HSR Health Services Research

DOI: 10.1111/1475-6773.12426
RESEARCH ARTICLE

Pay for Performance in Medicaid: Evidence from Three Natural Experiments

Meredith B. Rosenthal, Mary Beth Landrum, Jacob A. Robbins, and Eric C. Schneider

Objective. To examine the impact of pay for performance in Medicaid on the quality and utilization of care.

Data Sources. Medicaid claims and encounter data in three intervention states (Pennsylvania, Minnesota, and Alabama) and three comparison states.

Study Design. Difference-in-difference analysis with propensity score-matched comparison group. Primary outcomes of interest were Healthcare Effectiveness Data and Information Set (HEDIS)-like process measures of quality, utilization by service category, and ambulatory care–sensitive admissions and emergency department visits.

Principal Findings. In Pennsylvania, there was a statistically significant reduction of 88 ambulatory visits per 1,000 enrollee months compared with Florida. In Minnesota, there was a significant decrease of 7.2 hospital admissions per thousand enrollee months compared with Wisconsin. In Alabama, where incentives were not paid out until the end of a 2-year waiver period, there was a decline of 1.6 hospital admissions per thousand member months, and an increase of 59 ambulatory visits per 1,000 enrollees compared with Georgia. No significant quality improvements in intervention relative to control states.

Conclusions. Our findings are mixed, with no measurable quality improvements across the three states, but reductions in hospital admissions in two programs. As states move to value-based payment for patient-centered medical homes and Accountable Care Organizations, lessons learned from these pioneering states should inform program design.

Key Words. Pay for performance, Medicaid, value-based purchasing in provider relationships, shared savings

Pay-for-performance (P4P) programs for health care providers are now widely implemented in the hope of improving the quality of care and, increasingly, to control costs. The Affordable Care Act of 2010 is accelerating use of pay for performance in the Medicare payment system. Since October 2012, Medicare payments to hospitals have been adjusted based on measures of the quality of

care and patient experience initially, with cost measures entering in 2014 (Rau 2012). The Medicare program will phase in physician pay for performance—specifically targeting both cost and quality of care—with payment modifiers for physicians in groups of 100 or more beginning in 2015 (Federal Register 2012). Over the past decade, state Medicaid programs started to use P4P programs and related efforts to improve care delivery, especially in primary care (Kuhmerker 2007).

Despite the enthusiasm of policy makers, a number of literature reviews have underscored the lack of empirical support for pay for performance (Dudley et al. 2004; Petersen et al. 2006; Rosenthal and Frank 2006; Scott et al. 2011). P4P initiatives can have diverse program designs, which could lead to heterogeneous effects. Even with the proliferation of P4P programs among private and public payers, there have been few comparisons of different program schemes.

Some of the earliest published evaluations of pay for performance in health care were experiments conducted in Medicaid managed care. Two early, randomized studies by Hillman et al. (1998, 1999) examined the use of performance feedback and financial bonuses on the quality of preventive care in a Medicaid health maintenance organization (HMO). Neither initiative resulted in quality improvement despite substantial bonus potential (10–20 percent of capitation).

Several more recent studies have also examined the implementation of pay for performance in Medicaid managed care. In California, each of the seven Medicaid-focused health plans participating in the Local Initiative Rewarding Results Collaborative developed their own unique incentive programs that varied on many key dimensions, including the mix of provider and member incentives and the performance requirement necessary for receiving payments (Felt-Lisk and Smieliauskas 2006). Improvements on well-baby visit HEDIS scores ranged from 4 percent to 35 percent, and improvements on well-adolescent care HEDIS measures ranged from 8 percent to 12 percent (Highsmith and Rothstein 2006). However, within LIRR plans, the plans with the greatest adolescent well-care visits improvement did not implement any incentives until after the end of the program, suggesting that financial incen-

---

Address correspondence to Meredith B. Rosenthal, Ph.D., Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston MA 02115; e-mail: mrosenth@hsph.harvard.edu. Mary Beth Landrum, Ph.D., is with the Harvard Medical School, Boston, MA. Jacob A. Robbins, A.B., is with the Brown University, Providence, RI. Eric C. Schneider, M.D., M.Sc., is with The Commonwealth Fund, New York, NY.

tives may not have been responsible for the changes. Similar incentive programs were introduced by Hudson Health Plan, a Medicaid-focused plan in New York State. Studies looking at the impact of Hudson Health Plan's incentives for timely childhood immunization and evidence-based diabetes process and outcome measures found only modest improvements in care despite extensive outreach and quality improvement support for participating practices (Chien, Li, and Rosenthal 2010). Notably, all of the published evidence on pay for performance in Medicaid comes from the managed care setting and does not involve direct payments by state agencies to providers.

