HHS Public Access

Author manuscript

Am J Health Econ. Author manuscript; available in PMC 2018 March 26.

Published in final edited form as:

Am J Health Econ. 2018 ; 4(1): 105–130. doi:10.1162/ajhe_a_00095.

Incentive Design and Quality Improvements: Evidence from State Medicaid Nursing Home Pay-for-Performance Programs

R. Tamara Konetzka[a], **Meghan M. Skira**[b,1], and **Rachel M. Werner**[c,d]

[a]Department of Public Health Sciences, University of Chicago

[b]Department of Economics, University of Georgia

[c]Division of General Internal Medicine, University of Pennsylvania

[d]Center for Health Equity Research and Promotion, Crescenz VA Medical Center Philadelphia, PA

Abstract

Pay-for-performance (P4P) programs have become a popular policy tool aimed at improving health care quality. We analyze how incentive design affects quality improvements in the nursing home setting, where several state Medicaid agencies have implemented P4P programs that vary in incentive structure. Using the Minimum Data Set and the Online Survey, Certification, and Reporting data from 2001 to 2009, we examine how the weights put on various performance measures that are tied to P4P bonuses, such as clinical outcomes, inspection deficiencies, and staffing levels, affect improvements in those measures. We find larger weights on clinical outcomes often lead to larger improvements, but small weights can lead to no improvement or worsening of some clinical outcomes. We find a qualifier for P4P eligibility based on having few or no severe inspection deficiencies is more effective at decreasing inspection deficiencies than using weights, suggesting simple rules for participation may incent larger improvement.

Keywords

pay-for-performance; nursing home quality; long-term care; incentive design

I. INTRODUCTION

Consistent with current policy goals to shift from “paying for quantity” to “paying for quality,” pay-for-performance (P4P) incentive programs have become a popular policy tool aimed at improving health care quality in the United States. P4P provides a direct link between health care provider payment and quality of care and, as a result, attempts to focus provider attention on quality in lieu of or in addition to quantity of services provided. Typical P4P programs pay health care providers a bonus for performing well on one or more quality metrics, such as the provider’s rate of providing recommended care (e.g. influenza

[1]Corresponding author: skira@uga.edu. Department of Economics, Terry College of Business, University of Georgia, Athens, GA 30602.

JEL Codes: I11, I18

vaccination or cancer screening) or the provider's patients' outcomes (e.g. control of blood pressure or diabetes).

However, there is mixed evidence that P4P improves health care quality, with studies often finding little to no impact of these incentive programs across a variety of health care settings. For reviews of this literature, see Petersen et al. (2006); Rosenthal and Frank (2006); Mehrotra et al. (2009); Van Herck et al. (2010); Emmert et al. (2012); Eijkenaar (2013); Eijkenaar et al. (2013). The complexity of P4P design may be partly responsible, making it difficult for providers to improve in all the areas targeted by P4P. Because quality is multidimensional and no single measure captures the full breadth of provider quality, programs have increasingly turned to tying P4P incentives to an expanding number of performance measures. This is conceptually appealing as large numbers of measures may better capture information about a provider's underlying quality than a small number of measures. Furthermore, compared to programs that only reward one or very few metrics, it may prevent a disproportionate focus on a specific dimension of performance. However, as the complexity of reward systems and the number of targeted performance measures increase, the salience of any one quality metric may decrease along with the attention providers give to improving in that area.

One potential solution is to signal the relative importance of some measures over others by assigning weights to each performance measure used in the final bonus calculation. Program designers can also signal the importance of a particular measure by requiring providers to perform well on it as a qualifier for receiving any incentive payment. The goal of this paper is to analyze how providers respond to P4P programs that vary in incentive design. We do so in the setting of nursing home P4P, where a number of state Medicaid agencies have implemented state-specific programs that vary in incentive structure. Specifically, we examine how the weights put on various performance measures that are tied to P4P bonus incentives, such as clinical outcomes, inspection deficiencies, and staffing levels, affect improvements in those measures. We also analyze how nursing homes respond to the use of simple, dichotomous thresholds for P4P eligibility relative to the use of weights.

While there is significant heterogeneity across state nursing home P4P programs in the weights put on performance measures and the use of weights versus simple qualifiers for eligibility, the effects of these structural choices on provider performance are unknown. In fact, even outside the nursing home setting, surprisingly little is known about how these structural features of P4P matter for quality improvement. Many studies have noted that program design may play an important role in provider response to P4P and have systematically documented the large array of P4P design differences in various health care settings (see for example, Rosenthal et al. 2004; Van Herck et al. 2010; Emmert et al. 2012; Eijkenaar 2013; Eijkenaar et al. 2013), but we are unaware of studies that have empirically examined the impact of these design features directly. Our study attempts to fill this gap. Understanding how incentive structure affects quality improvement has important implications for policy and the future design of provider incentives.

To analyze how the structure of P4P programs impacts nursing home performance, we employ a difference-in-differences strategy, exploiting variation in the timing of P4P

implementation across states as well as variation in the weights states put on clinical outcomes, inspection deficiencies, and staffing ratios in their P4P bonus formula. We use facility-quarter-level data from 2001 to 2009 created from the Minimum Data Set and the Online Survey, Certification, and Reporting data. Our findings suggest P4P design matters considerably. Larger weights put on clinical outcomes sometimes lead to larger improvements in some clinical outcomes, but small (positive) weights often lead to no improvement and even worsening of some clinical outcomes. These results are consistent with standard multitasking theory, which predicts that providers will allocate effort towards those measures that are relatively more highly rewarded. We find a simple qualifier for P4P eligibility based on having few or no severe inspection deficiencies is more effective at decreasing inspection deficiencies than using weights, suggesting simple rules for participation may incent larger improvement. We then examine whether there are heterogeneous responses to P4P structure by nursing home characteristics and find that nursing homes historically associated with better quality—non-profits, non-chains, and facilities with low Medicaid resident populations—experience larger improvements in deficiencies (at any level of severity) in response to the use of deficiencies as a P4P qualifier. On the other hand, we find some evidence of larger improvements in immediate jeopardy deficiencies (i.e. those that are most severe) in response to deficiency qualifiers among for-profits. Our findings highlight the importance of design heterogeneity and that only examining the average effects of multi-faceted P4P programs without considering program structure may mask differential provider responses.

The paper proceeds as follows. Section II provides background on nursing home quality and state Medicaid nursing home P4P programs. Section III discusses a conceptual model that demonstrates how incentive design impacts quality improvements. We describe the data in Section IV and discuss our empirical strategy in Section V. We present the results in Section VI and sensitivity analysis in Section VII. In Section VIII we conclude.

II. BACKGROUND

II.A. Nursing Home Quality

Over 1.5 million people reside in US nursing homes at a cost of over $120 billion per year (Kaiser Family Foundation 2011). 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 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, Fought, and Ray 1994; Castle, Fogel, and Mor 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 2001).

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, Konetzka, and Liang 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 by state (Werner, Konetzka, and Polsky 2013). However, that evaluation treated P4P as a uniform, broad intervention and did not consider the structural differences in programs across states.

II.B. Nursing Home P4P

Between 2002 and 2009, 8 states adopted Medicaid-sponsored P4P programs in nursing homes[2], all of which primarily targeted quality of care for long-stay (or chronic-care) residents. The details of these programs have been previously described (see Werner, Konetzka, and Liang 2010). Briefly, each state uses a payment model based on a point system that is translated into bonus payments. States assign points to a nursing home based on performance on a combination of clinical quality measures, staffing measures, results from state inspections assessing regulatory compliance or deficiencies in compliance, and other metrics.[3]

For each measure included in the payment model, each nursing home is evaluated and earns points based on whether it has achieved a performance target. The number of points assigned to each quality measure varies across states. 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. The maximum add-on (and thus potential size of the financial incentive) varies by state. For example, Colorado's program used a $4 per diem maximum add-on during the study period, which translated to an approximate 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 3 percent maximum add-on (which is equivalent to approximately $3.58 per Medicaid patient day). Oklahoma used a $5.50 per diem add-on (or an approximate 5.7 percent increase in per diem rates).

Because each state Medicaid agency determines the quality measures used in its P4P program as well as the points assigned to those measures, there is variation across states in the weights assigned to various quality metrics. Table 1 shows the states with P4P programs in place as of the end of 2009, implementation dates, and the weights assigned to a subset of quality measures in each state.[4] Table A1 in the Online Appendix provides details about the

[2]We do not consider Vermont's P4P program and exclude nursing homes in the state of Vermont from our sample. Vermont implemented a P4P-like program in 2000, before our sample begins. Further, Vermont's P4P program was fundamentally different from those of other states. Initially, it did not use per diem add-ons, but instead gave flat bonuses to a maximum of 5 nursing homes that met quality targets.

