P while those farthest above the threshold worsened. This effect did not vary with the percentage of Medicaid residents in a nursing home.

Conclusions. Threshold-based P4P may provide perverse incentives for nursing homes above the performance threshold, but we do not find evidence to support concerns about the effects of performance thresholds on low-performing nursing homes.

Key Words. Performance-based thresholds, quality of care, pay-for-performance, nursing home quality, long-term care

The use of pay-for-performance (P4P) to improve health care quality has become commonplace in the United States. Despite the proliferation of P4P programs, there is mixed evidence to support their use, with the effects of P4P being variable across settings and programs.

One possible explanation for this mixed evidence is that provider response to P4P has been heterogeneous because of variation in how specific P4P programs are designed. Typical P4P incentives give providers financial bonuses or add-on payments for achieving prespecified quality goals, including achieving a target threshold (e.g., performance above a predetermined level), a relative rank (e.g., performance in the top 10 percent of all providers), or improvement on quality metrics (e.g., improving performance over the prior year's performance). While each of these quality goals has anticipated pros and cons, empirical evidence supporting their use is scarce. Understanding the tradeoffs of these design choices may enable design of more effective P4P programs.

Our objective was to investigate empirically the effect of using performance thresholds on provider response to financial incentives, an area of P4P design in which evidence is lacking. Performance thresholds are easy for payers to implement and transparent for participating providers, making them common in current P4P programs. However, the use of thresholds to determine bonus payments may have several downsides, assuming that improving scores requires investment of resources. First, threshold-based payments give providers no direct incentive to improve beyond the targeted threshold. Thus, a ceiling of quality improvement may be observed at the threshold, blunting the average effect of P4P on quality improvement. Second, they may give little incentive to low-performing providers to improve, as their performance may be far from the threshold with little chance of achieving the targeted threshold, making returns on investment in quality improvement unlikely. Indeed, threshold-based incentives may result in the most robust performance improvement among providers for whom it is easiest to obtain a financial bonus—those that are just below the threshold. This would result in P4P having limited impact on performance overall and failing to improve performance among those that would most benefit from improvement in terms of measured performance.

While threshold-based incentives have been shown to provide discontinuous incentives in non–health care settings (Grant 2010; Neal and Schanzenbach 2010), early studies of pay-for-performance in health care found that contrary to predictions, providers with the lowest baseline per-

Address correspondence to Rachel M. Werner, M.D., Ph.D., Division of General Internal Medicine, Perelman School of Medicine at the University of Pennsylvania, 1204 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104; e-mail: rwerner@upenn.edu. Rachel Werner is also with the Center for Health Equity Research and Promotion at the Crescenz VAMC, Philadelphia, PA. Meghan Skira, Ph.D., is with the Department of Economics, Terry College of Business, The University of Georgia, Athens, GA. R. Tamara Konetzka, Ph.D., is with the Department of Health Studies, University of Chicago, Chicago, IL.

formance improved the most. For example, Rosenthal et al. examined changes in physician performance after implementation of a P4P program in California and found that the largest improvements were among those physicians with baseline performance farthest below the threshold needed for financial rewards (Rosenthal et al. 2005; Mullen et al. 2010). Similarly, Beaulieu and Horrigan studied physicians enrolled in a threshold-based P4P program in New York, finding that those with the lowest baseline score had the largest improvements (Beaulieu and Horrigan 2005). While the findings that the lowest performing providers improved the most have been consistent, these studies have not accounted for the possibility of regression to the mean. That is, providers with the lowest performance may be most likely to improve even in the absence of P4P simply because their baseline performance is most extreme in the distribution of performance.

The objective of this study was to test provider response to the implementation of recent threshold-based P4P while controlling for regression to the mean. Specifically, we test whether providers farthest below the threshold are least likely to improve and whether providers close to the threshold are most likely to improve. We do so in the setting of nursing home P4P, comparing three states that recently introduced threshold-based P4P programs to three similar states without a P4P program in nursing homes.

BACKGROUND

Nursing Home Quality

Over 1.5 million people reside in U.S. nursing homes at a cost of over $120 billion per year (Kaiser Family Foundation 2007). Despite this frequent use and high cost of nursing home care, quality of care in nursing homes has long presented a policy challenge (Institute of Medicine 1986). Major regulatory policies aimed at improving nursing home quality were implemented in 1987 under the Nursing Home Reform Act or the Omnibus Budget Reconciliation Act (OBRA), a congressional act that mandated extensive regulatory controls. As a result of OBRA, each Medicare- or Medicaid-certified nursing home is inspected at least once every 15 months and is required to submit a comprehensive assessment of each chronic-care resident at least once per quarter. While researchers found that OBRA led to improved quality (Kane et al. 1993; Shorr et al. 1994; Castle et al. 1996; Fries et al. 1997; Mor et al. 1997; Snowden and Roy-

Byrne 1998), a follow-up report by the Institute of Medicine in 2000 concluded that significant problems remain (Wunderlich and Kohler 2000).

With regulation failing to fully reform nursing home quality, efforts have turned toward market-based reforms designed to improve quality of care. Since 2002, a number of state Medicaid agencies have implemented P4P programs based on the quality of chronic care delivered using financial incentives tied to Medicaid payment (Kane et al. 2007; Werner et al. 2010). A prior evaluation of this quality-improvement effort in nursing homes found that the effect of P4P on quality was inconsistent—performance on some quality measures improved more in states that had P4P compared to states that did not, while performance on other quality measures did not, and the effect varied state by state (Werner et al. 2013).

Nursing Home P4P

Between 2002 and 2009, eight states adopted Medicaid-sponsored P4P programs in nursing homes, all of which primarily targeted quality of care for long-stay (or chronic-care) residents. Four of these programs rewarded nursing homes based in part on the clinical quality of care they delivered. We focus our analyses on three of these four states—those that used threshold-based performance on clinical quality measures to determine bonus eligibility: Colorado, Georgia, and Oklahoma.

