HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 16/24

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

CR. Tamara Konetzka, Meghan M. Skira & Rachel M. Werner

August 2016

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

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

R. Tamara Konetzka[a], Meghan M. Skira[b1], 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

Acknowledgements: 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.

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

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

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, staffing levels, and inspection deficiencies, 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

JEL Codes: I11, I18

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 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, Woodard et al. (2006); Rosenthal and Frank (2006); Mehrotra, Damberg et al. (2009); Van Herck, De Smedt et al. (2010); Emmert, Eijkenaar et al. (2012); Eijkenaar (2013); Eijkenaar, Emmert 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, staffing levels, and inspection deficiencies, affect improvements in those measures. We also analyze how nursing homes respond to the use of simple, dichotomous thresholds used 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, Fernandopulle et al. 2004; Van Herck, De Smedt et al. 2010; Emmert, Eijkenaar et al. 2012; Eijkenaar 2013; Eijkenaar, Emmert 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, staffing ratios, and inspection deficiencies 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 a deficiency 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 and in response to relatively larger weights put on deficiencies among high-Medicaid census facilities. 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 in Section VII 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 2007). Despite this frequent use and high cost of nursing home care, quality of care in nursing homes has long presented a policy challenge (Institute of Medicine 1986). Major regulatory policies aimed at improving nursing home quality were implemented in 1987 under the 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, Williams et al. 1993; Shorr, Fought et al. 1994; Castle, Fogel et al. 1996; Fries, Hawes et al. 1997; Mor, Intrator et al. 1997; Snowden and Roy-Byrne 1998), a follow-up report by the Institute of Medicine in 2000 concluded that significant problems remain (Wunderlich and Kohler 2000).

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

[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.

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, Feng 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] Appendix Table 1 provides details about the construction of the weights. We focus on the weights put on clinical outcomes, staffing ratios, and inspection deficiencies 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. 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

[3] Other quality measures used in P4P programs include overall occupancy, Medicaid occupancy, consumer satisfaction, and culture change. See Werner, Konetzka et al. (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 FY2009, but were based on performance in the prior fiscal year. Given the program details were announced in 2008 and add-ons were based on FY2008 outcomes, we define the starting date of Colorado's program as July 2008. For add-ons distributed in FY2010, Colorado made some adjustments to their program. These changes were announced in 2009 and add-ons for FY 2010 were based on FY2009 outcomes. Thus, we allow these changes to be reflected in the weights starting in July 2009.

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 et al. (2010) which presents an application of the multitasking model of Holstrom and Milgrom (1991) to study P4P in the California physician medical group setting. Like Mullen et al. (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 $q = (q_1, q_2, \dots, q_J)$. *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 $y = (y_1, y_2, \dots, y_K)$. 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:

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

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,…,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

[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.

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:

$$\frac{\delta C}{\delta q_j} = 0, \qquad j = 1,2, \dots, J$$

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]

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

$$= \beta + \sum_{k=1}^{K} r_k Pr(y_k \geq T_k) - C(q)$$

$$= \beta + \sum_{k=1}^{K} r_k [F_k(\mu_k(q) - T_k)] - C(q)$$

The first order condition yields:

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

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,…,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).

[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, Feng et al. 2004; Miller, Mor 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 $E[R(y)] + \alpha B(q) - C(q)$ where $0 < \alpha < 1$.

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]

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:

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

and the nursing home's expected profits are:

$$E[R(y)] - C(q) =$$

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

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

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

$$\frac{\delta C}{\delta q_j} = f_1(\mu_1(q) - T_1)\frac{\delta \mu_1}{\delta q_j}\sum_{k=2}^{K} r_k[F_k(\mu_k(q) - T_k)]$$

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

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 comes 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 days. In addition, short-

[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.

stay patients are typically covered by Medicare, while long-stay patients are typically covered by Medicaid. As in prior work (Werner, Konetzka et al. 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, Moore 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, Fries et al. 1994), ADL scale (Morris, Fries et al. 1999), and Clinically Complex Scale (Kidder, Rennison 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, Yi 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]

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.