[3]Other quality measures commonly used in P4P programs include overall occupancy, Medicaid occupancy, consumer satisfaction, and culture change. See Werner, Konetzka, and Liang (2010) for details on the full set of performance metrics used in each state's P4P program.

[4]Colorado passed legislation establishing a P4P program in 2008. P4P add-ons were distributed starting in FY 2009, but were based on performance in the prior fiscal year. Given the program details were announced in 2008 and add-ons were based on FY 2008 outcomes, we define the starting date of Colorado's program as July 2008. For add-ons distributed in FY 2010, Colorado made some adjustments to their program. These changes were announced in 2009 and add-ons for FY 2010 were based on FY 2009 outcomes. Thus, we allow these changes to be reflected in the weights starting in July 2009.

construction of the weights. We focus on the weights put on clinical outcomes, inspection deficiencies, and staffing ratios because they are the most common dimensions of quality targeted by P4P and we observe these outcomes in our data. Four of the eight P4P states reward clinical outcomes, with weights ranging from 0.1 to 0.4 (from a total of 1). In terms of specific clinical outcomes, all four states that reward clinical outcomes target physical restraint use and pain, and three of them target pressure sores. Other clinical outcomes used less often include bladder catheterization, falls, and unexplained weight loss. With the exception of Minnesota's program, each individual clinical outcome is equally weighted within the P4P bonus formula. Most states reward staffing ratios, with weights ranging from 0.1 to 0.33. All states except Kansas base their P4P award in some way on inspection deficiencies, with four states assigning points to nursing homes based on the number and severity of their deficiency citations. Rather than assigning points to and putting weights on inspection deficiencies, three states (Colorado, Georgia, and Utah) require that facilities not have any severe deficiency citations in order to participate in the P4P program, with "severe" defined slightly differently across the states.[5] Thus, these states use inspection deficiencies as a qualifier for P4P eligibility. This particular incentive design feature allows us to examine how simple rules for P4P eligibility impact quality improvements.

III. CONCEPTUAL MODEL

To provide some intuition for how variation in P4P program structure may impact nursing home quality improvements, we present a simple multitasking model. Our model follows directly from that of Mullen, Frank, and Rosenthal (2010) which presents an application of the multitasking model of Holmstrom and Milgrom (1991) to study P4P in the California physician medical group setting. Like Mullen, Frank, and Rosenthal (2010), in order to make clear the key ideas, we abstract from quantity of care provided (in our case, the number of nursing home residents) and focus on the nursing home's choice of quality.

The nursing home chooses a quality level which is unobservable to the state Medicaid agency. Quality is multidimensional and is represented by the vector $q = (q_1, q_2, \ldots, q_J)$. $B(q)$ denotes the expected (and possibly unobserved) benefit that accrues to the state Medicaid agency from the nursing home's quality level choice. The nursing home incurs a cost, $C(q)$, which depends on its quality level. C is increasing in q, convex, and captures costs broadly (for example, wages paid to staff and nurses and infrastructure investments).

The nursing home's unobserved quality level generates a vector of observable indicators. Denote the vector of observable indicators as $y = (y_1, y_2, \ldots, y_K)$. Examples of these indicators include the proportion of residents who are physically restrained (or experience pain, have pressure sores, etc.), the facility's number of inspection deficiencies, and staffing levels. The observable indicators are noisy functions of q. That is, they are a function of q but do not perfectly reveal q, and they are given by:

[5]For example, in Utah, a nursing home that receives a violation at the "immediate jeopardy" level (i.e. a deficiency of scope and severity level J, K, or L) is ineligible for the bonus. In Colorado, no facility with "substandard quality of care" deficiencies (i.e. deficiencies of scope and severity level F, H, I, J, K, or L) on a regular annual, complaint, or any other Colorado Department of Public Health and Environment survey is considered for the bonus.

$$y=\mu(q)+\varepsilon$$

where $\mu:\Re_+^J \rightarrow \Re^K$ and is assumed to be concave. ε is a mean zero vector, and $\varepsilon_k \sim F_k$ for $k=1,2,\ldots,K$ where F_k is the cumulative density function of ε_k. To keep the model tractable, we assume $E(\varepsilon_k \varepsilon_k' | q) = 0$ for all k and k'. In this setting, μ can be thought of as the nursing home's production technology that converts effort (or unobserved quality) into observable signals of quality.

We denote as $R(y)$ the Medicaid reimbursement to the nursing home. For facilities in states without a P4P program in place, the reimbursement does not depend on the observable indicators of quality and can be represented by a flat rate, $R(y) = \beta$.[6] In this case, the nursing home chooses a quality level q that minimizes its costs, yielding the following first order condition:

$$\frac{\delta C}{\delta q_j}=0, \quad j=1,2,\ldots,J$$

Now we consider the optimization problem faced by nursing homes in states with P4P programs in place. To keep the model tractable while still illustrating how nursing homes may respond to different program structures, we first assume the P4P add-on rule is simple. The nursing home receives an add-on (or bonus), r_k, to its normal per diem rate when an observable indicator of quality y_k exceeds some threshold denoted by T_k. Different weights put on different observable indicators of quality can be represented by variation in the add-ons across the observable indicators. The nursing home maximizes expected profits which are given by:[7]

$$\begin{aligned} E[R(y)] - C(q) &= E[\beta + \textstyle\sum_{k=1}^{K} r_k I(y_k \geq T_k)] - C(q) \\ &= \beta + \sum_{k=1}^{K} r_k Pr(y_k \geq T_k) - C(q) \\ &= \beta + \sum_{k=1}^{K} r_k [F_k(\mu_k(q) - T_k)] - C(q) \end{aligned}$$

The first order condition yields:

$$\frac{\delta C}{\delta q_j}=\sum_{k=1}^{K} r_k \frac{\delta \mu_k}{\delta q_j}[f_k(\mu_k(q)-T_k)], \quad j=1,2,\ldots,J$$

[6]Since the late 1990s, most states reimburse nursing homes using a prospective or flat-rate payment system. The rates are set before the rate year and do not factor in costs incurred by the nursing homes during the rate year (Grabowski et al. 2004; Miller et al. 2009). Thus, we find it reasonable to assume a flat, pre-determined reimbursement (in the absence of P4P add-ons).

[7]We focus on the for-profit nursing home's maximization problem. However, the model could accommodate non-profit facilities by assuming they maximize $[(y)] + \alpha B(q) - C(q)$ where $0 < \alpha < 1$.

The nursing home chooses q such that the marginal cost of improving quality dimension j equals the expected marginal revenue from improving q_j for all $j=1,2,\ldots,J$.

The above first order condition makes clear there are several aspects to the marginal benefit of improving quality dimension j. First is the expected marginal increase in the observed indicator(s) of quality that results from the improvement in q_j. Second is the add-on(s) for performing above the threshold for the relevant observed quality indicator(s). Third is the probability of exceeding the eligibility threshold for the relevant observed quality indicator(s). Thus, when there are changes to the relative returns of various quality dimensions (that could occur due to a change in r_k, for example), the nursing home has an incentive to reallocate its resources across different dimensions of quality. Depending on the nursing home's production technology, it may be optimal for the facility to allocate resources toward rewarded indicators of quality at the expense of measures that are less rewarded. Even if a particular observable indicator of quality is rewarded by a P4P program, the nursing home may not significantly improve that measure (or could even let this measure deteriorate) if other indicators are rewarded more heavily or are less costly to improve. Thus, if small weights are put on certain observable indicators, it is possible that those indicators might deteriorate or see no improvement while indicators that are more heavily rewarded experience improvements, as it is the relative reward that matters.

It is important to note, however, that measures that are relatively less rewarded may improve if they share commonalities in production with measures that are more highly rewarded. In the model presented above, commonalities in production exist for two observable outcomes y_k and $y_{k'}$ if they both positively depend on some unobservable quality dimension q_j. For example, in some P4P programs, a large weight might be placed on staffing levels. If facilities respond by increasing their efforts to recruit and retain staff and if staffing levels impact the frequency of adverse clinical outcomes (for example, falls or pressure sores), then these clinical outcomes may improve even if they are relatively less rewarded. Little empirical evidence is available to illuminate particular commonalities in production of nursing home quality, but any such commonalities would reduce the net multitasking effect in our empirical results.[8] In addition, it is worth noting that there is little empirical evidence regarding the production function governing nursing home quality more generally. Thus, specific information available to guide nursing home managers in making investments in quality improvements is limited, and in reality, they likely make such investments facing some uncertainty about the production technology of care.

