HSR Health Services Research

DOI: 10.1111/1475-6773.12538
RESEARCH ARTICLE

The Impact of Nursing Home Pay-for-Performance on Quality and Medicare Spending: Results from the Nursing Home Value-Based Purchasing Demonstration

David C. Grabowski, David G. Stevenson, Daryl J. Caudry, A. James O'Malley, Lisa H. Green, Julia A. Doherty, and Richard G. Frank

Objective. To evaluate the impact of the Nursing Home Value-Based Purchasing demonstration on quality of care and Medicare spending.

Data Sources/Study Setting. Administrative and qualitative data from Arizona, New York, and Wisconsin nursing homes over the base-year (2008–2009) and 3-year (2009–2012) demonstration period.

Study Design. Nursing homes were randomized to the intervention in New York, while the comparison facilities were constructed via propensity score matching in Arizona and Wisconsin. We used a difference-in-difference analysis to compare outcomes across the base-year relative to outcomes in each of the three demonstration years. To provide context and assist with interpretation of results, we also interviewed staff members at participating facilities.

Principal Findings. Medicare savings were observed in Arizona in the first year only and Wisconsin for the first 2 years; no savings were observed in New York. The demonstration did not systematically impact any of the quality measures. Discussions with nursing home administrators suggested that facilities made few, if any, changes in response to the demonstration, leading us to conclude that the observed savings likely reflected regression to the mean rather than true savings.

Conclusion. The Federal nursing home pay-for-performance demonstration had little impact on quality or Medicare spending.

Key Words. Nursing homes, quality of care, pay-for-performance

Much recent policy attention has focused on the poor quality of care delivered in U.S. nursing homes (Centers for Medicare & Medicaid Services 2011; Office of Inspector General 2014). Historically, the main

government approach for ensuring acceptable levels of quality was regulation (Walshe and Harrington 2002), an emphasis reflecting skepticism that market forces alone would result in acceptable quality of care. In particular, the central role of government payment for nursing home services combined with the inability of many consumers to ascertain and monitor quality suggests an absent "business case" for providing high-quality care. Over the last 15 years, the federal government has attempted to create a business case for quality via the introduction of nursing home report cards such as the Nursing Home Compare five-star system (Konetzka et al. 2015). Another potential approach to creating a business case for quality is to institute pay-for-performance reimbursement strategies so that pecuniary concerns can be harnessed to motivate quality improvement in nursing homes. However, pay-for-performance initiatives have generally not been found to improve quality more broadly in health care (Rosenthal and Frank 2006) or in state Medicaid nursing home programs (Werner, Konetzka, and Polsky 2013).

In July 2009, the Centers for Medicare & Medicaid Services (CMS) launched a 3-year voluntary Nursing Home Value-Based Purchasing (NHVBP) demonstration in Arizona, New York, and Wisconsin to test how a performance-based reimbursement incentive impacted the quality of care. Performance was measured by hospitalization rates, quality measures, staffing, and survey inspections and was a blend of both end-of-year performance and improvement from baseline. Within each state, participating nursing homes' rankings on these scores determined the distribution of performance payments at the end of each year.

In this study, we report the results of a mixed-methods evaluation examining the impact of the NHVBP demonstration on quality of care and Medicare expenditures.

Address correspondence to David C. Grabowski, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; e-mail: grabowski@med.harvard.edu. David G. Stevenson, Ph.D., is with the Department of Health Policy, Vanderbilt University, Nashville, TN. Daryl J. Caudry, M.A., and Richard G. Frank, Ph.D., are with the Department of Health Care Policy, Harvard Medical School, Boston, MA. A. James O'Malley, Ph.D., is with The Department of Biomedical Data Science and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth College, Lebanon, NH. Lisa H. Green, Ph.D., and Julia A. Doherty, M.H.S.A., are with the L&M Policy Research, LLC, Washington, DC.

METHODS

Demonstration Design

CMS recruited nursing homes into the NHVBP demonstration via a two-step process. First, states were asked to apply for enrollment in the demonstration, resulting in four states—Arizona, Mississippi, New York, and Wisconsin—applying and all being selected for participation. Next, nursing homes in these four states were recruited to enroll in the demonstration voluntarily with the expectation that half of the facilities would be randomized to the treatment group and the other half to the control group. The goal was to recruit at least 100 nursing homes per state. Mississippi was excluded altogether from the demonstration due to insufficient enrollment. Ultimately, New York was the only state with sufficient facility enrollment to randomize, with 72 facilities assigned to the treatment group and 79 facilities assigned to the control group.

Due to the small number of facilities volunteering for the demonstration in Arizona ($N = 38$) and Wisconsin ($N = 61$), all participating facilities were enrolled in the treatment group and an identically sized comparison group was constructed via propensity score matching. CMS estimated separate logistic regression models for Arizona and Wisconsin to predict enrollment in the demonstration as a function of profit status, chain membership, hospital-based status (Wisconsin only), high Medicare census (Arizona only), total nursing hours per resident day, registered nurse hours per resident day (Wisconsin only), and whether the nursing home had a five-star health inspection rating on the Nursing Home Compare website. They matched each applicant to a nonapplicant based on nearest neighbor matching. In both states, the comparison group identified by the propensity score model was generally more similar to the treatment group than the full sample with respect to the measures included in the logistic regression models (see Table S1). Nevertheless, we acknowledge the limitation that some potentially important variables—such as the five-star rating—were still out of balance after matching.

For the purposes of the demonstration, each state was a separate "laboratory" in which to test the value-based purchasing concept. Nursing home performance was assessed using a 100-point scale with measures from four domains: nurse staffing (30 points), quality outcomes (20 points), survey deficiencies (20 points), and potentially avoidable hospitalization rates (30 points). The staffing domain allocated 10 points for registered nurse staffing, 5 points for licensed practical nurse staffing, 5 points for certified nurse aide staffing, and 10 points for overall staff turnover. In each state, nursing homes in the top

20 percent of the distribution could qualify for a reward payment based on their absolute performance level or their improvement relative to the prior year. The top 10 percent received a higher payout relative to the next decile of performers. Reward payments were allocated equally between the top absolute performers and the top improvers, and a nursing home could not receive a payout for both. All payouts were weighted by facility size.