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, staffing ratios, and inspection deficiencies. 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:

[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.

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

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 currently has a P4P program in place; $P4PClinical_{s,t}$ 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.[12] $X_{j,s,t}$ is a vector of facility characteristics, τ_t are time fixed effects (where time is measured in quarters), γ_j are facility 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.

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:

$$\Pr(Defic_{j,s,t} = 1) = \beta P4PDefic_{s,t} + \delta P4PDeficWeight_{s,t} + \omega P4PDeficQualifier_{s,t} + \varphi X_{j,s,t} + \tau_t + \gamma_j + \varepsilon_{j,s,t}$$

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). $P4PDefic_{s,t}$ is an indicator for whether the nursing home is in a state that rewards deficiencies in its P4P program. We omit the $P4P_{s,t}$ indicator because all P4P states except Kansas rewarded deficiencies in some form during our sample period, which does not allow us to 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:

[12] In practice, the weights on the performance measures we consider are usually less than 0.40.

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

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.[13] $X_{j,s,t}$, τ_t, γ_j, and $\varepsilon_{j,s,t}$ are as defined above.

The inclusion of time fixed effects in the above specifications controls for any systematic trends in clinical quality, staffing, or inspection deficiencies 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.[14] 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, Woolhandler et al. 2001; Hillmer, Wodchis et al. 2005; Comondore, Devereaux 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.

VI. RESULTS

[13] 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.
[14] The 75 percent of residents covered by Medicaid threshold was chosen because it corresponds to the 75th percentile of the distribution.

Table 4 presents the coefficient results from the estimation where clinical quality measures are the outcome of interest. 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*. 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. Table 5 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.[15]

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 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. 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.

Table 6 shows the coefficient results where the probability of having any deficiencies or the probability of having deficiencies at the immediate jeopardy level is the outcome of interest. While we find a negative coefficient on the weight put on deficiencies for both outcomes, the coefficient is not significant at conventional levels. We do find, however, that using deficiencies as a qualifier for P4P bonus receipt is associated with a significant decrease in the probability of having deficiencies at any level of severity. Table 7 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 as well as whether the state uses deficiencies to disqualify nursing homes from

[15] Table 5 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.

P4P bonuses.[16] Relative to nursing homes in states that do not reward lack of deficiencies (i.e. all states without P4P and the state of Kansas), nursing homes in states that use deficiencies as a P4P qualifier experience a 3.7 percentage point decrease in the probability of deficiencies at any level of severity and a non-significant decrease in the probability of deficiencies at the immediate jeopardy level.[17] These results suggest using deficiencies to disqualify nursing homes from the P4P bonus rather than putting weights on deficiencies is an 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.

Table 8 presents coefficient estimates where staffing hours per resident day are the outcome of interest. 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. The weight put on staffing levels has no significant effect on total staffing but significantly increases the level of skilled staffing (p=0.052). Table 9 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.[18] 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.

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.

Table 10 shows 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

[16] Table 7 shows the point estimate and standard error of the linear combination of the coefficient on *P4PDefic* plus the coefficient on *P4PDeficWeight* times the actual weight as well as the point estimate and standard error of the linear combination of the coefficients on *P4PDefic* plus *P4PDeficQualifier*.

[17] 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.

[18] Table 9 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.

of all facility types, but larger in magnitude for non-profits and facilities with a smaller share of residents covered by Medicaid. When we consider the range of effects given the smallest and largest weight put on deficiencies and whether deficiencies are used as a P4P qualifier in Table 11, we find significant improvements for all facility types, but the improvements are larger for facilities historically associated with better quality of care—non-chains, non-profits, and facilities with a smaller share of Medicaid-covered patients. 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.