The details of these programs have been previously described (Werner, Konetzka et al. 2010). While the programs are heterogeneous in many ways, each state uses a payment model based on a point system that is translated into bonus payments. For each measure included in the payment model, each nursing home is evaluated and earns points based on whether it has achieved a target level of performance. The earned points are summed across all measures and translated into a per-diem add-on for all Medicaid resident-days, where nursing homes with more points receive higher add-ons. Table 1 displays the clinical quality measures used by each state's P4P program.

The maximum add-on (and thus potential size of the financial incentive) varies by state. Colorado's program used a $4 per-diem maximum add-on during the study period, which translated to an approximately 2.8 percent increase in per-diem rates based on the state's average Medicaid per-diem rate in 2004 (Grabowski et al. 2008). Georgia's program used a 2 percent maximum add-on (which is equivalent to approximately $2.39 per Medicaid

Table 1: Clinical Quality Measures Included in State P4P Programs

% of Residents Who	*Mean (SD)*	*Colorado (7/2008 to Present)*	*Georgia (7/2007 to Present)*	*Oklahoma (7/2007 to Present)*
Were physically restrained	7.1 (7.8)	X	X	X
Developed pressure sores	12.9 (9.5)	X	X	X
Had moderate to severe pain	9.1 (9.7)	X	X	
Had bladder catheter inserted	6.7 (6.0)			X
Had falls	9.6 (5.9)			X
Had unexplained weight loss	9.2 (6.9)			X

patient day). Oklahoma used a $5.50 per-diem add-on (or an approximately 5.7 percent increase in per-diem rates).

Each state defines its threshold for bonus payment differently, and the transparency of what the threshold is during the performance period varies by state. Table 2 summarizes each state's approach to creating a performance threshold for payment. Colorado has the most transparent threshold, using predetermined thresholds for each clinical quality measure. Colorado sets two thresholds per clinical quality measure. While nursing homes are eligible to earn points toward their bonus payment for meeting either threshold, they can earn more points from achieving performance above the higher of the two

Table 2: Summary of Thresholds Used in Each P4P State

Colorado

- Set two prespecified thresholds for each quality measure
- NHs with performance equal to or above the higher threshold earned more points

  Note: only state with prespecified thresholds

Georgia

- Average performance on each measure in that period used as threshold

  Note: threshold calculations are based on concurrent state average

Oklahoma

- Uses several steps to create a composite measure of performance
  1 Calculates the percentile ranking on each clinical quality measure
  2 Averages the percentile rankings across clinical quality measures within nursing home
  3 Uses the median of the percentile average as the state threshold

Notes: threshold calculations are based on concurrent relative ranking; only state that uses a composite measure thresholds. In Georgia, the state sets the threshold at the state's average for each clinical quality measure in that year. Thus, nursing homes do not know exactly what the threshold is until after the year is over. Finally, in Oklahoma, the state creates a rank-based composite across all of the clinical quality measures included in the program. To do so, it first calculates each nursing home's percentile ranking on each measure, then averages the percentile rankings within each nursing home, and finally uses the median of this average across nursing homes as the threshold for that year. While only Colorado's threshold value is transparent to nursing homes during the measurement period, all nursing homes have access to their performance data and how they compare to other nursing homes in their state on the CMS website, Nursing Home Compare, which publicly rates all U.S. nursing homes on the clinical quality measures included in these P4P programs.

METHODS

Conceptual Approach

We conceptualize that nursing homes will respond to P4P incentives by attempting to improve performance if the expected payoff from P4P exceeds the cost of improving performance. We assume that quality improvement is costly for nursing homes, involving efforts such as staff hiring, staff training, or investments in infrastructure to support improved performance. We thus hypothesize that nursing homes with performance just under the P4P threshold will have the greatest incentive to improve, as a small increment in quality could result in a bonus. On the other hand, if quality improvement is costly, P4P may provide inadequate incentive for low-performing nursing homes to improve or high-performing nursing homes to maintain their high performance.

Overview of Empirical Approach

We use a difference-in-differences approach to identify changes in nursing home performance in states that implemented P4P compared to states that did not. For comparator states, we chose one neighboring control state for each P4P state. We selected the neighboring state that most closely matched the P4P state on their mean nursing home performance, standard deviation around the mean, and mean change in nursing home performance during the pre-P4P period, choosing Arizona (compared to Colorado), Tennessee (compared to Georgia), and Arkansas (compared to Oklahoma). We chose comparator states that were similar in these characteristics to ensure that P4P and comparator states had similar trends in nursing home performance in the absence of P4P (which is required for valid use of the difference-in-differences approach)[1] and, because we are interested in looking at changes in the distribution of performance around the thresholds, to ensure that the distributions of performance were similar at the baseline.

To investigate how changes in performance under P4P were related to the use of threshold-based P4P, we categorize nursing homes based on how far their prior performance was from the state's threshold and then test how nursing home performance changed relative to the threshold in states with and without P4P, before and after P4P was implemented. By using comparator states in a difference-in-differences model, we can test for changes in performance while controlling for the expected regression to the mean. We hypothesize that nursing homes with performance closest to the threshold will have the largest improvement in performance, those farthest below the threshold will have the smallest improvement in performance, and those farthest above the threshold will experience worsening quality.

Study Sample, Study Time Period, and Data

Our study sample covers the period from 2006 to 2009, spanning the years when P4P was implemented in the states we study (July 2007 in Georgia and Oklahoma; July 2008 in Colorado[2]). In each P4P state and its comparator, we test for changes in nursing home performance during the 1 year after P4P was implemented in that state compared to the 1 year prior to P4P implementation. We include all nursing homes in each P4P and comparator state. This included 887 P4P nursing homes and 670 comparator nursing homes. Characteristics of these nursing homes 1 year prior to P4P implementation in each state are displayed in Table 3.