Importantly, CMS designed the demonstration to be budget-neutral with respect to Medicare as mandated by the Office of Management and Budget. For a performance payment to be made in a particular state, improvements in quality must have resulted in a savings pool that could be used to fund the payments. Only savings that exceeded 2.3 percent of total Medicare expenditures were considered "true Medicare savings" and thus available for distribution to participants. Furthermore, the size of the performance payment pool could not exceed 5 percent of total Medicare expenditures. Providers received 80 percent of the savings above the threshold paid out until the 5 percent cap was reached. If no savings were generated for the treatment nursing homes in a state relative to the comparison group, then no incentive payment was made to any nursing home in that state regardless of any individual facility's activities or performance. High hospitalization nursing homes were ineligible for reward payments to help ensure that qualifying nursing homes contributed to the statewide savings pool. Importantly, the demonstration did not track spending by Medicaid (the dominant payer of nursing home care), Medicare Advantage, or other payers. As such, we acknowledge the limitation that we are not able to track the potential impact of the NHVBP—whether positive or negative—for other payers or non-Medicare beneficiaries.

The demonstration included all Medicare beneficiaries receiving care in participating nursing homes, even if their nursing home care was reimbursed by another payer. Medicare typically pays for short-stay, rehabilitative nursing home care, while Medicaid and private sources pay for long-stay, chronic care. The reward payments under the NHVBP were supplemental payments completely separate from the standard nursing home reimbursement system.

Data and Study Variables

This evaluation study examined all 3 years of the NHVBP: Year 1: July 1, 2009 through June 30, 2010; Year 2: July 1, 2010 through June 30, 2011; and Year 3: July 1, 2011 through June 30, 2012. We also had data for the baseline year prior to the beginning of the demonstration (July 1, 2008–June 30, 2009).

Data were obtained on expenditures and facility performance from several sources. First, Medicare fee-for-service eligibility and claims data were drawn from Medicare enrollment and claims files for all individuals residing in the treatment and control nursing homes during the baseline and all three demonstration years. Specifically, following the CMS rules for the demonstration, we calculated Medicare expenditures for these individuals based on their Medicare claims for skilled nursing facility (SNF) care, inpatient hospital care, outpatient hospital care, Part B (physician), and hospice. Home health care and durable medical equipment expenditures were excluded. Only those Medicare expenditures that occurred over the course of the nursing home stay and for up to 3 days following the end of the stay if the individual was discharged elsewhere were included. The top 1 percent of Medicare expenditures in each state was truncated in order to diminish the influence of cost outliers. Managed care enrollees and non-Medicare nursing home residents were also excluded. We adjusted Medicare spending for case mix using the CMS Hierarchical Condition Categories (HCC) model. Finally, expenditures were calculated separately for both short-stay (postacute) and long-stay (chronically ill) nursing home residents using a cutoff of 90 days of residence in the facility. As a potential limitation of the demonstration's 90-day short-stay measure, length of stay itself may be potentially endogenous to quality of care and hospitalization, and some censoring may exist if residents die.

Second, quality measures were drawn from the federally mandated Minimum Data Set (MDS) assessment instrument. The MDS was collected at time of admission and then at least quarterly thereafter for all nursing home residents. In this study, the evaluation team examined the full list of MDS-based outcomes that CMS identified and used in the demonstration to incentivize performance. For long-stay nursing home residents, the four measures included the following: the percent of residents whose need for help with activities of daily living (ADLs) had increased, the percent of high-risk residents with pressure ulcers, the percent of residents with a catheter inserted and left in their bladder, and the percent of residents who were physically restrained. For the short-stay residents, the three measures included the following: the percent of residents with improved level of ADL functioning, the percent with improved status on mid-loss ADL functioning, and the percent with failure to improve bladder incontinence. The long-stay measures mirrored those reported on Nursing Home Compare, the federal nursing home report card website on Medicare.gov, while the short-stay quality measures used in the NHVBP demonstration were distinct from those short-stay measures reported on Nursing Home Compare.

Third, Medicare claims were used to calculate potentially avoidable hospitalization rates for both short-stay and long-stay nursing home residents. A large literature has suggested that a substantial portion of hospital admissions of nursing home residents can be avoided through careful management of these conditions in the nursing home (O'Malley et al. 2007; Grabowski et al. 2008; Ouslander et al. 2010). Under the demonstration, "potentially avoidable" cases were defined as hospitalizations with any of the following primary or secondary diagnoses: coronary heart failure, electrolyte imbalance, respiratory disease, sepsis, urinary tract infection, and anemia (long-stayers only). Short stayers were defined based on episodes of fewer than 90 days, and we calculated the rate of hospitalizations per nursing home stay for this population. Long-stayers were defined as individuals with a nursing home episode greater than 90 days and the rate of hospitalization was calculated per 100 resident days. Hospitalizations that occurred up to 3 days after the end of the nursing home stay were included. Due to data limitations, the hospitalization measures for Year 3 of the demonstration could not be constructed. Although both the short- and long-stay hospitalization rates were risk-adjusted for medical acuity, functional impairment, and the frailty of nursing home residents using a series of measures from the claims and the MDS, we acknowledge that there still may be some remaining unobserved risk across the treatment and comparison groups. A full description of the risk adjustment is described elsewhere (Centers for Medicare & Medicaid Services 2012).