Finally, we consider the case where there is a simple rule for P4P eligibility based on a nursing home's performance on one observed indicator. We assume a nursing home is eligible to receive the bonus described above if some observable indicator, y_1, exceeds some threshold T_1. In this case, the nursing home reimbursement is given by:

$$R(y) = \beta + I(y_1 \geq T_1) \sum_{k=2}^{K} r_k I(y_k \geq T_k)$$

[8]For example, Mor et al. (2003) find relatively low levels of correlation among various nursing home performance measures.

and the nursing home's expected profits are:

$$E\left[R\left(y\right)\right]-C\left(q\right)=$$
$$\beta+F_1(\mu_1\left(q\right)-T_1)\sum_{k=2}^{K}r_k[F_k(\mu_k\left(q\right)-T_k)]-C\left(q\right)$$

recalling that $K(\varepsilon_k\varepsilon_k{}'|q) = 0$. The first order condition yields:

$$\frac{\delta C}{\delta q_j}=f_1(\mu_1\left(q\right)-T_1)\frac{\delta\mu_1}{\delta q_j}\sum_{k=2}^{K}r_k[F_k(\mu_k\left(q\right)-T_k)]$$
$$+F_1(\mu_1\left(q\right)-T_1)\sum_{k=2}^{K}r_k\frac{\delta\mu_k}{\delta q_j}[f_k(\mu_k\left(q\right)-T_k)],\quad j=1,2,\ldots,J$$

Relative to the case without a P4P qualifier, there is an additional marginal benefit from improving quality dimension q_J—a potential increase in the probability that observed quality indicator y_1 exceeds the threshold needed for eligibility. Thus, in the presence of such simple qualifiers, nursing homes may prioritize their efforts and allocate them first towards quality areas that increase their probability of P4P eligibility.

IV. DATA

We construct a facility-quarter level dataset from 2001 to 2009, including all Medicare and/or Medicaid-certified nursing homes in all states in the US (except Vermont). Our data come from two sources—the Minimum Data Set 2.0 (MDS) and the Online Survey, Certification, and Reporting (OSCAR) data.

The MDS contains detailed resident-level data obtained from regular assessments of residents in Medicare and/or Medicaid-certified nursing homes,[9] and it is also the data used by state Medicaid agencies to measure clinical quality and determine P4P bonuses. The MDS contains information on residents' health, activity of daily living (ADL) impairments, cognitive status, and behavioral problems. We use the MDS to construct quarterly facility-level measures of 6 clinical quality metrics, focusing on those most commonly used in P4P programs—the percentage of long-stay residents who were physically restrained, who had moderate to severe pain, who developed pressure sores, who had a bladder catheter inserted, who had unexplained weight loss, and who had falls. These outcomes are also commonly used in the literature as measures of nursing home quality.

We focus only on long-stay residents because P4P programs typically targeted these residents. Long-stay residents are usually chronically ill, require non-skilled care such as assistance with ADLs, and typically spend the remainder of their lives in a nursing home. We do not consider short-stay residents, individuals usually requiring rehabilitative or restorative care after a hospitalization and residing in the nursing home for less than 100

[9]Assessments occur upon admission to the nursing home, quarterly, annually, and when there is a significant change in the resident's status. We limit each resident to only one assessment per quarter to avoid putting excessive weight on residents who are in poorer health and have frequent assessments. In the event a resident has multiple assessments in a quarter, we include the most recent one. We also exclude admission assessments.

days. In addition, short-stay patients are typically covered by Medicare, while long-stay patients are typically covered by Medicaid. As in prior work (Werner, Konetzka, and Polsky 2013), we classify residents as long-stay in the MDS if we observe at least one quarterly or annual assessment in addition to an admission assessment or a prior quarterly or annual assessment.

In constructing the clinical measures, we follow the conventions set by the Centers for Medicare and Medicaid Services (Morris et al. 2003). We determine which resident assessments are eligible (or at risk) for the clinical outcome of interest to calculate the denominator, and then calculate the number of residents who had or experienced the outcome of interest among those who were eligible (or at risk) to create the numerator. We risk adjust these facility-level clinical measures following CMS conventions.

OSCAR contains facility data collected during state inspections of nursing homes. These inspections occur at least once every 15 months. We use OSCAR to construct measures of facility inspection deficiencies as well as staffing ratios. In terms of deficiencies, nursing homes are cited if state surveyors find they are in non-compliance with requirements or standards related to care practices and management. The severity of each citation is also recorded. We create an indicator for facilities that had deficiencies of any severity at their most recent past inspection as well as an indicator for facilities that had deficiencies at the immediate jeopardy level, the most serious deficiencies. We construct two staffing ratio measures—total staffing hours per resident day and skilled staffing hours per resident day. Total staff includes registered nurses (RNs), licensed practical nurses (LPNs), and nurse aides, while skilled staff includes RNs and LPNs.

We also use the MDS and OSCAR data to construct time-varying facility control variables. We use the resident-level MDS data and aggregate it to the facility level to construct the facility's average resident age, the percent of residents who are female, the percent of residents in particular racial and ethnic groups, and, the facility's average Cognitive Performance Scale (Morris et al. 1994), ADL scale (Morris, Fries, and Morris 1999), and Clinically Complex Scale (Kidder et al. 2002). We use the OSCAR data to construct the following variables: the facility's percent of residents covered by Medicare and percent of residents covered by Medicaid; ownership (for-profit, non-profit, or government-owned); whether the facility is hospital-based; whether the facility is part of a chain; and, the facility's total number of beds.

We follow criteria used by CMS and used in prior studies (Abt Associates Inc. 2001; Konetzka et al. 2004) to determine and exclude erroneous observations. We exclude facility-quarter observations where facilities reported more residents than total number of beds. We also exclude observations where facilities reported no RN hours but had 60 or more beds[10] as well as facilities that reported more than 12 total staff hours per resident day or less than 0.5 total staff hours per resident day.[11]

[10]Federal regulations require that facilities with 60 or more beds have an RN on duty 8 hours per day, 7 days per week.
[11]This restriction is made to avoid unreasonably high or low staffing hours.

Table 2 shows the mean values of the outcome variables we consider for nursing homes in P4P and non-P4P states in 2001, before any P4P programs were implemented. We find no evidence that nursing homes in states where P4P was eventually implemented had systematically worse or better quality than facilities in states that never implemented P4P during our sample period. For example, nursing homes in states that implemented P4P had a slightly lower probability of having inspection deficiencies, but also had lower staffing levels on average relative to facilities in states that did not implement P4P. In terms of clinical measures, we find no convincing pattern that would imply facilities in P4P states had better or worse outcomes compared to non-P4P states. These statistics offer suggestive evidence that P4P was not implemented in states with historically low (or high) quality in nursing homes, which is important for our identification strategy.[12]

Table 3 shows average facility characteristics over the full sample period, separately by states that implement P4P and states that do not implement P4P during our study period. There are 3,472 unique nursing homes that are in states that implement P4P, representing approximately 20 percent of the facilities in the data. Facilities in P4P states are more likely to be non-profits, smaller (in terms of total beds), and chain-affiliated, and they tend to have a less diverse patient-mix (in terms of race and ethnicity) compared to nursing homes in non-P4P states. We control for these observable characteristics in our empirical specifications, and we also allow for permanent unobserved differences in facility characteristics via the inclusion of nursing-home-specific fixed effects.

V. EMPIRICAL STRATEGY

We employ a difference-in-differences strategy to examine how the structure of P4P programs affects facility-level clinical quality measures, inspection deficiencies, and staffing ratios. Our identifying variation arises from several sources: whether or not a P4P state rewards a particular measure; the amount of weight given to that measure among the states that reward it; and variation in the timing of P4P implementation across states.