We use two datasets: the Minimum Data Set (MDS) and the Online Survey, Certification, and Reporting (OSCAR) data. The MDS contains detailed clinical data collected at regular intervals (usually quarterly) for every resident in a Medicare- and/or Medicaid-certified nursing home. The MDS is used by nursing homes to assess the needs and develop a care plan for each resident as well as for Medicare and Medicaid payment. We use the MDS data to identify residents included in this study and to measure nursing home quality and several covariates (defined below). OSCAR is derived from data collected in mandated state inspections of nursing homes

Table 3: Comparison of Nursing Homes in States with P4P (Colorado, Georgia, and Oklahoma) and without P4P (Arizona, Tennessee, and Arkansas) in the One Year Prior to Implementation of P4P in Each State

	CO	*AZ*		*GA*	*TN*	*p-*	*OK*	*AR*	*p-*
	(n = 204)	*(n = 128)*	*p-value*	*(n = 353)*	*(n = 315)*	*value*	*(n = 330)*	*(n = 227)*	*value*
Nursing home characteristics									
Percent Medicaid, mean (SD)	58.2 (21.9)	57.8 (28.7)	.872	73.5 (18.6)	65.6 (21.9)	.000	68.1 (18.8)	69.5 (17.8)	.389
Percent Medicare, mean (SD)	11.2 (12.9)	14.6 (14.5)	.026	11.4 (9.6)	16.3 (14.7)	.000	10.0 (13.2)	11.3 (13.4)	.263
Ownership, %			.118			.016			.413
Government	8.9	3.1		8.2	5.5		3.6	4.5	
Not for profit	20.8	20.5		25.5	18.1		9.4	12.6	
For profit	70.3	76.4		66.3	76.4		87.0	82.9	
Hospital-based, %	5.9	0.8	.019	13.3	5.5	.001	1.9	3.1	.377
Chain, %	58.4	57.8	.914	74.2	64.5	.007	30.8	50.0	.000
Total beds, mean (SD)	96.2 (45.0)	123.8 (51.9)	.000	112.9 (51.2)	119.0 (55.5)	.137	94.0 (39.5)	107.8 (34.3)	.000
Female residents, %	69.0	68.2	.646	71.7	72.3	.500	71.9	71.7	.875
Race of residents, %									
White	76.7	86.4	.000	68.9	85.2	.000	87.7	84.2	.017
Black	3.5	4.1	.519	30.0	14.3	.000	6.4	15.3	.000
Hispanic	8.5	11.3	.038	0.4	0.2	.000	1.0	0.2	.000
Other	1.5	7.8	.000	0.6	0.3	.138	4.9	0.3	.000
Age, mean (SD)	80.4 (8.0)	79.1 (8.7)	.182	78.9 (6.6)	79.9 (4.9)	.028	78.7 (5.9)	80.8 (4.3)	.000
Cognitive performance scale, mean (SD)	2.7 (0.6)	2.6 (0.6)	.045	3.3 (0.5)	3.1 (0.6)	.000	2.4 (0.5)	2.9 (0.5)	.000

Continued

Table 3. *Continued*

	CO	*AZ*		*GA*	*TN*	*p-*	*OK*	*AR*	*p-*
	(n = 204)	*(n = 128)*	*p-value*	*(n = 353)*	*(n = 315)*	*value*	*(n = 330)*	*(n = 227)*	*value*
ADL scale, mean (SD)	11.1 (1.9)	10.8 (1.5)	.114	11.3 (1.4)	11.5 (1.4)	.105	9.5 (1.3)	11.0 (1.4)	.000
Clinically complex scale, mean (SD)	0.8 (0.3)	0.6 (0.3)	.000	0.7 (0.3)	0.4 (0.2)	.000	0.2 (0.2)	0.5 (0.2)	.000
Nursing home performance measures									
% of residents who. . ., mean (SD)									
Had moderate to severe pain	9.2 (10.2)	11.0 (8.3)	.092	9.4 (7.5)	8.2 (6.8)	.030	12.5 (10.9)	6.9 (6.0)	.000
Developed pressure sores	10.2 (7.4)	11.2 (6.7)	.237	14.3 (7.5)	12.4 (6.5)	.000	15.8 (7.8)	12.8 (8.8)	.000
Were physically restrained	5.3 (8.6)	3.9 (4.7)	.110	6.2 (4.8)	7.6 (7.2)	.005	10.3 (7.1)	12.8 (8.2)	.002
Had falls	11.8 (4.0)	10.7 (5.2)	.023	9.4 (3.0)	8.6 (4.1)	.009	9.1 (4.0)	8.6 (3.4)	.095
Had bladder catheter inserted	8.0 (6.3)	8.1 (7.1)	.902	5.2 (3.7)	7.2 (6.4)	.000	7.7 (7.9)	6.2 (3.2)	.008
Had unexplained weight loss	10.0 (8.1)	7.9 (4.5)	.007	9.7 (4.0)	10.5 (4.5)	.015	8.7 (5.6)	8.5 (4.1)	.702

and contains facility-level characteristics such as payer mix and ownership status. We merge these two datasets using the closest OSCAR observation to each MDS assessment.

Nursing homes generally serve two populations—long-stay residents (who typically receive nonskilled care such as assistance with activities of daily living) and postacute residents (who receive rehabilitative care following an acute-care hospitalization). We limit our sample to long-stay (or chronic-care) nursing home residents, the focus of the P4P programs we evaluate. Because MDS includes data on both long-stay and short-stay nursing home residents, we identify long-stay residents as those having at least one quarterly or annual assessment in addition to an admission (or prior quarterly/annual) assessment. This ensures that the resident has been in the nursing home for at least 90 days, a commonly used cutoff to distinguish short-stay from long-stay residents.

Dependent Variables: Nursing Home Quality

Our dependent variables are the facility-level clinical quality metrics used by states to determine P4P payments derived from the MDS (the same source used by state Medicaid agencies to measure clinical quality and determine P4P incentive payments) following measure specifications from the Centers for Medicare and Medicaid Services (CMS) (Morris et al. 2003; Nursing Home Quality Initiative 2004).