In Table 12, 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 a large share of residents covered by Medicaid—but, these coefficients are generally not significant at conventional levels (except for for-profit nursing homes). Table 13 shows the range of effects given the linear combination of coefficients and the observed weights. Here, we find the linear combination of coefficients suggests for-profits in states that use deficiencies as a P4P qualifier experience a significant 1.8 percentage point decrease in the probability of having a deficiency at the immediate jeopardy level relative to non-P4P states (and Kansas). We also find when the largest weight put on deficiencies observed in our data (0.22) is used, there is a 1.4 percentage point decrease in the probability of having an immediate jeopardy-level deficiency for nursing homes with large shares of Medicaid residents relative to otherwise similar facilities in non-P4P states (and Kansas). Intuitively this makes sense as the P4P bonus is added to the facility per diem for all Medicaid resident days. Thus, the larger the Medicaid resident population, the larger is the potential add-on and the larger is the incentive to improve. Given the coefficient estimates were not individually significantly different from zero, we cautiously interpret these results as suggestive evidence that 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. 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.

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 little to no significant improvement when deficiencies had a large weight. 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 and Elster 2007; Konetzka and Werner 2009; Friedberg, Safran 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 non-profit, 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 and high-Medicaid 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 low-resource, low-quality nursing homes. Thus, P4P appears to be creating an incentive for improvement even among 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.

References

Abt Associates Inc. (2001). Report to Congress: appropriateness of minimum nurse staffing ratios in nursing homes phase II final report. Baltimore, MD, Centers for Medicare and Medicaid Services.

Casalino, L. P. and A. Elster (2007). "Will pay-for-performance and quality reporting affect health care disparities?" Health Affairs **26**(3): w405-w414.

Castle, N. G., B. S. Fogel, et al. (1996). "Study shows higher quality of care in facilities administered by ACHCA members." Journal Long Term Care Administration **24**(2): 11-16.

Comondore, V. R., P. Devereaux, et al. (2009). "Quality of care in for-profit and not-for-profit nursing homes: systematic review and meta-analysis." BMJ: British Medical Journal **339**(7717): 381-384.

Eijkenaar, F. (2013). "Key issues in the design of pay for performance programs." The European Journal of Health Economics **14**(1): 117-131.

Eijkenaar, F., M. Emmert, et al. (2013). "Effects of pay for performance in health care: a systematic review of systematic reviews." Health Policy **110**(2): 115-130.

Emmert, M., F. Eijkenaar, et al. (2012). "Economic evaluation of pay-for-performance in health care: a systematic review." The European Journal of Health Economics **13**(6): 755-767.

Friedberg, M. W., D. G. Safran, et al. (2010). "Paying for performance in primary care: potential impact on practices and disparities." Health Affairs **29**(5): 926-932.

Fries, B. E., C. Hawes, et al. (1997). "Effect of the National Resident Assessment Instrument on selected health conditions and problems." Journal of the American Geriatrics Society **45**(8): 994-1001.

Grabowski, D. C., Z. Feng, et al. (2008). "Medicaid nursing home payment and the role of provider taxes." Medical Care Research and Review **65**(4): 514-527.

Harrington, C., S. Woolhandler, et al. (2001). "Does investor ownership of nursing homes compromise the quality of care?" American Journal of Public Health **91**(9): 1452-1455.

Hillmer, M. P., W. P. Wodchis, et al. (2005). "Nursing home profit status and quality of care: is there any evidence of an association?" Medical Care Research and Review **62**(2): 139-166.

Holmstrom, B. and P. Milgrom (1991). "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." Journal of Law, Economics, & Organization **7**: 24-52.

Institute of Medicine (1986). Improving the quality of care in nursing homes. Washington, DC, National Academies Press.

Kaiser Family Foundation. (2007). "Medicaid and long-term care services and supports." Retrieved 8/20/2008, from http://www.kff.org/medicaid/upload/2186_05.pdf.