Fourth, the final two demonstration performance measures were drawn from the Online Survey, Certification, and Reporting (OSCAR) system. Collected and maintained by the CMS, the OSCAR data included information about whether nursing homes were in compliance with federal regulatory requirements. Every facility is required to have an initial survey to verify compliance. Thereafter, states were required to survey each facility no less often than every 15 months, with an average of about 12 months. Deficiencies are recorded in OSCAR when facilities are found to be out of compliance with federal regulatory standards. Each deficiency was categorized by the surveyor into one of 17 areas and rated by its scope and severity (on an "A" to "L" scale in order of increasing severity). In this paper, we report the total raw number of deficiencies, the number of deficiencies weighted by scope and severity, and deficiencies from complaint surveys. Staffing information from OSCAR was also analyzed and included registered nurses per resident day, licensed practical nurses per resident day, and certified nurse aide hours per resident day. Under the NHVBP,

treatment facilities reported staffing information from payroll data. Summary information from these payroll data for the treatment facilities is presented as part of this study as a check on the accuracy of the OSCAR system staffing data.

Finally, we obtained data on a range of potential covariates from the OSCAR, including payer mix, ownership status, membership in a continuing care retirement community (CCRC), chain membership, hospital-based affiliation, case mix, number of beds, and urban location.

Statistical Analysis

This study employed a "difference-in-differences" methodology, which compared the pre-post difference in the introduction of the demonstration in the treatment groups relative to the pre-post difference in the comparison groups. Thus, the model specification was as follows:

$$Y_{it} = \beta_1 \text{TREAT}_i * \text{POST}_{it} + \beta_2 \text{TREAT}_i + \beta_3 \text{POST}_t + \gamma X_{it} + \varepsilon_{it} \qquad (1)$$

where Y was an outcome for nursing home i at time t, TREAT was an indicator of enrollment in the treatment arm of the demonstration, POST was a dummy variable for postintervention, TREAT*POST was an interaction of the treatment and postintervention indicators, X was a vector of covariates, and ε was the randomly distributed errors. Once again, we estimated this model separately for each state and study year. Importantly, the same baseline year (July 1, 2008–June 30, 2009) was used to evaluate performance in demonstration Year 1 (2009–2010), Year 2 (2010–2011), and Year 3 (2011–2012). Unfortunately, we were not able to examine longer trends in performance prior to the demonstration. The key parameter of interest was β_1, the interaction term between the treatment and postintervention indicators, which provides us with any estimate of the pre-post difference in the treatment group relative to the pre-post difference in the comparison group over this same time period. The quality models, which controlled for the OSCAR-based covariates (X) discussed in the previous section, were estimated using least squares regression.

For the expenditure results, we replicated the approach taken by CMS in calculating potential savings by adjusting the spending total using the CMS HCC model and then running the differences-in-differences model without the facility-level covariates listed above. We acknowledge the possible limitation that there may be some remaining facility-level

differences across the treatment and control groups underlying the expenditure results.

Qualitative Analysis

The research team conducted interviews with a subset of participating providers during all 3 years of the demonstration in order to solicit contextual detail around how the demonstration was perceived and its influence on facility decision making and practice patterns. We conducted a series of 1-hour phone discussions using a semistructured protocol during each of the study years. In the first and second study years, we conducted 28 discussions with participating nursing home administrators across the three states (nine in both Arizona and Wisconsin, and 10 in New York). In the third year, the distribution of interviews was changed to allow for greater focus on soliciting information on what factors drove savings and quality improvement in Wisconsin in Year 2 of the demonstration. We spoke with 20 facilities in Wisconsin and reduced the number for Arizona and New York to five each. Given the open-ended nature of these conversations, some variation was present in the topics and issues covered across interviews. In general, however, discussions focused on the facility's perceptions of the demonstration, its impact on quality improvement and organizational activities, and any changes resulting from the demonstration.

The discussions engaged a combination of nursing home administrators, directors of nursing, and other staff involved with quality improvement activities or data submission for the demonstration. A senior team member led each discussion and was supported by a designated note-taker, who prepared transcript-style notes and coded text segments using a code-tree that mirrored the discussion protocol. We utilized Dedoose, a relational Web-based tool that facilitates computer-assisted qualitative data analysis, to house the notes, apply the codes to the text, and organize the text data. Given the very different state contexts within which the demonstration was situated, the team took a case-oriented approach to analyze the information collected, where each state reflected a case. We thus organized the coded text segments into tables arrayed by state and other facility characteristics, and three team members separately reviewed these tables to identify patterns and themes. These findings were then shared with the broader qualitative team to corroborate results, evaluate alternative explanations, and identify any negative cases disproving the patterns and themes identified in order to arrive at the qualitative data-supported conclusions.

RESULTS

Sample Characteristics

Based on observable characteristics, the treatment and comparison groups were roughly balanced at baseline within states (see Table 1). In general, large differences were not present in chain membership, hospital-based status, CCRC membership, ownership type (for-profit, nonprofit, government), payer mix, case mix, size, and location in urban areas. The treatment group in Arizona had a greater share of nonprofit facilities and facilities within a CCRC. Treatment facilities in New York had fewer Medicaid recipients, while the treatment facilities in Wisconsin were more likely to be for-profit and have a lower severity-adjusted deficiencies score.

The characteristics of nursing homes varied considerably across the participating states. For example, chain membership and for-profit ownership was highest in Arizona and lowest in New York State. Wisconsin had slightly fewer Medicaid residents, more rural facilities, and lower acuity residents overall. Finally, New York had much larger facilities on average that were more likely to be located in urban areas.

Average Medicare spending in the baseline period for long-stay residents (per day) and short-stay residents (per stay) is presented in Figure 1. Medicare spent $10,067 per short-stay episode in Arizona, $12,505 in New York, and $9,611 in Wisconsin. The bulk of the short-stay residents' spending was driven by SNF (ranging from 61.4 percent in New York to 74.6 percent in Wisconsin) and inpatient (16.6 percent in Wisconsin to 29.7 percent in New York) services. For long stays, Medicare spent $101 per long-stay day in Arizona, $83 in New York, and $56 in Wisconsin. The major Medicare spending categories among long-stay residents were SNF (ranging from 33.4 percent in New York to 43.3 percent in Wisconsin) and inpatient (24.1 percent in Wisconsin to 42 percent in New York) services.