To analyze the impact of P4P program structure on clinical quality measures we estimate the following difference-in-differences regression:

$$QM_{j,s,t}=\alpha P4P_{s,t}+\beta P4PClinical_{s,t}+\delta P4PClinicalWeight_{s,t}+\varphi X_{j,s,t}+\tau_t+\gamma_j+\varepsilon_{j,s,t}$$

where $QM_{j,s,t}$ is the fraction of residents at nursing home j in state s at time t that experience a particular clinical outcome. $P4P_{s,t}$ is an indicator for whether the nursing home is in a state that has a P4P program in place at time t; $P4PClinical_{s,t}$ is an indicator for whether the nursing home is in a state that rewards clinical outcomes in its P4P program; and, $P4PClinicalWeight_{s,t}$ is the weight a state puts on clinical outcomes in its P4P performance score, which in theory can take on values from 0 to 1 inclusive.[13] $X_{j,s,t}$ is a vector of facility characteristics, τ_t are time fixed effects (where time is measured in quarters), γ_j are facility

[12]The assumptions of the difference-in-differences model are weaker—common trends between the treatment and control groups in the absence of P4P—but, having similar baseline outcomes often improves the face validity of the comparability between the groups.
[13]In practice, the weights on the performance measures we consider are usually less than 0.40.

fixed effects, and $\varepsilon_{j,s,t}$ is a mean zero error term. We consider the 6 clinical measures mentioned above, and we estimate the above equation separately for each clinical measure.[14]

As mentioned previously, some states incorporate inspection deficiencies into their P4P programs by awarding points to nursing homes with zero or few deficiencies, while other P4P states use inspection deficiencies to determine if a nursing home is eligible for a P4P add-on. We analyze the impact of the weight put on inspection deficiencies as well as the impact of being in a state where deficiencies are used as a P4P qualifier on the probability that a facility has deficiencies at any level of severity as well as the probability a facility has deficiencies at the immediate jeopardy level. We estimate the following difference-in-differences linear probability model:

$$\Pr(Defic_{j,s,t}=1)=\alpha P4P_{s,t}+\delta P4PDeficWeight_{s,t}+\omega P4PDeficQualifer_{s,t}+\varphi X_{j,s,t}+\tau_t+\gamma_j+\varepsilon_{j,s,t}$$

where $Defic_{j,s,t}$ is an indicator for whether the facility has deficiencies at any level of severity (and in separate specifications it is an indicator for whether the facility has deficiencies at the immediate jeopardy level). We do not include an indicator for whether the nursing home is in a state that rewards lack of deficiencies (regardless of how) in its P4P program, $P4PDefic_{s,t}$, because all P4P states except Kansas incorporated deficiencies in some form in their program during our sample period. Thus, we cannot meaningfully separately identify the impact of $P4P_{s,t}$ from $P4PDefic_{s,t}$. $P4PDeficWeight_{s,t}$ is the weight a state puts on deficiencies in its P4P performance score, which can take on values from 0 to 1 inclusive, and $P4PDeficQualifier_{s,t}$ is an indicator for whether the state uses deficiencies to determine whether a nursing home is eligible for a P4P bonus. $X_{j,s,t}$, τ_t, γ_j, and $\varepsilon_{j,s,t}$ are as defined above.

We estimate the impact of P4P design on total staffing hours per resident day as well as total RN and LPN hours per resident day. We do so by estimating the following equation:

$$Staffing_{j,s,t}=\alpha P4P_{s,t}+\beta P4PStaffing_{s,t}+\delta P4PStaffingWeight_{s,t}+\varphi X_{j,s,t}+\tau_t+\gamma_j+\varepsilon_{j,s,t}$$

where $Staffing_{j,s,t}$ is either total staffing hours per resident day or the sum of RN and LPN hours per resident day at facility j at time t. $P4PStaffing_{s,t}$ is an indicator for whether the nursing home is in a state that rewards staffing levels in its P4P program, and $P4PStaffingWeight_{s,t}$ is the weight put on staffing ratios in the state's P4P performance score, which can take on values from 0 to 1 inclusive.[15] $X_{j,s,t}$, τ_t, γ_j, and $\varepsilon_{j,s,t}$ are as defined above.

[14]Alternatively, we could create a composite measure of nursing home quality and analyze how the weights put on certain metrics impact that composite measure. However, there is no empirical basis for the use of a composite measure. We could create such a measure using each state's P4P bonus formula, but unfortunately, we do not observe all the metrics that factor into each state's total performance score (such as culture change and consumer satisfaction). Furthermore, as mentioned earlier, the literature has found little correlation between the quality measures we consider, and we also find this is the case in our estimation sample. Thus, we find it more informative to analyze how P4P weights affect each outcome separately.

[15]We only consider the weight put on staffing ratios and do not include the weight put on staffing turnover, retention, or satisfaction. The OSCAR data does not contain information on those staffing measures.

The inclusion of time fixed effects in the above specifications controls for any systematic trends in clinical quality, inspection deficiencies, or staffing that affect all nursing homes. The facility fixed effects account for any facility time-invariant unobservables (and observables) that affect our outcomes of interest, and they also subsume state fixed effects. Thus, identification of the parameters of interest (i.e. those related to P4P) relies on within-facility variation in our outcomes of interest for nursing homes in P4P states before and after P4P was implemented compared to within-facility variation in those outcomes for facilities in states without P4P. In all specifications, standard errors are clustered at the facility level.

We then explore whether there are heterogeneous effects of P4P and P4P program design by facility characteristics. In particular, we analyze whether there are differential responses between for-profit facilities and non-profit facilities; chains and non-chains; and, facilities with a high percentage of residents who are covered by Medicaid (greater than 75 percent) and facilities with a low percentage of residents who are covered by Medicaid.[16] We are interested in differential responses by these facility characteristics because the literature has generally found for-profits, chains, and facilities with a relatively large Medicaid population deliver poorer quality of care (Harrington et al. 2001; Hillmer et al. 2005; Comondore et al. 2009). Furthermore, we explore differential responses by the share of residents who are covered by Medicaid because the P4P bonus is applied to the per diem for Medicaid resident days; thus, all else equal, the marginal revenue from quality improvements is predicted to be larger for nursing homes with a larger Medicaid census.[17]

VI. RESULTS

VI.A. Clinical Outcomes

The coefficient estimates on the *P4PClinicalWeight* variable suggest larger weights put on clinical outcomes are associated with larger improvements in the prevalence of pain, weight loss, and falls, *ceteris paribus* (see Panel A of Table 4). However, the full effect of P4P programs that reward clinical quality depends on the sum of the P4P-related coefficients as well as the actual weight put on the clinical measures. Panel B of Table 4 shows the effects and associated standard errors of rewarding clinical outcomes relative to P4P programs that do not reward clinical outcomes for the smallest (0.1) and largest (0.4) weight put on clinical measures observed in our sample period.[18] We show the range of effects because it is not immediately obvious from the coefficient estimates alone and it provides information about the practical magnitude of the effects given the weights observed.

Interestingly, we find significant improvements in physical restraint use for nursing homes in states with both the smallest and largest clinical weight, with the improvement being larger for nursing homes with the smaller weight. However, these effects are not significantly

[16]The 75 percent of residents covered by Medicaid threshold was chosen because it corresponds to the 75th percentile of the distribution.

[17]Ideally, we would additionally analyze differential responses to P4P by the bonus size, which also affects the marginal revenue from quality improvements. Unfortunately, we do not have this information for all P4P states, particularly the ones that specify the bonus as a percentage of the Medicaid per diem reimbursement, as time-varying Medicaid nursing home reimbursement rates are unavailable for most states.

[18]Panel B of Table 4 shows the point estimate and standard error of the linear combination of the coefficient on *P4PClinical* plus the coefficient on *P4PClinicalWeight* times the actual weight.

different from each other at conventional significance levels. This result suggests restraint usage improvements are not particularly sensitive to the weight put on clinical outcomes in the range of weights we observe. We attribute this result in part to the heavy emphasis placed on restraint use by public reporting and CMS, which may have provided nursing homes the impetus to reduce restraint use regardless of incentive size. The results also suggest the observed weights put on clinical outcomes do not generate significant improvements in pressure sore incidence, which is surprising given that all the P4P states that rewarded clinical measures targeted pressure sores. In the case of both restraint use and pressure sores, as performance has improved on average in most nursing homes, a performance ceiling may have been reached, making further improvements difficult regardless of the weight used. When the weight put on clinical outcomes is small, pain prevalence does not significantly change while weight loss and falls increase. All three of those measures significantly decrease when the weight put on clinical outcomes is larger. This result is of policy importance since it suggests low weights put on measures can lead to no improvements and even deteriorations in those measures. A likely explanation is that nursing homes allocate more effort to improving quality dimensions that are more heavily weighted in the P4P bonus formula and away from dimensions less heavily weighted, consistent with the multitasking theory outlined above. We find no evidence of significant improvements in catheter use given the clinical weights considered.