Kane, R. L., G. Arling, et al. (2007). "A quality-based payment strategy for nursing home care in Minnesota." The Gerontologist **47**(1): 108-115.

Kane, R. L., C. C. Williams, et al. (1993). "Restraining restraints: changes in a standard of care." Annual Review of Public Health **14**: 545-584.

Kidder, D., M. Rennison, et al. (2002). MegaQI Covariate Analysis and Recommendations: Identification and evaluation of existing quality indicators that are appropriate for use in longterm care settings. Cambridge, MA, Abt Associates Inc.

Konetzka, R. T. and R. M. Werner (2009). "Disparities in long-term care: Building equity into market-based reforms." Medical Care Research and Review **66**(5): 491-521.

Konetzka, R. T., D. Yi, et al. (2004). "Effects of Medicare payment changes on nursing home staffing and deficiencies." Health Services Research **39**(3): 463-488.
Mehrotra, A., C. L. Damberg, et al. (2009). "Pay for performance in the hospital setting: what is the state of the evidence?" American Journal of Medical Quality **24**(1): 19-28.
Miller, E. A., V. Mor, et al. (2009). "The Devil's in the details: Trading policy goals for complexity in Medicaid nursing home reimbursement." Journal of Health Politics, Policy and Law **34**(1): 93-135.
Mor, V., K. Berg, et al. (2003). "The quality of quality measurement in U.S. nursing homes." The Gerontologist **43**(2): 37-46.
Mor, V., O. Intrator, et al. (1997). "Changes in hospitalization associated with introducing the Resident Assessment Instrument." Journal of the American Geriatrics Society **45**(8): 1002-1010.
Morris, J. N., B. E. Fries, et al. (1994). "MDS Cognitive Performance Scale." Journal of Gerontology **49**(4): M174-182.
Morris, J. N., B. E. Fries, et al. (1999). "Scaling ADLs within the MDS." Journals of Gerontology Series A: Biological Sciences and Medical Sciences **54**(11): M546-553.
Morris, J. N., T. Moore, et al. (2003). Validation of Long-Term and Post-Acute Care Quality Indicators. Baltimore, MD, Centers for Medicare and Medicaid Services.
Mullen, K. J., R. G. Frank, et al. (2010). "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers." The RAND Journal of Economics **41**(1): 64-91.
Petersen, L. A., L. D. Woodard, et al. (2006). "Does pay-for-performance improve the quality of health care?" Annals of Internal Medicine **145**(4): 265-272.
Rosenthal, M. B., R. Fernandopulle, et al. (2004). "Paying for quality: providers' incentives for quality improvement." Health Affairs **23**(2): 127-141.
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): 135-157.
Shorr, R. I., R. L. Fought, et al. (1994). "Changes in antipsychotic drug use in nursing homes during implementation of the OBRA-87 regulations." JAMA **271**(5): 358-362.
Snowden, M. and P. Roy-Byrne (1998). "Mental illness and nursing home reform: OBRA-87 ten years later." Psychiatric Services **49**(2): 229-233.
Van Herck, P., D. De Smedt, et al. (2010). "Systematic review: effects, design choices, and context of pay-for-performance in health care." BMC Health Services Research **10**(1): 247-259.
Werner, R. M., R. T. Konetzka, et al. (2010). "State adoption of nursing home pay-for-performance." Medical Care Research and Review. **67**: 364-377.
Werner, R. M., R. T. Konetzka, et al. (2013). "The effect of pay-for-performance in nursing homes: evidence from state Medicaid programs." Health Services Research **48**(4): 1393-1414.
Wunderlich, G. S. and P. Kohler (2000). Improving the Quality of Long-Term Care. Washington, D.C., Division of Health Care Services, Institute of Medicine.