Medicare Spending

CMS estimated potential risk-adjusted Medicare savings under the NHVBP demonstration using a differences-in-differences approach for each state (see Table 2). Overall, only three of the nine state-years indicated sufficient savings to generate a reward payment to the top-performing nursing homes: Arizona and Wisconsin in Year 1, and Wisconsin in Year 2. In Year 1, the top nursing homes in Arizona and Wisconsin received a reward payment. In Arizona,

Table 1: Mean Characteristics of Participating Nursing Homes at Baseline: Treatment versus Comparison Facilities

	Arizona		*New York*		*Wisconsin*	
	Treatment	*Comparison*	*Treatment*	*Comparison*	*Treatment*	*Comparison*
Chain (%)	0.71	0.73	0.21	0.18	0.44	0.47
Hospital-based (%)	0	0.07	0.10	0.10	0.03	0.05
CCRC (%)	0.16	0.12	0.03	0.04	0.05	0.06
For-profit (%)	0.74	0.93	0.43	0.42	0.48	0.44
Nonprofit (%)	0.26	0.07	0.53	0.53	0.39	0.39
Government (%)	0	0	0.04	0.05	0.13	0.18
Medicaid (%)	0.67	0.62	0.64	0.68	0.59	0.59
Medicare (%)	0.12	0.16	0.14	0.14	0.14	0.13
Other payer (%)	0.20	0.22	0.22	0.18	0.27	0.28
Acuity score	10.55	10.31	10.61	10.60	9.53	9.62
ADL score	4.05	3.99	4.21	4.20	3.92	3.86
Total residents	97.21	95.93	198.46	197.63	89.21	83.42
Urban (%)	0.76	0.78	0.96	0.92	0.48	0.47
Deficiencies, raw count	12.42	11.63	3.78	4.16	5.02	7.10
Deficiencies, severity-adjusted score	76.26	63.46	23.96	30.19	38.10	67.05
RN hours/resident day	0.32	0.34	0.41	0.38	0.54	0.58
LPN hours/resident day	0.93	0.95	0.81	0.81	0.54	0.54
Nurse aide hours/resident day	2.07	2.04	2.31	2.19	2.28	2.32
Long-stay ADL worsening	15.14	12.73	15.90	14.81	16.29	14.83
Long-stay pressure ulcers, high risk	13.20	11.12	12.44	12.88	9.06	9.68
Long-stay catheters	5.63	6.58	3.88	4.71	6.10	7.09
Long-stay restraints	3.52	2.40	2.36	2.87	1.32	1.37
Short-stay failure to improve incontinence	0.51	0.49	0.53	0.54	0.54	0.53
Short-stay ADL improvement	0.41	0.37	0.35	0.37	0.38	0.34
Short-stay mid-loss ADL improvement	0.42	0.37	0.33	0.37	0.37	0.36
Long-stay avoidable hospitalization rate	0.21	0.18	0.19	0.19	0.12	0.12
Short-stay avoidable hospitalization rate	0.15	0.15	0.18	0.20	0.12	0.14
N	38	41	72	79	61	62

ADL, activity of daily living; CCRC, continuing care retirement community.

Figure 1: Medicare Spending in Baseline Period by State [Color figure can be viewed at wileyonlinelibrary.com]

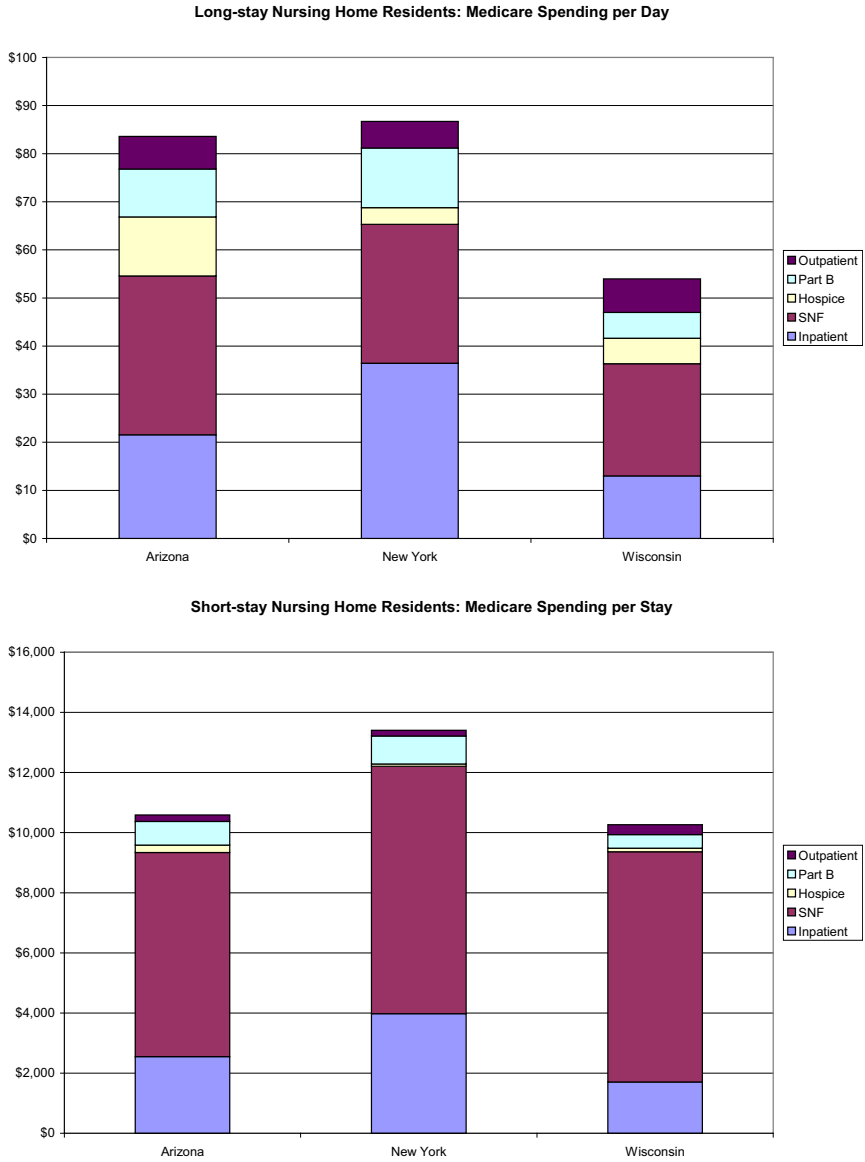

Table 2: Differences in Risk-Adjusted Medicare Spending for Long-Stay and Short-Stay Residents across Treatment and Comparison Nursing Homes