Expanding the evidence on pay for performance in Medicaid is important for at least three reasons. First, the Medicaid context is different from commercial insurance and Medicare in important ways, including the nature of patient clinical and social needs, churning of insurance coverage, and access issues that arise from low payment levels (Kaiser Commission on Medicaid and the Uninsured 2013). Medicaid covers some of the poorest individuals and families in the nation, and Medicaid patients have significantly worse health status than the privately insured. In addition, there is a significant amount of turnover in insurance coverage, with 20–40 percent of Medicaid recipients failing to reenroll every year (Merrill and Rosenbach 2006). Due to low payment rates, about 20 percent of physicians do not accept new Medicaid patients (Borck, Cheh, and Lucy 2011). Among physicians who have at least one Medicaid patient, about 60 percent derive less than 20 percent of their practice revenue from Medicaid patients. These unique aspects of Medicaid may make P4P less effective. For example, if Medicaid patients only make up a small fraction of physicians' panels, Medicaid P4P programs may provide only weak incentives. Or if enrollees are enrolled in Medicaid for only a few months, physicians may be unable to reap any benefits from changing their practice styles.

Second, with many states in the midst of a broad expansion of Medicaid eligibility, improving the delivery and controlling the cost of care to the Medicaid population is more than ever a critical national priority. Third, with the proliferation of different P4P designs, it is important to distinguish between effective and ineffective designs, which can provide lessons for the future design of state initiatives.

In this study, we empirically examine the impact of prototypical P4P programs initiated by Medicaid agencies in three states: Pennsylvania, Minnesota, and Alabama. The three programs are among the pioneers in implementation of pay for performance in Medicaid, having initiated distinctive programs that rely on a diverse array of performance measures, incentive

designs, and payment amounts. The programs were leading contemporaneous examples of Medicaid P4P programs that directed financial incentives to physicians, as opposed to hospitals or managed care organizations. Two of the programs focus on individual physician incentives, with the third focusing on physician group incentives. The performance measures used across the states include structural measures related to being a medical home, quality of care, and utilization. As these and other Medicaid agencies adopt or update P4P programs, analysis of the impact of these early experiences can yield evidence about the effectiveness of alternative approaches.

Description of Interventions and Comparison Sites

The three interventions profiled here were selected for study for two principal reasons. First, these study sites all have several years of experience with provider (as opposed to health plan) P4P programs. Second, the three programs have a similar focus on physicians but offer very distinct models of incentives: one is a "medical home" structural incentive and shared savings model (Alabama), the second rewards intermediate health outcomes in two different ways for fee-for-service and managed care (Minnesota), and the third rewards collaboration with disease management and management of chronically ill patients based on process measures of quality (Pennsylvania). We should note that states in which we identified P4P programs to study also had initiatives for other providers such as hospitals and managed care plans, but we chose to focus on physician incentives.

We evaluate each of the three efforts using a difference-in-difference approach that compares changes in outcomes in intervention sites to those experienced in selected control states. Because each state has a degree of latitude to design its Medicaid program, there is heterogeneity in terms of payment system, eligibility, spending per enrollee, and program benefits. Due to this heterogeneity, we compare each intervention state to a matched comparison state selected to be as similar as possible to the intervention state. We selected comparison patients from similar Medicaid populations in states where pay for performance had not yet been introduced (this eliminated about half of all states from being selected as a comparison group). In selecting comparison states, we required an exact match by type of program; comparison states needed to have a sizeable population of enrollees in the same type of funding arrangement as the "intervention" state (fee-for-service, primary care

case management [PCCM], managed care), comparable use of disease management, and the absence of another major Medicaid program (including, but not limited to pay for performance) that would affect the outcomes of interest. To the extent possible, we also looked for comparability in terms of geography, population demographics and physician supply.