Following CMS convention to avoid overweighting sicker residents who may have more frequent assessments, we limit each resident to only one assessment per quarter, choosing the most recent assessment in a quarter if a resident has more than one assessment per quarter[3]; and we do not include admission assessments, as patient outcomes on admission cannot be attributed to the admitting nursing home's quality of care. We then follow the CMS specifications to determine which assessments are eligible for each clinical quality measure (creating the denominator of the measure) and, among those eligible, which had the outcome of interest (the numerator). Each facility-level quality measure is risk adjusted according to CMS specifications. The resulting facility-level measures are expressed as percentages. A list of the specific clinical outcomes included in state P4P programs is included in Table 1, including a descriptive summary of each measure. Because these rates vary significantly from one another in mean and variance, we used normalized quality measures in all regressions by converting them to *z*-scores.

Key Independent Variables: Nursing Home Distance from Payment Threshold

Our key independent variable is the interaction between a P4P indicator variable and a measure of how far a nursing home was from the performance threshold in the prior year. By using each nursing home's prior year average to determine distance from the threshold (rather than prior quarter), we obtain more stable estimates of baseline performance and avoid capturing idiosyncratic shocks in the baseline average. The P4P indicator is simply a time-varying dummy variable for each nursing home indicating whether it was in a state with a P4P program in that time period. It thus equals one in P4P states after P4P has been implemented and zero otherwise.

To measure how far a nursing home was from its threshold, we first calculated the nonnormalized performance threshold in year t for each P4P state and its comparator state. In states with P4P after P4P was initiated, this is the actual threshold that was used to determine incentive payments at the end of year t once P4P was initiated. In P4P states prior to the implementation of P4P and in comparator states, this is a hypothetical threshold that would have been used had the P4P program been in place. In two P4P states (Georgia and Oklahoma) and their comparator states (Tennessee and Arkansas), these thresholds were based on the state's mean of each quality measure and median of the combined quality measures, respectively. In Colorado, we use the thresholds predefined by its P4P program (testing the effect of both the higher and lower thresholds in separate regressions) and assigned those same thresholds to the pre-P4P period and the comparator state.

We next calculated how far nursing homes were from the threshold in year $t - 1$. To do this, we first calculated nursing home performance on each measure in year $t - 1$. We then subtracted the state's threshold in year t from that performance to get the "distance" between a nursing home's performance last year and the threshold they face this year. Next, we categorized nursing homes in each year, first dividing nursing homes into two groups based on whether the facility was above or below the threshold and then dividing each group of nursing homes into three subgroups based on how far from the threshold the nursing home was—the 25 percent closest to the threshold, the 25 percent farthest from the threshold, and the middle 50 percent. In Colorado and Arizona, over 25 percent of nursing homes had no restraint use, resulting in more than 25 percent of nursing homes in the highest performing quantile. In this case, nursing homes with no restraint use were put in the top quantile and the remaining nursing homes with performance above the threshold were equally divided into the two remaining quantiles. In all cases, this resulted in dummy variables representing the six groups of nursing homes, with three groups above the threshold and three below the threshold.

We interact the dummy variables for these six groups of nursing homes with the P4P dummy variable described above to test how nursing home performance changed in P4P states when P4P was implemented relative to how far above or below the threshold the nursing home was, compared to nursing homes in states without P4P that were similarly above or below their counterfactual threshold.

Covariates

To account for heterogeneity in facility characteristics across nursing homes, we included the following covariates from OSCAR in all analyses: a facility's percent of residents covered by Medicare, percent of residents covered by Medicaid, ownership; whether the facility is hospital-based, part of a chain; and its total number of beds. To account for heterogeneity in resident characteristics across nursing homes, we include the following resident characteristics from the MDS, which are aggregated at the facility level: each facility's mean resident age, percent of residents who are female, percent in each racial and ethnic group, and the facility's mean Cognitive Performance Scale (Morris et al. 1994), ADL scale (Morris et al. 1999), and Clinically Complex Scale (Kidder et al. 2002).

Empirical Specifications

We implement a difference-in-differences framework, using linear models. First, we estimate the main effect of P4P on nursing home performance using the following equation:

$$QM_{j,t} = \alpha P4P_{j,t} + \varphi X_{j,t} + \tau_t + \gamma_j + \varepsilon_{j,t} \tag{1}$$

where *QM* is nursing home *j*'s performance in period *t* (where *t* is a quarter) and *P4P* is an indicator for whether a nursing home was in a state with P4P in time *t*. We also include facility-level time-varying covariates, quarterly fixed effects, and nursing home fixed effects. The coefficients on the *P4P* indicator reflect the difference-in-differences estimates of the overall response to P4P implementation among nursing homes in P4P states compared to similar nursing homes in comparator states.

We then estimate the effects of interest—the effects of using performance thresholds:

$$QM_{j,t} = \alpha P4P_{j,t} + \beta \sum\nolimits_{Q=2}^{6} Quantile_{j,t-1} + \delta P4P_{j,t} * \sum\nolimits_{Q=2}^{6} Quantile_{j,t-1} + \varphi X_{j,t} + \tau_t + \gamma_j + \varepsilon_{j,t} \qquad (2)$$

In equation 2, we add a set of five dummy variables indicating which performance quantile the nursing home was in period $t - 1$ and the interactions between the P4P indicator and the quantile indicators. The coefficients on the interactions are the difference-in-differences estimates of the response to P4P implementation among nursing homes compared to similar nursing homes in a comparator state, compared to the omitted category of nursing homes (nursing homes far above the threshold), after P4P was implemented compared to before. We hypothesize that nursing homes in quantile 4 (those nursing homes just below the threshold) will have the largest improvements in P4P states after P4P was implemented. Additionally, if nursing homes are uncertain about their position relative to the threshold, we hypothesize there will also be large improvements among nursing homes in quantile 3 (those just above the threshold). We also hypothesize the nursing homes in quantile 1 (farthest above the threshold) will have declines in performance and nursing homes in quantile 6 (farthest below the threshold) will have little improvement. As nursing home performance measures target adverse outcomes, a negative coefficient would indicate improvement in nursing home performance among nursing homes in a P4P state compared to similar nursing homes in a non-P4P state, compared to nursing homes in quantile 1 and compared to before P4P was implemented.