	Arizona		*New York*		*Wisconsin*	
	Comparison	*Treatment*	*Comparison*	*Treatment*	*Comparison*	*Treatment*
Long-stay residents						
Base-year	$75.29	$90.85	$83.97	79.92	$49.42	$55.73
Year 1	$81.12	$93.28	$85.61	83.29	$52.78	$53.38
Year 2	$113.85	$137.41	$110.78	112.87	$77.73	$76
Year 3	$96.95	$108.68	$94.11	94.65	$55.92	$58.15
Short-stay residents						
Base-year	$10,753	$10,152	$13,122	$12,740	$10,409	$10,151
Year 1	$10,589	$9,916	$13,300	$12,839	$9,848	$9,831
Year 2	$13,016	$14,085	$13,598	$13,970	$10,993	$10,777
Year 3	$12,037	$11,281	$14,359	$14,490	$10,446	$10,338

Notes. The spending total is risk-adjusted using the CMS Hierarchical Condition Categories (HCC) model.

Medicare spending per day among long-stay residents increased $2.44 (or 2.68 percent) in the treatment group and $5.83 (or 7.74 percent) in the control group. Over the 326,618 long-stay days in the treatment group, these findings suggested $1.5 million in savings. For short-stay residents, Medicare expenditures per stay decreased $236.28 (or 2.33 percent) in the treatment group and $164.54 (or 1.53 percent) in the control group. Over 5,079 short-stay episodes, the treatment facilities realized $411,071 in savings. Inpatient expenditures (8.1 percent reduction) were the main driver of short-stay savings, while reduced SNF and hospice use were the main drivers of long-stay savings. In Arizona, the 38 treatment facilities realized $1,912,143 in Year 1 savings; this amount was only slightly above the 2.3 percent savings threshold set by CMS, meaning just $27,032 was distributed to the 12 highest performing facilities. The average payout amount in Arizona was $2,253, ranging from $802 to $3,810 per nursing home.

In Wisconsin in Year 1, Medicare spending per day among long-stay residents decreased $2.35 (or 4.22 percent) in the treatment group and increased $3.36 (or 6.79 percent) in the control group, suggesting $8,516,701 in savings over the 1,387,474 long-stay days. This decrease was predominantly driven by a reduction in SNF (−15.2 percent) and inpatient (−11.3 percent) spending. Medicare spending per short-stay episode decreased $320.38 (or 3.16 percent) in the treatment group and $560.79 (or 5.39 percent) in the control group, suggesting $1,062,109 in increased Medicare expenditures over 4,688 short-stay episodes. Thus, the estimated overall Year 1 Medicare savings totaled

$7,454,591, of which almost $3.5 million was distributed to the 19 highest performing facilities. On average, these high-performing facilities received a payment of $183,371, ranging from a low payout of $39,281 to a high payout of $369,970.

In Wisconsin in Year 2, spending per day among long-stay residents increased $19.78 (or 35.67 percent) in the treatment group and increased $24.11 (or 47.21 percent) in the control group, suggesting $4,166,583 in savings over the 778,031 long-stay days. Medicare spending per short-stay episode increased $703.25 (or 6.89 percent) in the treatment group and $1,172.57 (or 11.50 percent) in the control group, suggesting $1,834,229 in Medicare savings over 4,965 short-stay episodes. Thus, the estimated overall Year 2 Medicare savings realized by the treatment group totaled $6,000,812, of which roughly $3 million was distributed to the 17 highest performing facilities. On average, these high-performing facilities received a payment of $171,789, ranging from a low payout of $65,519 to a high payout of $361,369.

In New York, the participating providers never generated sufficient savings to qualify for a reward payment under the demonstration. In Year 2, the treatment facilities generated almost a half-million dollars in Medicare savings; however, because this amount was below the 2.3 percent savings threshold ($9,582,775), the top-performing nursing homes in New York received no payout.

Quality of Care

Using a differences-in-differences regression framework (see Table 3), we examined the effect of the NHVBP demonstration on a range of quality measures. Only three of the 108 quality regressions generated statistically significant ($p < .05$) results. Specifically, the severity-adjusted deficiencies score was higher in the treatment group in Wisconsin in demonstration Years 2 and 3, while the count of deficiencies in Wisconsin was higher in the treatment group Year 3. Importantly, if we were to make a correction for the multiple comparisons issue, these three statistically significant findings would no longer be significant at conventional levels.

To evaluate the degree of precision in our estimates, we multiplied our standard error values by ± 1.96 to get an estimate of what effect size would have counterfactually been significant. For example, when we examined the impact of the NHVBP on the count of deficiencies in New York in Year 1, we would have to obtain an effect size of 1.65 fewer deficiencies ($=-1.96*0.84$) to achieve statistical significance. Relative to the dependent variable mean, 1.65