VI.B. Inspection Deficiencies

We find a negative and significant coefficient on the weight put on deficiencies when the probability of having any deficiencies is the outcome of interest, and we find that using deficiencies as a qualifier for P4P bonus receipt is also associated with a significant decrease in the probability of having any deficiencies (see Panel A of Table 5). None of the P4P related coefficients are statistically significant when immediate jeopardy level citations are the outcome of interest.[19] Panel B of Table 5 shows the range of effects and associated standard errors given the smallest (0.1) and largest (0.22) weights put on deficiencies observed in our sample period relative to P4P programs that do not reward lack of deficiencies.[20] We find both the smallest and largest weights generate significant decreases in the probability of having any deficiency, with the effects ranging from a 1.1 to 2.6 percentage point decrease. However, the decrease in the probability of having any deficiencies that results from using deficiencies as a qualifier is even larger in magnitude (a 6.5 percentage point decrease versus a 2.6 percentage point decrease when the largest weight is used). These results suggest using deficiencies to disqualify nursing homes from the P4P bonus rather than putting weights on deficiencies may be a more effective means of generating improvements. While improvements in deficiencies are likely of a different nature and generated from a different production process than improvements in clinical quality or staffing, these results suggest using certain dimensions of quality as disqualifiers from P4P could be important for improving quality.

[19]Given that only about 3 percent of observations in our sample have a deficiency at the immediate jeopardy level and identification comes from within-facility variation in deficiencies, it is not surprising we find less precise estimates when deficiencies at the immediate jeopardy level are the outcome of interest.

[20]Panel B of Table 5 shows the point estimate and standard error of the coefficient on *P4PDeficWeight* times the actual weight.

VI.C. Staffing Ratios

We find the coefficient on the level effect of having a P4P program that rewards staffing levels is negative and significant when considering both total staffing and skilled staffing levels (see Panel A of Table 6). The weight put on staffing levels has no significant effect on total staffing but significantly increases the level of skilled staffing (p=0.052). Panel B of Table 6 shows the effects and associated standard errors of the effects of rewarding staffing levels relative to P4P programs that do not reward staffing levels for the smallest (0.1) and largest (0.33) weight put on staffing levels observed in our sample period.[21] We find the observed weights do not generate significant increases in staffing levels. In fact, when the smaller weight is used, nursing homes experience significant decreases in staffing levels. Similar to our clinical quality results, these results are consistent with standard multitasking theory and highlight that small weights put on quality dimensions can actually lead to deteriorations in these measures, particularly if nursing homes allocate effort away from improving such dimensions and instead focus on areas that are more heavily weighted.

VI.D. Heterogeneous Effects

We then consider heterogeneous responses to P4P and P4P program structure by facility characteristics such as chain affiliation, ownership, and the share of patients covered by Medicaid. The full results from these analyses are available upon request. In what follows we discuss and present the findings where we find systematic heterogeneous effects.

The coefficient estimates in Panel A of Table 7 show evidence of heterogeneous effects by facility characteristics when we consider the probability of having inspection deficiencies at any level of severity. We find the coefficient on using deficiencies as a P4P qualifier is negative and significant for nursing homes of all facility types, but larger in magnitude for facilities historically associated with better quality of care—non-chains, non-profits, and facilities with a smaller share of Medicaid-covered patients. We also find the coefficient on the weight put on deficiencies is negative and significant only for that subset of facilities. If such facilities already deliver relatively high quality of care, they may be able to allocate more effort to eliminating all deficiencies relative to facilities that have more dimensions of quality that need improvement, and hence more areas that require an investment of resources and effort. Related, perhaps these facilities are able to make these deficiency improvements at lower cost relative to chains, for-profits, and facilities with a large share of Medicaid residents. When we consider the range of effects given the smallest and largest weight put on deficiencies in Panel B of Table 7 and whether deficiencies are used as a P4P qualifier, we again find that using deficiencies as a qualifier generates larger improvements than the largest weight put on deficiencies observed in our sample.

In Panel A of Table 8, we show the coefficient estimates when we explore heterogeneous effects on the probability of having deficiencies at the immediate jeopardy level. The coefficient on using deficiencies to disqualify facilities from P4P is negative for all facility types historically associated with poor quality of care—chains, for-profits, and facilities with

[21]Panel B of Table 6 shows the point estimate and standard error of the linear combination of the coefficient on *P4PStaffing* plus the coefficient on *P4PStaffingWeight* times the actual weight.

a large share of residents covered by Medicaid—but, these coefficients are generally not significant at conventional levels (except for for-profit nursing homes). Panel B of Table 8 shows the range of effects given the linear combination of coefficients and the observed weights. Here, we find the observed weights put on deficiencies do not generate significant decreases in the probability of having a deficiency at the immediate jeopardy level relative to states that do not incorporate deficiencies in their programs. We cautiously interpret the result that using deficiencies as a qualifier generates significant improvements among for-profit nursing homes as suggestive evidence that some facilities typically associated with poor quality allocate effort to eliminate their most severe deficiencies, perhaps because such improvements are "low-hanging fruit," while facilities typically associated with better quality allocate effort to eliminate all deficiencies.

VII. SENSITIVITY ANALYSES

Above we consider how rewarding certain quality metrics in P4P programs affects those specific metrics. However, the multitasking model implies that these outcomes may be jointly determined. For example, clinical outcomes may not only be impacted by the weight put on clinical measures, but also by the weights put on staffing ratios and inspection deficiencies. To explore this idea, we reestimate the specifications above maintaining a consistent and comprehensive set of P4P related independent variables. Specifically, we allow each performance metric (clinical outcomes, the probability of having deficiencies, and staffing ratios) to be a function of $P4P_{s,t}$, $P4PClinical_{s,t}$, $P4PClinicalWeight_{s,t}$, $P4PDeficQualifier_{s,t}$, $P4PDeficWeight_{s,t}$, $P4PStaffing_{s,t}$, and $P4PStaffingWeight_{s,t}$. In this way, we allow for the weights put on certain metrics to have an impact on other performance measures.[22] Results from these regressions are presented in Tables A2, A3, and A4 in the Online Appendix.

When we consider clinical outcomes, we find the coefficients related to rewarding clinical outcomes are qualitatively and quantitatively similar to our baseline estimates (see Table A2 in the Online Appendix). The range of effects generated from the observed weights put on clinical outcomes is qualitatively similar to the baseline effects, except we find in some cases, such as pain prevalence and weight loss, the largest weight no longer generates significant improvements. We find using deficiencies as a qualifier leads to significant decreases in physical restraint use, pressure sores, pain, weight loss, and catheterizations. This makes sense as these adverse clinical outcomes can trigger quality of care deficiency citations. Generally, we do not find that weights put on deficiencies or staffing lead to systematic improvements in clinical outcomes.

When we consider deficiencies, we again find using deficiency citations as a P4P qualifier generates a significant decrease in the probability of having any deficiencies relative to programs that do not target deficiencies (see Table A3 in the Online Appendix). We also find that the larger the weight put on deficiencies, the larger the decrease in the probability of having any deficiencies. In this case, we find the largest weight put on deficiencies generates

[22]To the extent weights put on other rewarded measures like consumer satisfaction and culture change impact clinical outcomes, deficiencies, or staffing, those effects would be broadly captured in the coefficient on $P4P_{S,t}$.

improvements similar in magnitude to those generated from using deficiencies as a qualifier. We also find evidence that larger weights put on staffing levels significantly decrease the probability of having a deficiency.

We find the coefficients related to rewarding staffing levels are qualitatively similar to our baseline estimates when staffing levels are the outcome of interest (see Table A4 in the Online Appendix). When we consider ranges of effects, we again find that small weights put on staffing can lead to significant decreases in staffing levels. We find that the use of deficiencies as a qualifier decreases total staffing, potentially a multitasking issue.

Overall, when we allow our outcomes to be a function of a comprehensive set of P4P related variables, we find inconclusive and inconsistent evidence for weights put on certain metrics having spillover effects on other performance measures. Because the inclusion of these potential spillovers increases the complexity of the model while leaving our main conclusions largely unchanged, we rely on the model without spillovers as our preferred specification.

VIII. DISCUSSION AND CONCLUSION

Despite the prominence of P4P programs aimed at improving the quality of health care in the United States, prior evidence on the effectiveness of various ways of structuring P4P programs is sparse. Our analysis of Medicaid P4P programs in nursing homes begins to fill this gap. Using a difference-in-differences design capitalizing on both within-state changes over time in the existence of P4P programs and across-state differences in program structure, we estimate the effects on quality-related outcomes of two key program features, the use of weights and the use of qualifiers for bonus eligibility. Our results have several important implications for policy.