For our main analyses, we focus on three clinical quality measures as dependent variables: the percent of residents who had moderate to severe pain, developed pressure sores, and were physically restrained. We chose these three measures because they are common across the states in their P4P programs[4] and because a prior evaluation of these P4P programs showed improvement in nursing home performance on these three measures after P4P was implemented, although the results were inconsistent across states, suggesting heterogeneity in state response (Werner et al. 2013). We estimate this equation separately for each performance measure and for each P4P and comparator state pair. We also estimate this equation separately for each performance measure, but pooling the three P4P and three comparator states.

The size of the maximum financial incentive is directly related to the size of the Medicaid population in each nursing home. Therefore, we also run the main analysis stratifying by the percentage of residents covered by Medicaid in each nursing home. To do this, we identify nursing homes with the percentage of Medicaid residents in the lowest quartile, the highest quartile, and in the interquartile range. We also test the robustness of our results by using a false P4P implementation date, moving the date 1 year earlier in each state and ending the post-P4P period prior to the actual P4P implementation.

We adjust for clustering of observations within nursing homes over time using the Huber White sandwich estimator in all regressions.

RESULTS

Table 3 shows a comparison of facility characteristics, resident characteristics, and performance outcomes for nursing homes in the P4P and comparator states in the year prior to P4P implementation. We control for differences in facility and resident characteristics by including these characteristics as covariates in all regressions and by including nursing home fixed effects, which control for unobserved time-invariant differences across nursing homes. As expected when comparing just a few states on numerous attributes, we note some residual differences in baseline nursing home performance despite choosing comparator states with performance most similar to that of P4P states. However, there do not seem to be systematic differences in baseline performance across P4P and comparator nursing homes. In some cases, P4P states exhibit worse baseline performance, and in others cases they exhibit better performance depending on the clinical outcome.

We then estimate the main effect of P4P—that is, the average effect across all nursing homes. As expected, the implementation of P4P was associated with heterogeneous changes in nursing home performance (see Table S2). While rates of pressure sores declined across all three states, changes in rates of physical restraint use were more varied. Georgia achieved statistically significant reductions in restraint use, while Colorado and Oklahoma experienced increases in restraint use compared to the comparator states. Changes in pain were small and not statistically significant. This heterogeneous average response to nursing home P4P replicates previous findings (Werner et al. 2013).

In Table 4 we show the effect of performance thresholds on nursing home performance. These results pool the three P4P states and three comparator states, and we find two main results—as predicted, nursing homes with performance highest above the threshold at baseline experienced larger declines in performance in P4P states than in comparator states and, contrary to predictions, nursing homes in P4P states that were farthest below the threshold had the largest gains in performance under P4P compared to similar nursing homes in comparator states. The coefficient on p4p is positive, suggesting that in P4P states nursing homes highest above the threshold (or in quantile 1, the omitted category) experienced a decline in performance after P4P was implemented, compared to similar nursing homes in comparator states, with declines ranging from 0.20 to 0.23 standard deviations. Looking at the combination of the coefficient on p4p and the coefficient on p4p*Quantile 2 interaction, we see the effect of P4P implementation on nursing homes in quantile 2 in P4P states compared to similar nursing homes in comparator states. Similar to quantile 1 nursing homes, the second-highest performing

Table 4: Regression Results Comparing All Three P4P States to Three Comparator States

	Pain	*Pressure Sores*	*Restraints*
p4p	0.228*** (0.0736)	0.231*** (0.0773)	0.204*** (0.0426)
Quantile 2	−0.0222 (0.0531)	−0.0162 (0.0497)	0.0504 (0.0459)
Quantile 3	−0.0249 (0.0681)	−0.0366 (0.0597)	0.103** (0.0518)
Quantile 4	−0.0623 (0.0654)	−0.0841 (0.0640)	0.0873* (0.0515)
Quantile 5	0.0569 (0.0723)	−0.0771 (0.0664)	0.103* (0.0567)
Quantile 6	0.325*** (0.107)	0.0302 (0.0934)	0.246*** (0.0809)
p4p*Quantile 2	−0.117 (0.0763)	−0.226*** (0.0854)	−0.106** (0.0494)
p4p*Quantile 3	−0.242*** (0.0837)	−0.198** (0.0993)	−0.277*** (0.0748)
p4p*Quantile 4	−0.174** (0.0831)	−0.405*** (0.0945)	−0.158*** (0.0554)
p4p*Quantile 5	−0.312*** (0.0777)	−0.309*** (0.0841)	−0.223*** (0.0474)
p4p*Quantile 6	−0.369*** (0.122)	−0.548*** (0.109)	−0.479*** (0.0821)
Constant	2.589 (1.870)	−1.122 (1.268)	−1.339* (0.792)
Covariates	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes
Nursing home FE	Yes	Yes	Yes
N	7,746	11,778	11,817
Adj. R^2	0.078	0.038	0.097

Notes. Robust standard errors in parentheses. Nursing homes in quantile 1 (the omitted category) are farthest above the threshold and, thus, the highest performers. The threshold is between quantiles 3 and 4. Nursing homes in quantile 6 are farthest below the threshold and, thus, the lowest performers.