Table 3: Differences-in-Differences: Quality Regression Results

	Arizona			*New York*			*Wisconsin*		
Outcomes	*Year 1*	*Year 2*	*Year 3*	*Year 1*	*Year 2*	*Year 3*	*Year 1*	*Year 2*	*Year 3*
Deficiencies, count	0.55 (2.46)	−0.05 (1.05)	−0.14 (0.70)	−0.57 (0.84)	0.30 (0.38)	−0.21 (0.30)	1.87 (1.22)	0.94 (0.57)	0.89* (0.45)
Deficiencies, severity-adjusted score	4.21 (23.04)	−3.57 (9.18)	−1.84 (5.99)	1.97 (17.15)	−2.89 (5.50)	−3.80 (4.79)	28.64 (15.06)	14.45* (6.68)	12.01* (5.41)
RN hours/ resident day	−0.023 (0.074)	0.014 (0.04)	0.007 (0.027)	−0.018 (0.043)	−0.014 (0.023)	−0.009 (0.015)	−0.010 (0.082)	0.014 (0.036)	0.012 (0.024)
LPN hours/ resident day	−0.17 (0.11)	−0.035 (0.065)	−0.023 (0.043)	0.023 (0.056)	0.026 (0.028)	0.018 (0.018)	−0.078 (0.066)	−0.038 (0.033)	−0.026 (0.022)
Nurse aide hours/ resident day	−0.10 (0.22)	−0.002 (0.134)	−0.002 (0.090)	−0.038 (0.106)	−0.022 (0.051)	−0.014 (0.034)	−0.011 (0.132)	0.028 (0.072)	0.021 (0.048)
Long-stay ADL worsening	−0.006 (0.021)	−0.011 (0.012)	−0.004 (0.009)	−0.010 (0.013)	−0.009 (0.007)	−0.006 (0.004)	−0.022 (0.013)	−0.0001 (0.0075)	−0.006 (0.005)
Long-stay pressure ulcers, high risk	−0.018 (0.027)	−0.014 (0.013)	−0.010 (0.009)	0.006 (0.011)	0.0015 (0.0059)	−0.001 (0.003)	0.007 (0.012)	0.005 (0.006)	0.0001 (0.0037)
Long-stay catheters	0.016 (0.018)	0.006 (0.008)	0.007 (0.006)	0.002 (0.007)	0.0006 (0.0032)	0.001 (0.002)	−0.006 (0.009)	0.0005 (0.0043)	−0.0003 (0.0028)
Long-stay restraints	−0.0063 (0.0100)	−0.004 (0.005)	−0.002 (0.003)	−0.001 (0.007)	−0.0017 (0.0032)	0.0005 (0.0021)	0.0020 (0.0047)	0.0026 (0.0023)	0.0017 (0.0017)
Short-stay failure to improve incontinence	0.023 (0.052)	−0.009 (0.026)	−0.033 (0.017)	0.029 (0.033)	0.016 (0.016)	0.013 (0.011)	−0.022 (0.038)	−0.019 (0.021)	−0.015 (0.013)
Short-stay ADL improvement	−0.052 (0.044)	−0.024 (0.025)	0.009 (0.015)	0.014 (0.031)	−0.011 (0.017)	−0.010 (0.010)	−0.018 (0.035)	−0.014 (0.016)	0.002 (0.011)
Short-stay mid-loss ADL improvement	0.013 (0.042)	−0.006 (0.023)	0.008 (0.015)	0.024 (0.029)	−0.004 (0.015)	−0.0006 (0.0099)	0.011 (0.031)	0.012 (0.017)	−0.00004 (0.0108)
Long-stay avoidable hospitalization rate	−0.020 (0.037)	−0.0011 (0.018)	—	0.0092 (0.017)	0.0006 (0.0094)	—	−0.000002 (0.0176)	−0.0040 (0.0080)	—
Short-stay avoidable hospitalization rate	−0.005 (0.020)	−0.0098 (0.010)	—	0.0144 (0.0179)	0.009 (0.009)	—	0.0169 (0.0175)	−0.0018 (0.0081)	—
N	158			302			244		

Notes. Standard errors are presented in parentheses. ADL, activity of daily living.
*Statistically significant at 5% level. All regressions include the variables listed in the upper panel of Table 1.

fewer deficiencies would constitute a 40 percent decline in deficiencies; yet most would consider a range of values less than 1.65 to be relevant for both policy makers and clinicians. Thus, we acknowledge the limited precision in many of our estimates due to the small sample size.

Qualitative Findings

In our discussions with nursing homes across all three states, and across all years of the demonstration, administrators and DONs explained that most facility changes in areas targeted by the demonstration reinforced internal priorities and areas of focus and were attributable to the increasing external pressures to contain costs and improve quality. A Wisconsin administrator said that the demonstration "was rewarding quality that was already being provided . . . reducing readmission, surveys, those were focused on initiatives that were in place anyways. Whether the payout happened or not, we would be focused on same things." An administrator in New York agreed, saying, "Nothing was done specifically for the demonstration; however, demonstration issues are things that are covered every month, like restraints, catheters, etc."

With respect to the experience of participating in the demonstration, several nursing home administrators noted the burden of data collection, particularly early on in the demonstration, and the significant lag in receiving quality reports. For some nursing homes, the CMS quality reports received through participation in the demonstration were the only source of these benchmarking-type quality metrics, so they found them useful despite the delays in receipt. One facility even mentioned using data from the demonstration showing their high ranking within the state as they approached ACOs to present a case for inclusion on their preferred provider list. In general, nursing homes were not only interested in benchmarking but also in receiving more information about best practices and other suggestions for how to continuously improve around the demonstration performance metrics.

Quality reports were provided once a year and reflected information that was sometimes as much as 18 months old, making it difficult to predict likelihood of payment or target-specific areas for improvement during each subsequent demonstration year. One administrator hypothesized that his good scores were more likely the result of the luck, because they had not had many patients' conditions worsen severely or family members pushing for residents to be admitted to a hospital when they became ill that year.

One stakeholder summed up the feelings of many demonstration participants saying, "This was an absolute missed opportunity." This stakeholder felt that their nursing home association had a lot to offer in terms of leadership but was only enlisted to help recruit nursing homes and was not consistently invited to listen in on the quarterly demonstration calls. In general, administrators and stakeholders felt that the demonstration was a good idea but lacked the necessary communication, direction, and leadership to really impact quality measures.

Rather than being incented to change practices because of the possibility of a payout, many facilities saw the demonstration as a reinforcement of actions they were already planning to take or had already begun implementing. Most nursing homes did not change their actions because of the demonstration; rather, some hoped to be rewarded for things that they were already doing or thought their involvement in the demonstration would just be an opportunity to learn from other nursing homes, or prepare for what is to come from CMS, moving forward.