Tables and Figures

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	7/2007 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 Showing Effect of P4P on Clinical Quality Outcomes

	Physically Restrained	Pressure Sores	Pain	Weight Loss	Catheter Inserted	Falls
P4P	0.00991***	-0.00218*	-0.00510***	0.00150*	-0.00162*	0.00310***
	(0.00124)	(0.00130)	(0.00192)	(0.000906)	(0.000934)	(0.000759)
P4PClinical	-0.0207***	0.00563	0.0144*	0.0175***	0.00232	0.00866***
	(0.00486)	(0.00479)	(0.00751)	(0.00324)	(0.00295)	(0.00253)
P4PClinicalWeight	0.00523	-0.0186	-0.0646***	-0.0601***	0.000724	-0.0278***
	(0.0139)	(0.0135)	(0.0207)	(0.00910)	(0.00814)	(0.00701)
Constant	0.0215**	0.220***	0.383***	0.0165	0.109***	0.0487***
	(0.0102)	(0.0176)	(0.0184)	(0.0133)	(0.0160)	(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

Notes: Standard errors are clustered at the facility level and shown in parentheses.
* p<.1, ** p<.05, *** p<.01

Table 5: Range of Effects of Programs that Reward Clinical Outcomes

	Physically Restrained	Pressure Sores	Pain	Weight Loss	Catheter Inserted	Falls
Effect given smallest clinical weight observed (0.10)	-0.02021*** (0.00362)	0.00377 (0.00359)	0.00793 (0.00563)	0.01150*** (0.0024)	0.00239 (0.00224)	0.00588*** (0.00192)
Effect given largest clinical weight observed (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: Effects are relative to having a P4P program that does not reward clinical outcomes.

* p<.1, ** p<.05, *** p<.01

Table 6: Results from Regressions Showing Effect of P4P on the Probability of Having Inspection Deficiencies

	Any Deficiencies	Any Immediate Jeopardy Deficiencies
P4PDefic	0.0278**	0.0110
	(0.0130)	(0.0128)
P4PDeficWeight	-0.118	-0.0751
	(0.0810)	(0.0618)
P4PDeficQualifier	-0.0649***	-0.0179
	(0.0155)	(0.0136)
Constant	0.763***	0.0672***
	(0.0270)	(0.0161)
Time FEs	Yes	Yes
Facility Covariates	Yes	Yes
N	518249	518237
R2	0.00534	0.00106

Notes: Standard errors are clustered at the facility level and shown in parentheses.
* p<.1, ** p<.05, *** p<.01

Table 7: Range of Effects of Programs that Reward Lack of Deficiencies

	Any Deficiencies	Any Immediate Jeopardy Deficiencies
Effect given smallest deficiency weight observed (0.10)	0.01604***	0.00345
	(0.00620)	(0.00682)
Effect given largest deficiency weight observed (0.222)	0.00166	-0.00570**
	(0.00765)	(0.00272)
Effect if deficiencies used as a P4P qualifier	-0.03710***	-0.00699
	(0.00866)	(0.00468)

Notes: Effects are relative to not having a P4P program that rewards deficiencies (i.e. all non-P4P states and Kansas).
* p<.1, ** p<.05, *** p<.01

Table 8: Results from Regressions Showing Effect of P4P on Staffing Ratios

	Total Staffing HPRD	RN+LPN HPRD
P4P	0.0103	-0.000998
	(0.0256)	(0.0125)
P4PStaffing	-0.0815**	-0.0345*
	(0.0353)	(0.0182)
P4PStaffingWeight	0.199	0.153*
	(0.145)	(0.0790)
Constant	3.089***	1.326***
	(0.144)	(0.0922)
Time FEs	Yes	Yes
Facility Covariates	Yes	Yes
N	518249	518249
R2	0.0248	0.0226

Notes: Standard errors are clustered at the facility level and shown in parentheses.
* p<.1, ** p<.05, *** p<.01