*p < .10, **p < .05, ***p < .01.

group of nursing homes also experienced a small decline in performance relative to similar nursing homes in comparator states, ranging from 0.005 to 0.1 standard deviations.[5] The performance of nursing homes in quantiles 3 and 4 (just above and below the threshold) generally changed very little with the implementation of P4P compared to nursing homes in comparator states, though nursing homes in P4P states that were just below the threshold for pressure sores experienced an improvement in performance compared to nursing homes in non-P4P states. The largest changes were seen in nursing homes farthest below the threshold, which had improvements in performance ranging from 0.14 to 0.32 standard deviations.[6]

Next, we examine the effects state by state. In Colorado, the state that uses prespecified thresholds, we find only one statistically significant change in nursing home performance after P4P was implemented compared to before and compared to the control state without P4P using the higher threshold (Table S3) or the lower threshold (results not shown). However, even though the coefficient on p4p*Quantile 5 is statistically significant and negative, the sum of the coefficients on p4p and p4p*Quantile 5 is zero. In Georgia, where the state uses its concurrent mean performance as a threshold, we find that most quantiles of nursing homes improve across the three measures used by Georgia's program, and the largest improvements are among those nursing homes farthest below the performance threshold in the prior period (Table S4). Results from Oklahoma are similar to Georgia (Table S5). For the summary measure used to set the state's threshold, those nursing homes farthest below the threshold had the largest gains in performance compared to the non-P4P state of Arkansas, although these gains were similar in size to those nursing homes that were just below the threshold. Similarly, for most of the individual clinical quality measures that make up the summary measure, nursing homes farthest below the threshold generally had the largest improvement in performance.

Finally, we stratify by the percentage of residents covered by Medicaid and see similar effects (Table 5). Across most regressions, quantile 6 nursing homes have the largest improvements in performance, though after stratifying, some of the effects are no longer statistically significant. Additionally, the magnitude of the effect in the high-Medicaid group is similar to that for the mid- and low-Medicaid groups for all three quality measures.

When using a false P4P implementation date, we find support for our empirical strategy. P4P nursing homes in the highest performing quantile continue to have a decline in performance that is larger than in comparison states for pressure sores, but not for the other two measures (Table S6). Addi-

Table 5: Regression Results Combing All Three P4P and Comparator States and Stratifying by the Percentage of Nursing Home Residents with Medicaid, Where Low Is Less Than 25%, Midrange Is 25%–75%, and High Is >75%

	Pain			*Pressure Sores*			*Restraints*		
	% of Nursing Home Residents with Medicaid			*% of Nursing Home Residents with Medicaid*			*% of Nursing Home Residents with Medicaid*		
	Low	*Mid*	*High*	*Low*	*Mid*	*High*	*Low*	*Mid*	*High*
p4p	0.286 (0.220)	0.285** (0.127)	0.147 (0.0965)	0.239 (0.270)	0.241*** (0.0832)	0.191 (0.143)	0.272*** (0.0703)	0.110* (0.0564)	0.213*** (0.0752)
Quantile 2	0.111 (0.141)	−0.0764 (0.0605)	0.0414 (0.0698)	0.0409 (0.135)	−0.0277 (0.0680)	−0.00333 (0.101)	0.0425 (0.0874)	−0.00286 (0.0470)	−0.0174 (0.0776)
Quantile 3	0.0202 (0.190)	−0.0278 (0.0749)	−0.109 (0.101)	0.113 (0.151)	−0.0256 (0.0791)	−0.198 (0.131)	0.181 (0.117)	0.0352 (0.0635)	0.0188 (0.0894)
Quantile 4	0.109 (0.128)	−0.0596 (0.0800)	−0.117 (0.106)	−0.0918 (0.160)	−0.124 (0.0829)	0.0675 (0.139)	0.0917 (0.106)	−0.000403 (0.0600)	0.0818 (0.102)
Quantile 5	0.320** (0.140)	0.0533 (0.0897)	0.00368 (0.101)	−0.0325 (0.180)	−0.0497 (0.0814)	−0.107 (0.160)	0.0340 (0.133)	0.0459 (0.0644)	0.117 (0.102)
Quantile 6	0.514*** (0.177)	0.293** (0.143)	0.113 (0.172)	0.140 (0.271)	0.0585 (0.110)	−0.0417 (0.216)	0.206 (0.192)	0.246*** (0.0937)	0.0313 (0.169)
p4p* Quantile 2	−0.340 (0.219)	−0.131 (0.135)	−0.143 (0.0951)	−0.326 (0.297)	−0.175* (0.102)	−0.227 (0.143)	−0.0569 (0.110)	−0.0449 (0.0667)	−0.0579 (0.0757)
p4p* Quantile 3	−0.389* (0.226)	−0.279** (0.141)	−0.00212 (0.152)	−0.281 (0.286)	−0.169 (0.121)	−0.121 (0.197)	−0.147 (0.144)	−0.226** (0.0964)	−0.258** (0.130)
p4p* Quantile 4	−0.360 (0.226)	−0.270* (0.150)	−0.130 (0.133)	−0.342 (0.284)	−0.428*** (0.114)	−0.542*** (0.202)	−0.180* (0.1000)	−0.00530 (0.0857)	−0.210** (0.0897)
p4p* Quantile 5	−0.457** (0.220)	−0.395*** (0.134)	−0.325*** (0.121)	−0.298 (0.272)	−0.307*** (0.101)	−0.373** (0.162)	−0.231*** (0.0807)	−0.0894 (0.0710)	−0.218** (0.100)
p4p* Quantile 6	−0.325 (0.295)	−0.448*** (0.170)	−0.512* (0.289)	−0.561* (0.310)	−0.576*** (0.137)	−0.640*** (0.218)	−0.516*** (0.199)	−0.351*** (0.105)	−0.565*** (0.133)
Constant	3.538 (3.501)	5.608** (2.729)	−2.934* (1.667)	−3.448* (1.991)	−0.0130 (2.093)	0.786 (2.038)	−0.991 (0.875)	1.580 (1.271)	−3.569*** (1.318)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nursing Home FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	2,166	3,997	1,583	3,049	6,231	2,498	3,083	6,234	2,500
Adj. R^2	0.130	0.078	0.064	0.046	0.042	0.034	0.091	0.104	0.138

Notes. Robust standard errors in parentheses. Nursing homes in quantile 1 (the omitted category) are farthest above the threshold and, thus, the highest performers. The threshold is between quantiles 3 and 4. Nursing homes in quantile 6 are farthest below the threshold and, thus, the lowest performers.
$*p < .10$, $**p < .05$, $***p < .01$.

tionally, the coefficients on the interactions between (the false) P4P and the performance quantiles are generally insignificant and much smaller in magnitude than in our main analysis, providing evidence that our main results are not simply capturing regression to the mean or underlying differences in secular trend. These results also support the use of the chosen comparator states.