Although there were some outliers, these impressions did not vary notably across key facility characteristics, such as size or profit status. The Wisconsin and Arizona facilities, however, were more likely than the New York facilities to report being actively engaged in the quality improvement activities reinforced by (though not motivated by) the demonstration and felt that they were operating at a higher than average level of efficiency as a result of a long history of conducting these activities. The qualitative analyses indicated very little direct effort on the part of demonstration facilities toward improving quality and lowering Medicare expenditures in direct response to the demonstration.

DISCUSSION

Based on our mixed-methods evaluation, we concluded that the demonstration did not directly lower Medicare spending nor did it improve quality for nursing home residents. Two important questions emanate from this conclusion. First, how did Arizona (Year 1) and Wisconsin (Year 1 and 2) generate savings if nursing homes generally did not respond directly to the NHVBP demonstration? And, second, why did the treatment facilities appear not to respond to the payment incentives under the NHVBP demonstration?

The answer to the first question might relate to the design of the demonstration. New York was the only state in which facilities that applied to

participate were randomized across the treatment and comparison groups. Thus, observed savings in Arizona and Wisconsin may reflect differences between treatment facilities and comparison groups selected by propensity scores in these states. Indeed, the difference in base-year Medicare spending for long-stay residents between the treatment and comparison facilities was much larger in Arizona and Wisconsin than in New York. Specifically, long-stay spending per day in Arizona was $15.56 (20.7 percent) higher in the treatment group in the base-year, while it was $6.31 (12.8 percent) higher in Wisconsin. By comparison, base-year spending for long-stayers in New York was $4.05 (4.8 percent) lower per day in the treatment group. Thus, the observed savings in Arizona and Wisconsin may simply reflect a "regression toward the mean." That is, when a variable has an extreme value on its first measurement, it will tend to be closer to the average on its second measurement. That attributed as savings due to the demonstration may have simply reflected relatively higher baseline spending in the treatment facilities.

Toward the second question of why treatment facilities did not appear responsive to NHVBP incentives, nursing homes may not have altered behaviors under the demonstration for a variety of reasons. First, incentive-based payment systems work well when providers understand how effort links to performance and ultimately to a reward payment. The NHVBP demonstration had a very complex payment and reward system based on a number of measures and relative and absolute performance standards. Nursing homes may not have understood or been able to predict how their efforts toward improving quality would result in a better performance score and ultimately a reward payment.

Second, the size of potential reward payments under a pay-for-performance program inevitably influences providers' response (Werner and Dudley 2012). Once again, the top performing nursing homes received 80 percent of the savings between 2.3 and 5 percent of total Medicare expenditures. CMS put these sharing rules in place—especially the 2.3 percent threshold—to ensure that any payments made to facilities reflected true savings on the part of the participating nursing homes and not chance differences. However, by applying these sharing rules, the payouts under the demonstration may have been too small to incentivize major changes in quality. Of the $15.4 million in relative savings achieved by facilities in Arizona and Wisconsin in Years 1 and 2, over $8 million was retained by CMS under the demonstration's shared savings rules.

Third, a well-designed incentive system minimizes uncertainty among participating providers as to the likelihood that their efforts under the program

will result in a reward payment. Under the demonstration, a payout was made only if the treatment nursing homes as a whole generated savings relative to the comparison facilities in that state-year period. Thus, in the context of this uncertainty, many nursing homes may have decided not to act in direct response to the demonstration because their likelihood of a payout depended on other nursing homes in the state generating savings.

Fourth, facilities are likely most responsive to real-time payouts that allow them to recoup quality improvement investments relatively quickly. Yet, due in part to the use of administrative data to determine savings and performance, payouts to top-performing nursing homes took up to 18 months, potentially lowering the salience of any potential rewards to treatment facilities. Moreover, the corresponding lag in feedback to the participating facilities on their performance during the demonstration discouraged facilities from benchmarking their performance against their own prior performance or their peers.

Fifth, many researchers have argued that incentive payments do not work well in the context of complicated tasks (Gneezy, Meier, and Rey-Biel 2011). The idea is that poor performance relates both to misaligned payment and also to a lack of on-the-ground knowledge on how to improve performance. The NHVBP demonstration was designed to address misaligned incentives, but nursing homes may still have lacked the infrastructure and expertise to improve performance. This issue, along with the limited resources available to direct toward significant operational changes without a guarantee of a reward for their investment, was corroborated during qualitative discussions with many nursing home administrators by the study team. As CMS intended, the demonstration provided little guidance and education to nursing homes as to how to improve quality outside of quarterly update calls for participating facilities. The rationale for this decision was that the demonstration was designed to encourage broad innovation on the part of the participating nursing homes. Moreover, considering the logistics of an eventual national program launch, CMS is limited somewhat in the extent they can educate and guide 16,000 nursing homes nationwide.

Finally, it is important to consider the broader policy context in which the NHVBP demonstration occurred. In particular, the 2009–2012 demonstration period saw bolstered emphasis on nursing home public reporting with the Five-Star Quality Rating System (which focused on many of the same performance measures) and a number of changes put in place with the 2010 passage of the Affordable Care Act (ACA). These ACA-related changes included a more intense focus on hospital readmission and development of several

delivery-system innovation programs such as accountable care organizations and the bundled payments for care improvements initiative. Although these changes presumably would have affected treatment and control facilities alike, their collective impact could have swamped any changes related to the NHVBP.

All of these factors may have contributed to the limited quality improvement and savings found under the demonstration. The results might convey more about specific design features of the NHVBP demonstration rather than the potential of nursing home pay-for-performance more generally. As the Medicare program moves forward with the pay-for-performance concept in the nursing home setting (e.g., current statute requires establishment of a SNF value-based purchasing program in the coming years), it should consider changes to optimize the response to payment incentives to improve quality. Modifications to the design of any future nursing home pay-for-performance program might include the following: (1) simplified payment and reward rules that link facility effort, the performance scores, and the likelihood of payout; (2) increased payout pools; (3) relaxation/elimination of budget neutrality restrictions such that the likelihood of payout does not hinge on the efforts of other participating facilities; (4) offering more immediate payouts relative to when performance gains occur; (5) real-time feedback on performance and quality activity results; and (6) providing increased education and guidance on best practices to providers.