We find that the use of weights on clinical quality outcomes has consequences that were unintended by policymakers. First, stronger weights sometimes lead to more improvement, as expected, but this is not always the case. Second, small (but positive) weights lead, in several cases, to a *decline* in performance on some clinical measures. This is consistent with the theory of multitasking in which the relative importance of a targeted measure matters. Although policymakers might assume that a small weight would still induce positive change, when resources for quality improvement are scarce, this assumption appears to be incorrect. Health care providers may simply focus on the measures that bring the highest relative rewards. It is also possible that observed improvements in quality do not reflect true improvements in quality. Nursing homes might use alternative low-cost methods to improve their performance by simply improving (or changing) the coding in the data used for the P4P performance metrics. In this case, weights might increase the attention given to certain performance metrics without resulting in substantive improvements in quality.

Furthermore, we find that the use of a deficiency threshold as a qualifier for eligibility for any bonuses under the P4P program is more effective than using deficiencies in a weighting scheme. Nursing home providers exhibited significant improvement in deficiency-defined quality when used as a qualifier, but smaller improvement when deficiencies received a

weight in the bonus formula. The key to the effectiveness of using a quality measure as a qualifier may lie in its simplicity. Given scarce resources for quality improvement, simple rules lessen the uncertainty associated with choosing areas for quality improvement and incentivize nursing homes to prioritize their efforts towards improvements that increase the probability of P4P eligibility. The effectiveness of using a qualifier is also consistent with multitasking theory, as meeting the qualifying criterion has the highest return.

Finally, our results have distributional implications. One often-expressed fear of P4P programs is that they will reward health care providers that already provide better quality and have more resources and that low-resource providers will not be able to achieve the bonuses even with effort, such that P4P might increase the gap between high- and low-quality providers (Casalino et al. 2007; Konetzka and Werner 2009; Friedberg et al. 2010). To some extent, our results support this concern. Where we do see significant effects of P4P —for example, in the use of deficiencies as a qualifier—we see larger improvement among nursing homes that are nonprofit, non-chain, and with a lower Medicaid census, all attributes traditionally associated with higher quality and better financial performance on average. However, in our analysis of immediate jeopardy deficiencies, our results are suggestive of the opposite—that for-profit facilities exhibit greater improvement. Although this may be due in part to ceiling effects in that the higher quality nursing homes have few, if any, immediate jeopardy deficiencies to begin with, it still reflects improvement among some low-resource, low-quality nursing homes. Thus, P4P appears to be creating an incentive for improvement among some nursing homes that are in the lower tiers of quality. It is not clear, however, whether improvement on this margin would translate into bonuses if these lower-quality nursing homes are unable to improve across all the highly weighted measures or to achieve the overall deficiency qualifier. To create a sustainable incentive for improvement over time, policymakers might consider rewarding improvement in areas most needed by each nursing home rather than a one-size-fits-all approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grant R01-AG034182 from the National Institute on Aging (NIA). Rachel Werner was supported in part by grant K24-AG047908 from the NIA.

References

Abt Associates Inc. Appropriateness of Minimum Nurse Staffing Ratios in Nursing Homes, Report to Congress: Phase II Final Report. Baltimore, MD: Centers for Medicare and Medicaid Services; 2001.

Casalino, Lawrence P., Elster, Arthur, Eisenberg, Andy, Lewis, Evelyn, Montgomery, John, Ramos, Diana. Will Pay-for-Performance and Quality Reporting Affect Health Care Disparities? Health Affairs. 2007; 26(3):w405–w414. [PubMed: 17426053]

Castle, Nicholas G., Fogel, Barry S., Mor, Vincent. Study Shows Higher Quality of Care in Facilities Administered by ACHCA Members. Journal of Long Term Care Administration. 1996; 24(2):11–16.

Comondore, Vikram R., Devereaux, PJ., Zhou, Qi, Stone, Samuel B., Busse, Jason W., Ravindran, Nikila C., Burns, Karen E., Haines, Ted, Stringer, Bernadette, Cook, Deborah J., Walter, Steven D., Sullivan, Terrence, Berwanger, Otavio, Bhandari, Mohit, Banglawala, Sarfara, Lavis, John N., Petrisor, Brad, Schünemann, Holger, Walsh, Katie, Bhatnagar, Neera, Guyatt, Gordon H. Quality of Care in For-Profit and Not-for-Profit Nursing Homes: Systematic Review and Meta-Analysis. British Medical Journal. 2009; 339(7717):381–384.

Eijkenaar, Frank. Key Issues in the Design of Pay for Performance Programs. The European Journal of Health Economics. 2013; 14(1):117–131. [PubMed: 21882009]

Eijkenaar, Frank, Emmert, Martin, Scheppach, Manfred, Schöffski, Oliver. Effects of Pay for Performance in Health Care: A Systematic Review of Systematic Reviews. Health Policy. 2013; 110(2):115–130. [PubMed: 23380190]

Emmert, Martin, Eijkenaar, Frank, Kemter, Heike, Esslinger, Adelheid S., Schöffski, Oliver. Economic Evaluation of Pay-for-Performance in Health Care: A Systematic Review. The European Journal of Health Economics. 2012; 13(6):755–767. [PubMed: 21660562]

Friedberg, Mark W., Safran, Dana Gelb, Coltin, Kathryn, Dresser, Marguerite, Schneider, Eric C. Paying for Performance in Primary Care: Potential Impact on Practices and Disparities. Health Affairs. 2010; 29(5):926–932. [PubMed: 20439882]

Fries, Brant E., Hawes, Catherine, Morris, John N., Phillips, Charles D., Mor, Vincent, Park, Pil S. Effect of the National Resident Assessment Instrument on Selected Health Conditions and Problems. Journal of the American Geriatrics Society. 1997; 45(8):994–1001. [PubMed: 9256854]

Grabowski, David C., Feng, Zhanlian, Intrator, Orna, Mor, Vincent. Recent Trends in State Nursing Home Payment Policies. Health Affairs (Supplemental Web Exclusives). 2004 W4-363-373.

Grabowski, David C., Feng, Zhanlian, Intrator, Orna, Mor, Vincent. Medicaid Nursing Home Payment and the Role of Provider Taxes. Medical Care Research and Review. 2008; 65(4):514–527. [PubMed: 18369236]

Harrington, Charlene, Woolhandler, Steffie, Mullan, Joseph, Carrillo, Helen, Himmelstein, David U. Does Investor Ownership of Nursing Homes Compromise the Quality of Care? American Journal of Public Health. 2001; 91(9):1452–1455. [PubMed: 11527781]

Hillmer, Michael P., Wodchis, Walter P., Gill, Sudeep S., Anderson, Geoffrey M., Rochon, Paula A. Nursing Home Profit Status and Quality of Care: Is There Any Evidence of an Association? Medical Care Research and Review. 2005; 62(2):139–166. [PubMed: 15750174]

Holmstrom, Bengt, Milgrom, Paul. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, & Organization. 1991; 7:24–52.

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

Kaiser Family Foundation. Medicaid and Long-Term Care Services and Supports. 2011. Report No 2186-08https://kaiserfamilyfoundation.files.wordpress.com/2013/01/2186-08.pdf

Kane, Robert L., Arling, Greg, Mueller, Christine, Held, Robert, Cooke, Valerie. A Quality-Based Payment Strategy for Nursing Home Care in Minnesota. The Gerontologist. 2007; 47(1):108–115. [PubMed: 17327546]

Kane, Robert L., Williams, Carter C., Williams, T Franklin, Kane, Rosalie A. Restraining Restraints: Changes in a Standard of Care. Annual Review of Public Health. 1993; 14:545–584.

Kidder, David, Rennison, Melissa, Goldberg, Henry, Warner, David, Bell, Barbara, Hadden, Louise, Morris, John, Jones, Richard, Mor, Vincent. MegaQI Covariate Analysis and Recommendations: Identification and Evaluation of Existing Quality Indicators that are Appropriate for Use in Long-Term Care Settings. Cambridge, MA: Abt Associates Inc; 2002.