Table 9: Range of Effects of Programs that Reward Staffing Ratios

	Total Staffing HPRD	RN+LPN HPRD
Effect given smallest staffing weight observed (0.10)	-0.06157**	-0.01914
	(0.02835)	(0.01410)
Effect given largest staffing weight observed (0.333)	-0.01522	0.01662
	(0.03740)	(0.01937)

Notes: Effects are relative to having a P4P program that does not reward staffing ratios.
* p<.1, ** p<.05, *** p<.01

Table 10: Heterogeneous Effects on the Probability of Having Inspection Deficiencies at Any Level of Severity by Facility Characteristics

	Non-Chain	Chain	For-Profit	Non-Profit	Low Medicaid	High Medicaid
P4PDefic	0.0217	0.0467**	0.0242	0.0281	0.0275*	0.0249
	(0.0191)	(0.0199)	(0.0162)	(0.0266)	(0.0152)	(0.0251)
P4PDeficWeight	-0.143	-0.198*	-0.0763	-0.175	-0.125	-0.0845
	(0.129)	(0.114)	(0.0958)	(0.187)	(0.0963)	(0.156)
P4PDeficQualifier	-0.0691***	-0.0752***	-0.0516***	-0.0961***	-0.0718***	-0.0471*
	(0.0252)	(0.0226)	(0.0189)	(0.0338)	(0.0193)	(0.0280)
Constant	0.775***	0.779***	0.818***	0.723***	0.771***	0.870***
	(0.0391)	(0.0387)	(0.0351)	(0.0491)	(0.0340)	(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	0.00604	0.00504	0.00412	0.00773	0.00694	0.00278

Notes: Standard errors are clustered at the facility level and shown in parentheses.
* p<.1, ** p<.05, *** p<.01

Table 11: Range of Heterogeneous Effects on Inspection Deficiencies at Any Level of Severity by Facility Characteristics

	Non-Chain	Chain	For-Profit	Non-Profit	Low Medicaid	High Medicaid
Effect given smallest weight observed (0.10)	0.00737 (0.00864)	0.02698*** (0.00997)	0.01654** (0.00805)	0.01056 (0.01127)	0.01498** (0.00723)	0.01642 (0.01213)
Effect given largest weight observed (0.222)	-0.01008 (0.01314)	0.00289 (0.00965)	0.00722 (0.00860)	-0.01081 (0.01905)	-0.00024 (0.00920)	0.00611 (0.01482)
Effect if deficiencies used as a P4P qualifier	-0.04741*** (0.01651)	-0.02845*** (0.01077)	-0.02746*** (0.00994)	-0.06798*** (0.02109)	-0.04430*** (0.01196)	-0.02226* (0.01292)

Notes: Effects are relative to not having a P4P program that rewards deficiencies (i.e. all non-P4P states and Kansas).
* p<.1, ** p<.05, *** p<.01

Table 12: Heterogeneous Effects on the Probability of Having Inspection Deficiencies at the Immediate Jeopardy Level by Facility Characteristics

	Non-Chain	Chain	For-Profit	Non-Profit	Low Medicaid	High Medicaid
P4PDefic	0.00288	0.0167	0.0214	-0.0152	0.00533	0.0471
	(0.0187)	(0.0182)	(0.0205)	(0.0128)	(0.0128)	(0.0386)
P4PDeficWeight	-0.0288	-0.0959	-0.123	0.0603	-0.0379	-0.278
	(0.0922)	(0.0877)	(0.0969)	(0.0667)	(0.0633)	(0.180)
P4PDeficQualifier	-0.00934	-0.0239	-0.0395*	0.0359**	-0.00762	-0.0585
	(0.0208)	(0.0190)	(0.0212)	(0.0178)	(0.0144)	(0.0392)
Constant	0.0503**	0.0757***	0.0945***	0.0143	0.0516***	0.0686
	(0.0240)	(0.0224)	(0.0251)	(0.0180)	(0.0166)	(0.0446)
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	0.000868	0.00123	0.00143	0.00106	0.00160