With respect to this last point, the program could become more prescriptive by mandating that participating providers undertake specific training or best practices to qualify for a reward payment. Toward this end, a recent nursing home pay-for-performance program in Minnesota was structured around a provider-initiated quality improvement approach rather than incentivizing performance based on different outcomes. That is, nursing homes propose targeted, 1- to 3-year quality improvement projects to the state for funding, with nursing homes at risk of losing 20 percent of the funds if the project objectives are not achieved. An evaluation suggested the Minnesota program has broadly improved nursing home performance (Arling et al. 2013).

Many evaluations of CMS demonstrations are not published in the peer-reviewed literature (Grabowski 2006). Instead, the results of these demonstrations are typically only available in a final report on the CMS website, thus going underpublicized in the broader research community. To increase overall awareness of the demonstrations, CMS should be encouraged to make submission to a peer-reviewed journal a necessary step in the

independent evaluation of these demonstrations. This requirement also has the potential to improve the quality of the research.

One challenge to publishing research conducted under CMS-funded evaluations in the peer-reviewed literature is the limited time CMS allows researchers under their Data Use Agreement (DUA). With most funders, researchers can extend a DUA until the research is published. With a CMS-funded project, however, researchers can only extend the DUA for a limited period following the end of the funding period, after which the researchers lose access to the study data. Given the standard review period at most journals, this limited window of time may not allow the study team sufficient time to address multiple rounds of reviewer comments, if necessary. Indeed, we lost access to study data toward the latter part of the review process for this manuscript, which limited our ability to respond to some reviewer comments. Moving forward, the benefits to research seem to outweigh the limited risk of keeping the DUA open for a longer period.

In sum, the NHVBP tested whether a nursing home pay-for-performance program could improve quality while also generating savings for Medicare. The program was not found to improve quality of care, and while some Medicare savings were achieved by the participating facilities, it is unclear how much, if any, of these could be attributed to the demonstration. Our qualitative analyses suggested that the participating nursing homes engaged in few direct activities to lower Medicare spending. When we combine these qualitative results with the high base-year spending in Arizona and Wisconsin, we concluded that the observed savings likely reflected a "regression to the mean" rather than true savings for the program. Future CMS nursing home pay-for-performance initiatives can address some of the potential design flaws inherent in the NHVBP in order to encourage the intended outcomes.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was prepared under CMS Contract HHSM-500-2006-0009i/TO 7 by L&M Policy Research, LLC and its partner Harvard Medical School. The opinions presented here are those of the author and do not necessarily represent the views or policies of the Centers for Medicare & Medicaid Services. David Stevenson's time on the project was supported in part by a Career Development Award from the National Institutes of Health (grant no. NIA K01 AG038481).

Disclosures: None.
Disclaimers: None.

REFERENCES

Arling, G., V. Cooke, T. Lewis, A. Perkins, D. C. Grabowski, and K. Abrahamson. 2013. "Minnesota's Provider-Initiated Approach Yields Care Quality Gains at Participating Nursing Homes." *Health Affairs* 32 (9): 1631–8.

Centers for Medicare & Medicaid Services. 2011. "Center for Strategic Planning, Policy and Data Analysis Group Policy Insight Report: Dual Eligibles and Potentially Avoidable Hospitalizations" [accessed on October 27, 2011]. Available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/downloads/segal_policy_insight_report_duals_pah_june_2011.pdf

Centers for Medicare & Medicaid Services. 2012. "Risk Adjustment for Hospitalization Measures: Nursing Home Value Based Purchasing (NHVBP) Demonstration" [accessed on October 28, 2012]. Available at https://innovation.cms.gov/Files/x/NHP4P-Hospitalization-Risk-Adjustment.pdf

Gneezy, U., S. Meier, and P. Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4): 191–210.

Grabowski, D. C. 2006. "The Cost-Effectiveness of Noninstitutional Long-Term Care Services: Review and Synthesis of the Most Recent Evidence." *Medical Care Research and Review* 63 (1): 3–28.

Grabowski, D. C., K. A. Stewart, S. M. Broderick, and L. A. Coots. 2008. "Predictors of Nursing Home Hospitalization: A Review of the Literature." *Medical Care Research and Review* 65 (1): 3–39.

Konetzka, R. T., D. C. Grabowski, M. C. Perraillon, and R. M. Werner. 2015. "Nursing Home 5-Star Rating System Exacerbates Disparities in Quality, by Payer Source." *Health Affairs* 34 (5): 819–27.

Office of Inspector General. 2014. *Adverse Events in Skilled Nursing Facilities: National Incidence among Medicare Beneficiaries.* Washington, DC: U.S. Department of Health and Human Services, OEI-07-06-00540.

O'Malley, A. J., E. R. Marcantonio, R. L. Murkofsky, D. J. Caudry, and J. L. Buchannan. 2007. "Deriving a Model of the Neccessity to Hospitalize Nursing Home Residents." *Research on Aging* 29 (6): 606–25.

Ouslander, J. G., G. Lamb, M. Perloe, J. H. Givens, L. Kluge, T. Rutland, A. Atherly, and D. Saliba. 2010. "Potentially Avoidable Hospitalizations of Nursing Home Residents: Frequency, Causes, and Costs." *Journal of the American Geriatrics Society* 58 (4): 627–35.

Rosenthal, M. B., and R. G. Frank. 2006. "What Is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review* 63 (2): 135–57.

Walshe, K., and C. Harrington. 2002. "Regulation of Nursing Facilities in the United States: An Analysis of Resources and Performance of State Survey Agencies." *Gerontologist* 42 (4): 475–87.

Werner, R. M., and R. A. Dudley. 2012. "Medicare's New Hospital Value-Based Purchasing Program Is Likely to Have Only a Small Impact on Hospital Payments." *Health Affairs* 31 (9): 1932–40.

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.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.
Table S1. Results of Propensity Score Matching.