Konetzka, R Tamara, Werner, Rachel M. Disparities in Long-Term Care: Building Equity into Market-Based Reforms. Medical Care Research and Review. 2009; 66(5):491–521. [PubMed: 19228634]

Konetzka, R Tamara, Yi, Deokhee, Norton, Edward C., Kilpatrick, Kerry E. Effects of Medicare Payment Changes on Nursing Home Staffing and Deficiencies. Health Services Research. 2004; 39(3):463–488. [PubMed: 15149474]

Mehrotra, Ateev, Damberg, Cheryl L., Sorbero, Melony E S., Teleki, Stephanie S. Pay for Performance in the Hospital Setting: What is the State of the Evidence? American Journal of Medical Quality. 2009; 24(1):19–28. [PubMed: 19073941]

Miller, Edward Alan, Mor, Vincent, Grabowski, David C., Gozalo, Pedro L. The Devil's in the Details: Trading Policy Goals for Complexity in Medicaid Nursing Home Reimbursement. Journal of Health Politics, Policy and Law. 2009; 34(1):93–135.

Mor, Vincent, Berg, Katherine, Angelelli, Joseph, Gifford, David, Morris, John, Moore, Terry. The Quality of Quality Measurement in U.S. Nursing Homes. The Gerontologist. 2003; 43(2):37–46.

Mor, Vincent, Intrator, Orna, Fries, Brant E., Phillips, Charles, Teno, Joan, Hiris, Jeffrey, Hawes, Catherine, Morris, John. Changes in Hospitalization Associated with Introducing the Resident Assessment Instrument. Journal of the American Geriatrics Society. 1997; 45(8):1002–1010. [PubMed: 9256855]

Morris, John N., Fries, Brant E., Mehr, David R., Hawes, Catherine, Phillips, Charles, Mor, Vincent, Lipsitz, Lewis A. MDS Cognitive Performance Scale. Journal of Gerontology. 1994; 49(4):M174–M182. [PubMed: 8014392]

Morris, John N., Fries, Brant E., Morris, Shirley A. Scaling ADLs within the MDS. Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 1999; 54(11):M546–M553.

Morris, John N., Moore, Terry, Jones, Rich, Mor, Vincent, Angelelli, Joseph, Berg, Katherine, Hale, Christine, Morris, Shirley, Murphy, Katharine M., Rennison, Melissa. Validation of Long-Term and Post-Acute Care Quality Indicators. Baltimore, MD: Centers for Medicare and Medicaid Services; 2003.

Mullen, Kathleen J., Frank, Richard G., Rosenthal, Meredith B. Can You Get What You Pay for? Pay-for-Performance and the Quality of Healthcare Providers. The RAND Journal of Economics. 2010; 41(1):64–91. [PubMed: 21667575]

Petersen, Laura A., Woodard, LeChauncy D., Urech, Tracy, Daw, Christina, Sookanan, Supicha. Does Pay-for-Performance Improve the Quality of Health Care? Annals of Internal Medicine. 2006; 145(4):265–272. [PubMed: 16908917]

Rosenthal, Meredith B., Fernandopulle, Rushika, Song, HyunSook Ryu, Landon, Bruce. Paying for Quality: Providers' Incentives for Quality Improvement. Health Affairs. 2004; 23(2):127–141. [PubMed: 15046137]

Rosenthal, Meredith B., Frank, Richard G. What is the Empirical Basis for Paying for Quality in Health Care? Medical Care Research and Review. 2006; 63(2):135–157. [PubMed: 16595409]

Shorr, Ronald I., Fought, Randy L., Ray, Wayne A. Changes in Antipsychotic Drug Use in Nursing Homes During Implementation of the OBRA-87 Regulations. Journal of the American Medical Association. 1994; 271(5):358–362. [PubMed: 8283585]

Snowden, Mark, Roy-Byrne, Peter. Mental Illness and Nursing Home Reform: OBRA-87 Ten Years Later. Psychiatric Services. 1998; 49(2):229–233. [PubMed: 9575011]

Van Herck, Pieter, De Smedt, Delphine, Annemans, Lieven, Remmen, Roy, Rosenthal, Meredith B., Sermeus, Walter. Systematic Review: Effects, Design Choices, and Context of Pay-for-Performance in Health Care. BMC Health Services Research. 2010; 10(1):247–259. [PubMed: 20731816]

Werner, Rachel M., Konetzka, R Tamara, Liang, Kevin. State Adoption of Nursing Home Pay-for-Performance. Medical Care Research and Review. 2010; 67:364–377. [PubMed: 19923629]

Werner, Rachel M., Konetzka, R Tamara, Polsky, Daniel. The Effect of Pay-for-Performance in Nursing Homes: Evidence from State Medicaid Programs. Health Services Research. 2013; 48(4): 1393–1414. [PubMed: 23398330]

Wunderlich, Gooloo S., Kohler, Peter O. Improving the Quality of Long-Term Care. Committee on Improving Quality in Long-Term Care, Division of Health Care Services, Institute of Medicine; Washington, DC: National Academy Press; 2001.

Table 1

Summary of States Implementing P4P between 2001 and 2009; Implementation Dates; and, Weights Assigned to Clinical Outcomes, Staffing Ratios, and Inspection Deficiencies by Each Program

State	Dates of Program	Weights Put On:		
		Clinical Outcomes	Staffing Ratios	Inspection Deficiencies
Colorado	7/2008 to Present	0.27 (FY 2008) 0.25 (FY 2009)	0	Qualifier
Georgia	4/2007 to Present	0.40	0.33	Qualifier
Iowa	7/2002 to Present	0	0.182	0.182
Kansas	7/2005 to Present	0	0.222	0
Minnesota	10/2006 to 9/2008	0.40 (FY 2006) 0.35 (FY 2007)	0 (FY 2006) 0.10 (FY 2007)	0.10
Ohio	7/2006 to Present	0	0.111	0.222
Oklahoma	7/2007 to Present	0.10	0.10	0.10
Utah	7/2003 to Present	0	0	Qualifier

Notes: Weights can take on values from 0 to 1 inclusive. Other quality measures used in P4P programs include overall occupancy, Medicaid occupancy, consumer satisfaction, and culture change, among others.

Table 2

Facility-Level Clinical Outcome, Inspection Deficiency, and Staffing Averages in Non-P4P and P4P States in 2001

	Non-P4P States	P4P States
% of residents who:		
were physically restrained (SD)	10.40 (10.94)	8.47 (8.88)
developed pressure sores (SD)	14.56 (11.36)	12.62 (10.28)
had moderate to severe pain(SD)	11.72 (11.35)	14.28 (11.63)
had unexplained weight loss (SD)	10.27 (8.79)	9.27 (7.40)
had a bladder catheter inserted (SD)	6.76 (7.89)	6.45 (7.19)
had falls (SD)	8.84 (6.07)	10.17 (6.07)
% of facilities that had:		
any deficiencies	90.32	88.88
deficiencies at the immediate jeopardy level	2.47	2.42
Staffing ratios:		
total staff hours per resident day (SD)	3.19 (1.18)	2.96 (1.02)
RN + LPN hours per resident day (SD)	1.10 (0.71)	1.03 (0.58)

Notes: All measures are at the facility-level and summarized for all quarters in 2001, prior to P4P implementation in the states and time period we consider.

Table 3

Descriptive Statistics for Facilities in Non-P4P and P4P States (2001–2009)

	Non-P4P States	P4P States
# of unique facilities	13,816	3,472
% of patients covered by Medicaid (SD)	63.73 (21.99)	61.14 (19.95)
% of patients covered by Medicare (SD)	13.39 (13.31)	10.21 (11.51)
Ownership		
Government, %	5.67	6.12
Non-profit, %	25.24	29.31
For-profit, %	69.09	64.57
Hospital-based, %	5.86	6.98
Chain, %	53.61	56.92
Average # of total beds (SD)	116.07 (71.05)	96.56 (56.14)
% of female patients (SD)	71.38 (13.30)	71.42 (12.55)
Race of patients		
White, % (SD)	82.35 (23.33)	89.19 (17.85)
Black, % (SD)	11.51 (18.73)	8.55 (17.18)
Hispanic, % (SD)	4.07 (10.87)	1.16 (3.82)
Other, % (SD)	2.06 (8.27)	1.11 (4.13)
Average age of patients (SD)	80.68 (7.41)	80.92 (6.90)
Average patient Cognitive Performance Scale (SD)	2.86 (0.63)	2.86 (0.60)
Average patient Activities of Daily Living Scale (SD)	11.34 (1.80)	11.21 (1.78)
Average patient Clinically Complex Scale (SD)	0.56 (0.33)	0.64 (0.35)

Notes: All measures are at the facility-quarter level for the time period 2001 to 2009.