Notes: Standard errors are clustered at the facility level and shown in parentheses.
* p<.1, ** p<.05, *** p<.01

Table 13: Range of Heterogeneous Effects on Inspection Deficiencies at the Immediate Jeopardy Level by Facility Characteristics

	Non-Chain	Chain	For-Profit	Non-Profit	Low Medicaid	High Medicaid
Effect given smallest weight observed (0.10)	-0.000004 (0.00987)	0.00711 (0.00972)	0.00913 (0.01107)	-0.00913 (0.00659)	0.00153 (0.00679)	0.01938 (0.02102)
Effect given largest weight observed (0.222)	-0.00351 (0.00464)	-0.00459 (0.00373)	-0.00588* (0.00344)	-0.00178 (0.00411)	-0.00310 (0.00312)	-0.01448** (0.00618)
Effect if deficiencies used as a P4P qualifier	-0.00645 (0.00918)	-0.00718 (0.00569)	-0.01806*** (0.00520)	0.02076* (0.01245)	-0.00229 (0.00644)	-0.01137 (0.00754)

Notes: Effects are relative to not having a P4P program that rewards deficiencies (i.e. all non-P4P states and Kansas).
* p<.1, ** p<.05, *** p<.01

Appendix

Appendix Table 1: Details Regarding Weight Construction

State	Program and Weight Description
Colorado	*General information*: 100 total points available. *Clinical outcome weight calculation*: -Add-ons distributed in FY2009 were based on FY2008 clinical outcomes. 27 points total allocated to clinical measures (9 points for high-risk pressure ulcers; 9 points for physical restraint use; 9 points for pain). -Thus, 27% weight put on clinical outcomes in FY 2008. -Add-ons distributed in FY2010 were based on FY2009 clinical outcomes. 25 points total allocated to clinical measures (5 points for high-risk pressure ulcers; 5 points for physical restraint use; 5 points for pain; 5 points for urinary tract infections; 5 points for falls). -Thus, 25% weight put on clinical outcomes in FY 2009. *Staffing ratio weight calculation:* -No points allocated to staffing ratios. *Inspection deficiency weight or qualifier calculation:* -No facility with "substandard quality of care deficiencies" on a regular annual, complaint, or any other Colorado Department of Public Health and Environment survey is considered for the bonus. -Thus, inspection deficiencies used as a qualifier.
Georgia	*General information:* Maximum of a 3% per diem add-on available. -Composed of a staffing incentive that is a 1% add-on based on staffing ratios and a separate 2% add-on based on a 10 total point performance score. *Clinical outcome weight calculation:* -6 of the 10 points that are part of the 2% add-on are allocated to clinical measures (1 point for high-risk long-stay residents who have pressure ulcers; 1 point for long-stay physical restraint use; 1 point for long-stay residents with pain; 1 point for short-stay residents with pain; 1 point for residents who receive flu vaccine; 1 point for low-risk long-stay residents who have pressure ulcers). -Thus, 60% of the 2% add-on incentive is based on clinical outcomes, which means 40% of the total 3% add-on is based on clinical outcomes. *Staffing ratio weight calculation:* -1% add-on requires staffing hours are at least 2.5 hours per resident day. -Thus, 33% of the total 3% add-on is based on staffing ratios.