Pennsylvania and Florida

The Pennsylvania Office of Medical Assistance Programs (OMAP) purchases services through contracts with managed care organizations, an enhanced PCCM vendor and under a traditional, fee-for-service system for nearly 1.8 million Pennsylvania residents. The PCCM program, Access-Plus, serves 280,000 members, 32,000 of whom have chronic diseases covered by the vendor's disease management program. In Pennsylvania, managed care is mandatory in urban counties where HMOs serving Medicaid are relatively plentiful and voluntary where there are few HMOs; Access-Plus is the only option for Medicaid members in some rural counties with no managed care plans.

Office of Medical Assistance Programs has been actively engaged in value-based purchasing approaches for a number of years. Since 2000, OMAP has published a report card that compares Medicaid HealthChoices managed care plans on HEDIS and CAHPS results. More recently, OMAP has been developing pay-for-performance programs for managed care plans, hospitals, and physicians. Pay for performance was introduced in Access-Plus through OMAP's disease management vendor in 2006. The vendor has developed a program to reward providers for active participation in five disease management programs (coronary artery disease [CAD], congestive heart failure [CHF], asthma, chronic obstructive pulmonary disease [COPD], and diabetes).

Payment incentives in Pennsylvania are at the physician level and focus on two separate goals—support of chronic disease management and optimal chronic care (see Table 1). Providers have a variety of incentives to enroll patients in disease management. There are also incentives for delivering optimal treatment, measured by HEDIS "process of care" measures: beta blockers for patients with CHF, aspirin for patients with diabetes and CAD, a controller medication for patients with asthma, and LDL for patients with diabetes.

The initial implementation of the program began in September 2005, with 337 participating providers and 10,222 patients. By August 1, 2008, there were over 1,450 participating providers, caring for more than 15,000 chroni-

Table 1: Description of Pay for Performance Programs

	Pennsylvania–Florida	*Minnesota–Wisconsin*	*Alabama–Georgia*
Preintervention	November 2003–October 2005	July 2004–July 2006	October 2002–September 2004
Postintervention	November 2005–October 2007	July 2006–December 2007	October 2004–September 2006
Participating Medicaid population	"Access-Plus" PCCM patients with one of five chronic diseases: CAD, CHF, asthma, COPD, and diabetes	"BTE" diabetics enrolled in managed care plans	"Patient 1st" PCCM patients
Providers	Physicians	Medical groups	Physicians
Financial incentives	Practitioners are given a one-time $200 payment to support the program, $30 to provide contact information to the vendor and encourage patient participation, $40 to identify candidates for disease management, and $60 for completing a chronic care feedback form and care plan. There are also incentives for delivering optimal treatment: $17 per process per patient per year for delivering recommended treatment. The "process" performance measures are HEDIS specifications: beta blockers for CHF, aspirin for diabetes and CAD, a controller medication for asthma, LDL for diabetes	Participating practices receive $100 per diabetic Medicaid enrollee if the practices exceed a threshold percentage of diabetics who receive "optimal" diabetes care. A patient is said to receive ODC if and only if they meet all of the following standards: (1) HbA1c less than or equal to 7.0%; (2) blood pressure less than 130/80 mmHg; (3) LDL less than 100 mg/dl; (4) daily aspirin use for ages 41–75; and (5) documented tobacco free. The threshold was 20% in 2006 and 2007	A tiered case management fee pays providers the following specific amounts per member per month for providing items related to providing a medical home: $0.45 for being an EPSDT Provider, $0.10 Vaccines for Children Participant, $0.10 Medical Home CME, $0.85 24/7 Coverage, $0.30 Hospital Admitting Privileges, $0.10 In-Home Monitoring, $0.50 InfoSolutions (E-prescribing) Participant, $0.05 Receives Electronic Notices, $0.15 Receives Electronic Education Materials. In addition, there is a shared savings program, which shares 50 percent of any documented savings with primary care physicians in Patient 1st

Continued

Table 1. *Continued*

	Pennsylvania–Florida	*Minnesota–Wisconsin*	*Alabama–Georgia*
Sample	Individuals enrolled in the PCCM program are identified as having any of the following five chronic diseases: asthma, CHF, CAD, COPD, or diabetes. Enrollees under the age of 18 or older than 64, with less than 6 months of Medicaid eligibility, and pregnant women are excluded	Diabetics enrolled in Medicaid HMOs. Enrollees under the age of 18 or older than 64, with less than 6 months of Medicaid eligibility, and pregnant women are excluded	Individuals enrolled in Patient 1st. Enrollees under the age of 18 or older than 64, with less than 6 months of Medicaid eligibility, and pregnant women are excluded
No. of physicians	Pennsylvania—2,206 Florida—2,206	Minnesota—746 Wisconsin—746	Alabama—905 Georgia—905
No. of enrollees	Pennsylvania—14,408 Florida—72,187 (10,570 matched)	Minnesota—7,287. Wisconsin—10,436 (5,690 matched)	Alabama—90,763 Georgia—220,829 (48,162 matched)