Discussion

In theory, threshold-based P4P may provide perverse incentives for providers with performance that is either far below or far above the threshold required for bonus payment, where average improvements from P4P programs might be driven by providers close to the threshold while those far above the threshold might worsen and those far below the threshold would not improve. In this case, P4P might miss its goals of improving care among those who most need to improve while maintaining high-quality care among others.

We find mixed evidence to support these concerns in the setting of nursing home P4P. Examining changes in nursing home performance in three states that implemented threshold-based P4P incentives, we find that in two of the three states, the highest performing nursing homes experienced a decline in performance once performance thresholds were put in place. We also find that improved performance under P4P is driven primarily by nursing homes farthest below the threshold (rather than those close to the threshold). The one state with no effect from P4P (Colorado) on these outcomes was also the only state to use predefined transparent thresholds, making it difficult to comment on the effect of threshold transparency.

While worsening performance among high-performing nursing homes and improving performance among low-performing nursing homes may seem consistent with regression to the mean, these changes were observed in P4P states above and beyond any changes that occurred in comparator states or any changes that occurred prior to the implementation of P4P that could be attributed to regression to the mean in comparator states and were not seen when using an earlier, false P4P implementation date.

Our findings of a strategic response to threshold-based incentives among high-performing nursing homes but not low-performing nursing homes may be puzzling. However, prior literature has suggested that high performance is highly correlated with organizational factors such as strong leadership and strong finances (Park and Werner 2011). These factors may also allow high-performing providers to be more strategic in response to new incentives.

It is also possible that what appears to be a difference in a general strategic response to P4P among high- versus low-performing nursing homes is rather a difference in the specific strategies nursing homes implemented. It is commonly assumed that improving performance is costly. If we expect that nursing homes far below the threshold would have to invest such large sums to see their investment pay off in P4P rewards, they might opt out. However, the large improvements we find among this group of nursing homes suggest that nursing homes may be able to achieve significant improvements through less costly mechanisms. Those farthest below the threshold may have relatively low costs of improvement because they have not yet implemented the simplest solutions, and P4P gives them an incentive to do so. It is also possible that nursing homes used alternative low-cost methods to improve their performance by simply improving (or changing) the coding in the data used for the P4P performance metrics or by shifting resources away from areas with relatively high performance to areas with low performance that are targeted by P4P. If either of these were the case, we would expect to see exactly what we found—improvements across the distribution of baseline performance, but especially among the poorest performing nursing homes. We might also expect to see worsening of unmonitored care, a phenomenon we do not test for here.

Several limitations of this work should be noted. First, this is an observational study. While we use rigorous quasi-experimental design to account for the usual biases resulting from observational data, it remains possible that our results are subject to endogeneity biases. In particular, P4P is not randomly assigned to states; nursing home performance (and changes in nursing home performance) may be correlated with the decision to adopt P4P. Similarly, while we selected control states that closely matched P4P states on several factors, they are imperfect counterfactuals. It is also important to note that we do not address whether quality truly improved or whether the documentation of these outcomes changed; in this study, we examine changes in nursing home scores in response to P4P incentives but remain agnostic about whether the chosen metrics reflect true quality. Finally, in this paper, we focus on several selected measures of quality that are relevant to threshold-based bonuses and do not directly consider other outcomes targeted by P4P programs, such as regulatory deficiencies, or other heterogeneous design features such as bonus amounts.

Our findings provide important new evidence of the heterogeneous effect of P4P in a recent, large-scale P4P program and may help assuage fears that the use of clinical quality thresholds will harm the lowest performing providers. Indeed, the large improvements in performance among the lowest performing nursing homes provide some support for the continued use of thresholds in P4P programs. However, we also found declining performance among those highest above the threshold. In light of this, continuous incentives may be more effective as they are more likely to prevent ceiling effects associated with threshold-based incentives. Although the role of cost and the underlying drivers of improvement among the lowest scoring providers is undoubtedly complicated, it is reassuring that the use of threshold-based incentives in nursing home P4P appears to have provided the intended incentives for improvement where it is most needed.

Acknowledgments

Joint Acknowledgment/Disclosure Statement: This research was supported by grant R01-AG034182 from the National Institute on Aging. Rachel Werner is supported in part by grant K24-AG047908, also from the National Institute on Aging.

Disclosures: None.

Disclaimers: None.

Notes

1. Tests for differences in pre-P4P trends of nursing home performance between P4P and comparator states are displayed in Table S1. Trends in nursing home performance were not statistically different between P4P states and comparator states for most performance measures prior to the implementation of P4P. The one exception was the use of physical restraints in Georgia, where restraint use declined slightly faster in Georgia than in Tennessee prior to implementation of P4P in Georgia.
2. In Colorado, while payment incentives began in July 2009, the thresholds were released to the participating nursing homes in July 2008. Because we expect nursing homes to respond to the thresholds at their release, we use July 2008 as the start date for P4P in that state.
3. Each long-stay resident is assessed at least quarterly in the MDS, including on admission, annually, quarterly, and for a significant status change.
4. The measures of pressure sores and physical restraint use were included in all three P4P programs; the pain measure was used in two of the three programs (see Table 1).
5. The change in performance for quantile 2 nursing homes in P4P states compared to quantile 2 nursing homes in non-P4P states is calculated as the combined effect of

the p4p coefficient with the p4p*Quantile 2 coefficient: 0.228–0.117 = 0.111 standard deviation decline for pain; 0.231–0.226 = 0.005 for pressure sores; and 0.204–0.106 = 0.098 for restraints.

6. The change in performance for quantile 6 nursing homes in P4P versus non-P4P states is: 0.228–0.369 = −0.141 standard deviation change for pain (where negative effects represent improvement); 0.231–0.548 = −0.317 for pressure sores; and 0.204–0.479 = −0.275 for restraints.