	Inspection deficiency weight or qualifier calculation: -Facility placed on the Special Focus List produced by CMS cannot earn the add-on until its next standard or compliant survey does not have a deficiency over Level E in scope and severity; and, the facility's second standard or compliant survey after being placed on the list does not have a deficiency over Level E in scope and severity. -Thus, inspection deficiencies used as a qualifier.
Iowa	*General information*: 11 total points available. *Clinical outcome weight calculation*: -No points allocated to clinical outcomes. *Staffing ratio weight calculation:* -Facility can earn up to 2 points if their CMI-adjusted nursing hours are at or above the 75th percentile (1 point if the hours fall between the 50th and 75th percentiles). -Thus, the weight put on staffing ratios is 18.2%. *Inspection deficiency weight or qualifier calculation:* -Facility with a deficiency-free survey receives 2 points (1 point for regulatory compliance with the survey but not deficiency free). -Thus, the weight put on inspection deficiencies is 18.2%.
Kansas	*General information*: 9 total points available. *Clinical outcome weight calculation*: -No points allocated to clinical outcomes. *Staffing ratio weight calculation:* -Facility can earn up to 2 points if its CMI-adjusted nurse staffing ratio is greater than or equal to 120% of the state median (1 point if the ratio is between 110 and 120% of the state median). -Thus, the weight put on staffing ratios is 22.2%. *Inspection deficiency weight or qualifier calculation:* -No points allocated to inspection deficiencies and deficiencies are not used as a qualifier.
Minnesota	*General information*: 100 total points available in both FY 2006 and FY 2007. *Clinical outcome weight calculation*: -In FY 2006, a facility could earn up to 40 points if its Quality Indicator (QI) score was above a certain threshold. (QI score ranged from 0-100 points and was based on 24 different quality indicators). -Thus, the weight put on clinical outcomes was 40% in FY 2006. -In FY 2007, a facility could earn up to 35 points if its QI score was above a certain threshold.

	-Thus, the weight put on clinical outcomes was 35% in FY 2007. *Staffing ratio weight calculation:* -No points allocated to staffing ratios in FY 2006. -In FY 2007, 10 points allocated to CMI-adjusted staffing hours. Thresholds required to receive the points varied with facility type (standard, hospital-attached, and boarding care homes). -Thus, the weight put on staffing ratios was 10% in FY 2007. *Inspection deficiency weight or qualifier calculation:* -In both FY 2006 and 2007, up to 10 points could be earned if a facility's deficiencies were all below Level F in scope and severity (5 points if highest deficiencies were at Level F or G). -Thus, the weight put on inspection deficiencies was 10%.
Ohio	*General information*: 9 total points available. *Clinical outcome weight calculation*: -No points allocated to clinical outcomes. *Staffing ratio weight calculation:* -Facility with nursing hours per resident day above the state average receives 1 point. -Thus, the weight put on staffing ratios is 11.1%. *Inspection deficiency weight or qualifier calculation:* -1 point can be earned if there are no health deficiencies with scope and severity greater than Level E on the facility's most recent standard survey. -An additional 1 point can be earned by a facility that is deficiency-free on its most recent standard survey. -Thus, the weight put on inspection deficiencies is 22.2%
Oklahoma	*General information*: 10 total points available. *Clinical outcome weight calculation*: -Facility can receive 1 point if it is above the 50^{th} percentile on the following measures: falls, catheters, physical restraints, weight loss, and pressure ulcers. -Each of the clinical measures is percentile-ranked within the state and then combined for a composite percentile score to determine if the facility meets the threshold to receive the 1 point. -Thus, the weight put on clinical outcomes is 10%. *Staffing ratio weight calculation:* -Facility can receive 1 point if its nursing staff per resident day is above the 50^{th} percentile. -Thus, the weight put on staffing ratios is 10%.

	Inspection deficiency weight or qualifier calculation: -Facility can earn 1 point for either being deficiency free or for having no deficiencies worse than Level D in scope and severity in care-related areas and no deficiencies worse than Level E in scope and severity in non-care related areas on its state survey. -Thus, the weight put on inspection deficiencies is 10%.
Utah	*General information*: Not a point-based program. *Clinical outcome weight calculation*: -Clinical outcomes do not factor into the bonus. *Staffing ratio weight calculation:* -Staffing ratios do not factor into the bonus. *Inspection deficiency weight or qualifier calculation:* -To qualify, a facility must not have deficiencies that reach an "immediate jeopardy" level at the most recent re-certification survey. -Thus, inspection deficiencies are used as a qualifier.