Table 4

Results from Regressions Estimating the Effect of P4P on Clinical Quality Outcomes

	Physically Restrained	Pressure Sores	Pain	Weight Loss	Catheter Inserted	Falls
Panel A: Coefficient Estimates						
P4P	0.00991*** (0.00124)	−0.00218* (0.00130)	−0.00510*** (0.00192)	0.00150* (0.000906)	−0.00162* (0.000934)	0.00310*** (0.000759)
P4PClinical	−0.0207*** (0.00486)	0.00563 (0.00479)	0.0144* (0.00751)	0.0175*** (0.00324)	0.00232 (0.00295)	0.00866*** (0.00253)
P4PClinicalWeight	0.00523 (0.0139)	−0.0186 (0.0135)	−0.0646*** (0.0207)	−0.0601*** (0.00910)	0.000724 (0.00814)	−0.0278*** (0.00701)
Constant	0.0215** (0.0102)	0.220*** (0.0176)	0.383*** (0.0184)	0.0165 (0.0133)	0.109*** (0.0160)	0.0487*** (0.00757)
Time FEs	Yes	Yes	Yes	Yes	Yes	Yes
Facility Covariates	Yes	Yes	Yes	Yes	Yes	Yes
N	518243	514433	518222	517984	515469	514795
R2	0.146	0.0364	0.0592	0.0274	0.0514	0.00421
Panel B: Range of Effects						
Effect given smallest clinical weight (0.10)	−0.02021*** (0.00362)	0.00377 (0.00359)	0.00793 (0.00563)	0.01150*** (0.00244)	0.00239 (0.00224)	0.00588*** (0.00192)
Effect given largest clinical weight (0.40)	−0.01865*** (0.00223)	−0.00182 (0.00215)	−0.01145*** (0.00301)	−0.00653*** (0.00149)	0.00261* (0.00137)	−0.00245** (0.00118)

Notes: Standard errors are clustered at the facility level and shown in parentheses. In Panel B, the effects are relative to having a P4P program that does not reward clinical outcomes.

*p<.1,

**p<.05,

***p<.01

Table 5

Results from Regressions Estimating the Effect of P4P on the Probability of Having Inspection Deficiencies

Panel A: Coefficient Estimates		
	Any Deficiencies	**Any Immediate Jeopardy Deficiencies**
P4P	0.0280*** (0.00891)	0.00457 (0.00653)
P4PDeficWeight	−0.116** (0.0578)	−0.0422 (0.0327)
P4PDeficQualifier	−0.0646*** (0.0123)	−0.0115 (0.00797)
Constant	0.763*** (0.0270)	0.0672*** (0.0161)
Time FEs	Yes	Yes
Facility Covariates	Yes	Yes
N	518249	518237
R2	0.00541	0.00105
Panel B: Range of Effects		
Effect given smallest deficiency weight (0.10)	−0.01159** (0.00578)	−0.00422 (0.00327)
Effect given largest deficiency weight (0.222)	−0.02573**** (0.01282)	−0.00936 (0.00725)

Notes: Standard errors are clustered at the facility level and shown in parentheses.

In Panel B, the effects are relative to having a P4P program that does not reward deficiencies.

* p<.1,

** p<.05,

*** p<.01

Table 6

Results from Regressions Estimating the Effect of P4P on Staffing Ratios

Panel A: Coefficient Estimates		
	Total Staffing HPRD	**RN+LPN HPRD**
P4P	0.0103 (0.0256)	−0.000998 (0.0125)
P4PStaffing	−0.0815** (0.0353)	−0.0345* (0.0182)
P4PStaffingWeight	0.199 (0.145)	0.153* (0.0790)
Constant	3.089*** (0.144)	1.326*** (0.0922)
Time FEs	Yes	Yes
Facility Covariates	Yes	Yes
N	518249	518249
R2	0.0248	0.0226
Panel B: Range of Effects		
Effect given smallest staffing weight (0.10)	−0.06157** (0.02835)	−0.01914 (0.01410)
Effect given largest staffing weight (0.333)	−0.01522 (0.03740)	0.01662 (0.01937)

Notes: Standard errors are clustered at the facility level and shown in parentheses.

In Panel B, the effects are relative to having a P4P program that does not reward staffing ratios.

* p<.1,

** p<.05,

*** p<.01

Table 7

Heterogeneous Effects on the Probability of Having Inspection Deficiencies at Any Level of Severity by Facility Characteristics

Panel A: Coefficient Estimates						
	Non-Chain	**Chain**	**For-Profit**	**Non-Profit**	**Low Medicaid**	**High Medicaid**
P4P	0.0415*** (0.0147)	0.0210* (0.0118)	−0.000768 (0.00974)	0.0595*** (0.0181)	0.0260*** (0.00971)	0.0296 (0.0270)
P4PDeficWeight	−0.244** (0.0992)	−0.0678 (0.0730)	0.0468 (0.0633)	−0.339*** (0.130)	−0.114* (0.0648)	−0.108 (0.156)
P4PDeficQualifier	−0.0881*** (0.0219)	−0.0492*** (0.0158)	−0.0268* (0.0138)	−0.126*** (0.0275)	−0.0697*** (0.0153)	−0.0517* (0.0296)
Constant	0.773*** (0.0391)	0.778*** (0.0387)	0.818*** (0.0351)	0.721*** (0.0490)	0.770*** (0.0340)	0.870*** (0.0571)
Time FEs	Yes	Yes	Yes	Yes	Yes	Yes
Facility Covariates	Yes	Yes	Yes	Yes	Yes	Yes
N	237183	281066	353467	134930	354441	163808
R2	0.00621	0.00503	0.00410	0.00814	0.00701	0.00281
Panel B: Range of Effects						
Effect given smallest deficiency weight (0.10)	−0.02444***** (0.00992)	−0.00678*** (0.00730)	0.00468*** (0.00633)	−0.03390****** (0.01302)	−0.01138**** (0.00648)	−0.01075*** (0.01561)
Effect given largest deficiency weight (0.222)	−0.05426** (0.02203)	−0.01505 (0.01621)	0.01040 (0.01405)	−0.07526*** (0.02890)	−0.02527* (0.01439)	−0.02387 (0.03466)

Notes: Standard errors are clustered at the facility level and shown in parentheses. In Panel B, the effects are relative to having a P4P program that does not reward deficiencies.

* p<.1,

** p<.05,

*** p<.01

Table 8

Heterogeneous Effects on the Probability of Having Inspection Deficiencies at the Immediate Jeopardy Level by Facility Characteristics

Panel A: Coefficient Estimates						
	Non-Chain	**Chain**	**For-Profit**	**Non-Profit**	**Low Medicaid**	**High Medicaid**
P4PDefic	−0.00103 (0.00879)	0.00651 (0.00996)	0.00886 (0.0102)	−0.00143 (0.00900)	0.00119 (0.00645)	0.0241 (0.0244)
P4PDeficWeight	−0.00810 (0.0444)	−0.0448 (0.0500)	−0.0603 (0.0489)	−0.0148 (0.0494)	−0.0167 (0.0335)	−0.162 (0.113)
P4PDeficQualifier	−0.00547 (0.0126)	−0.0136 (0.0114)	−0.0269** (0.0114)	0.0222 (0.0153)	−0.00348 (0.00905)	−0.0355 (0.0254)
Constant	0.0504** (0.0240)	0.0755*** (0.0224)	0.0947*** (0.0251)	0.0144 (0.0180)	0.0517*** (0.0165)	0.0674 (0.0447)
Time FEs	Yes	Yes	Yes	Yes	Yes	Yes
Facility Covariates	Yes	Yes	Yes	Yes	Yes	Yes
N	237180	281057	353459	134926	354434	163803
R2	0.00168	0.000859	0.00122	0.00139	0.00106	0.00154
Panel B: Range of Effects						
Effect given smallest deficiency weight (0.10)	−0.00081 (0.00444)	−0.00448 (0.00500)	−0.00603 (0.00489)	−0.00148 (0.00494)	−0.00167 (0.00335)	−0.01617 (0.01129)
Effect given largest deficiency weight (0.222)	−0.00180 (0.00986)	−0.00994 (0.01111)	−0.01338 (0.01086)	−0.00328 (0.01098)	−0.00370 (0.00744)	−0.03590 (0.02506)

Notes: Standard errors are clustered at the facility level and shown in parentheses. In Panel B, the effects are relative to having a P4P program that does not reward deficiencies.

* p<.1,

** p<.05,

*** p<.01