REFERENCES

Beaulieu, N. D., and D. R. Horrigan. 2005. "Putting Smart Money to Work for Quality Improvement." *Health Services Research* 40 (5 Pt 1): 1318–34.

Castle, N. G., B. S. Fogel, and V. Mor. 1996. "Study Shows Higher Quality of Care in Facilities Administered by ACHCA Members." *Journal of Long Term Care Administration* 24 (2): 11–6.

Fries, B. E., C. Hawes, J. N. Morris, C. D. Phillips, V. Mor, and P. S. Park. 1997. "Effect of the National Resident Assessment Instrument on Selected Health Conditions and Problems." *Journal of the American Geriatrics Society* 45 (8): 994–1001.

Grabowski, D. C., Z. Feng, and V. Mor. 2008. "Medicaid Nursing Home Payment and the Role of Provider Taxes." *Medical Care Research and Review: MCRR* 65 (4): 514–27.

Grant, D. P. 2010. "The Essential Economics of Thresholds: Theory and Estimation." Available at SSRN 1597331.

Institute of Medicine. 1986. *Improving the Quality of Care in Nursing Homes.* Washington, DC: National Academies Press.

Kaiser Family Foundation. 2007. "Medicaid and Long-Term Care Services and Supports" [accessed on August 20, 2008]. Available at http://www.kff.org/medicaid/upload/2186_05.pdf

Kane, R. L., G. Arling, C. Mueller, R. Held, and V. Cooke. 2007. "A Quality-Based Payment Strategy for Nursing Home Care in Minnesota." *The Gerontologist* 47 (1): 108–15.

Kane, R. L., C. C. Williams, T. F. Williams, and R. A. Kane. 1993. "Restraining Restraints: Changes in a Standard of Care." *Annual Review of Public Health* 14: 545–84.

Kidder, D., M. Rennison, H. Goldberg, D. Warner, B. Bell, L. Hadden, J. Morris, R. Jones, and V. Mor. 2002. *MegaQI Covariate Analysis and Recommendations: Identification and Evaluation of Existing Quality Indicators That Are Appropriate for Use in Long-Term Care Settings.*. Contract No. 500-95-0062 TO #4. Baltimore, MD: Abt Associates Inc.

Mor, V., O. Intrator, B. E. Fries, C. Phillips, J. Teno, J. Hiris, C. Hawes, and J. Morris. 1997. "Changes in Hospitalization Associated with Introducing the Resident Assessment Instrument." *Journal of the American Geriatrics Society* 45 (8): 1002–10.

Morris, J. N., B. E. Fries, D. R. Mehr, C. Hawes, C. Phillips, V. Mor, and L. A. Lipsitz. 1994. "MDS Cognitive Performance Scale." *Journal of Gerontology* 49 (4): M174–82.

———. 1999. "Scaling ADLs within the MDS." *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 54 (11): M546–53.

Morris, J. N., T. Moore, R. Jones, V. Mor, J. Angelelli, K. Berg, C. Hale, S. Morris, K. M. Murphy, and M. Rennison. 2003. *Validation of Long-Term and Post-Acute Care Quality Indicators.* Baltimore, MD: Centers for Medicare and Medicaid Services.

Mullen, K. J., R. G. Frank, and M. B. Rosenthal. 2010. "Can You Get What You Pay for? Pay-for-Performance and the Quality of Healthcare Providers." *RAND Journal of Economics* 41 (1): 64–91.

Neal, D., and D. W. Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92 (2): 263–83.

Nursing Home Quality Initiative. 2004. *Quality Measures Resource Manual: Enhanced Set of Quality Measures.* Baltimore, MD: Nursing Home Quality Initiative.

Park, J., and R. M. Werner. 2011. "Changes in the Relationship between Nursing Home Financial Performance and Quality of Care under Public Reporting." *Health Economics* 20: 783–801.

Rosenthal, M. B., R. G. Frank, Z. Li, and A. M. Epstein. 2005. "Early Experience with Pay-for-Performance: From Concept to Practice." *Journal of the American Medical Association* 294 (14): 1788–93.

Shorr, R. I., R. L. Fought, and W. A. Ray. 1994. "Changes in Antipsychotic Drug Use in Nursing Homes during Implementation of the OBRA-87 Regulations." *Journal of the American Medical Association* 271 (5): 358–62.

Snowden, M., and P. Roy-Byrne. 1998. "Mental Illness and Nursing Home Reform: OBRA-87 Ten Years Later. Omnibus Budget Reconciliation Act." *Psychiatric Services (Washington, DC)* 49 (2): 229–33.

Werner, R. M., R. T. Konetzka, and D. Polsky. 2013. "The Effect of Pay-for-Performance in Nursing Homes: Evidence from State Medicaid Programs." *Health Services Research* 48 (4): 1393–414.

Werner, R. M., R. T. Konetzka, and K. Liang. 2010. "State Adoption of Nursing Home Pay-for-Performance." *Medical Care Research & Review* 67: 364–77.

Wunderlich, G. S., and P. Kohler. 2000. *Improving the Quality of Long-Term Care.* Washington, DC: Division of Health Care Services, Institute of Medicine.

Supporting Information

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

Appendix SA1: Author Matrix.

Table S1: Results of Tests for Differences in Trends of Nursing Home Performance Prior to Implementation of P4P.

Table S2: Regression Results Showing the Main Effect of P4P on Nursing Home Performance for Each State Individually (Compared to Its Comparator State) and for the Three States Combined.

Table S3: Regression Results Comparing Colorado (Using the Higher Predetermined Thresholds in Their P4P Program) to Arizona.

Table S4: Regression Results Comparing Georgia (Using Average Performance in the State as the Threshold in Their P4P program) to Tennessee.

Table S5: Regression Results Comparing Oklahoma (Using Thresholds Based on Median Performance in Their P4P Program) to Arkansas.

Table S6: Regression Results Using a False P4P Implementation Date (One Year Prior to the Actual P4P Implementation in Each State).