BTE, Bridges to Excellence; CAD, coronary artery disease; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; HbA1c, Hemoglobin A1c; HMO, health maintenance organization; ODC, optimal diabetes care; PCCM, primary care case management.

cally ill patients, with over $3.2 million in incentives paid to enrolled providers over and above standard fees.

We selected Florida as a comparison state because of Florida's large Medicaid PCCM population from which a matched comparison group could be identified. States that were demographically and geographically closer to Pennsylvania either did not have PCCM programs, had contemporaneous quality improvement efforts in the state Medicaid program, or in one case, were missing physician identifiers that were unique over time.

Minnesota and Wisconsin

During the study period, the Minnesota Department of Human Services (DHS) served over 750,000 Medicaid enrollees. Approximately half of all MHCP enrollees were covered by a Medicaid HMO program. These HMO enrollees are the focus of the intervention and our study.

Previous cost containment and quality improvement programs in Minnesota have taken place in the commercial insurance environment. In 1988, the Buyers Health Care Action Group was formed by a coalition of large private Minnesota employers and initially contracted with health plans with a focus on quality improvement. Later, the Buyers Health Care Action Group contracted directly with groups of providers focused around primary care physicians and hospitals ("care systems"). In 1993, the Institute for Clinical Systems Improvement was established, which focused on developing and publishing best practice health care guidelines for Minnesota physicians. In 2002, the Minnesota Community Measurement group formed to gather data about health plan performance and cost. Every year Minnesota Community Measurement releases a health care quality report measuring the performance of medical groups on performance measures. Finally, in 2004 a "Smart Buy Alliance" was formed by both private and public health care purchasers in Minnesota to purchase health care with a focus on quality.

Beginning in July of 2006, the DHS joined private employers through the Smart Buy Alliance as a participant in the Bridges to Excellence (BTE) Diabetes Care link for the managed care population, with payments to providers beginning in 2007. BTE is a national, employer-led P4P program that provides a model and infrastructure for health plans, employers, and other payers to reward doctors for optimal care of diseases such as diabetes and heart disease. BTE in Minnesota, however, is a customized program that utilizes existing Minnesota-based infrastructure and the optimal diabetes care (ODC) measure. The most important difference between the Minnesota BTE pro-

gram and other BTE initiatives is that it rewards providers based on an "all-or-none" measure (the ODC); thus, if the practice does not achieve all five components in the measure, they don't qualify for P4P. For enrollees in managed care programs, DHS is participating in BTE under the same model as private employers.

The main performance measure of BTE in Minnesota is a measure of ODC, described in detail in Table 1. Unlike many other P4P programs that pay for process of care (such as the percentage of individuals receiving HbA1c tests), BTE also pays for outcomes—it is not enough that patients get regular blood tests; they must have blood sugar and lipid levels controlled. The performance incentives paid by BTE are at the practice level. Medical groups receive $100 per diabetic Medicaid enrollee served by those medical groups or clinics that meet a predetermined threshold of performance.

We selected the Wisconsin Medicaid program as the comparison state for Minnesota. Wisconsin has a substantial Medicaid managed care population, and it is demographically and geographically similar to Minnesota.

Alabama and Georgia

Alabama Medicaid covers health care services for over 900,000 individuals, about 400,000 of whom are in its PCCM program, Patient 1st. The agency received federal approval in October 2004 for two significant changes to Patient 1st (see Table 1). The first change, implemented in January 2005, was the introduction of a tiered case management fee. The tiered case management fee pays providers for specific items related to providing a medical home.

The second significant change in Alabama was the institution of a shared savings program, which shares 50 percent of any documented savings with primary care physicians in the PCCM program, Patient 1st. The State provides primary care physicians with quarterly reports that show them how they compare to other physicians on three risk-adjusted measures of performance: generic medication use, emergency department utilization, and number of office visits. In addition, the physicians are profiled on their actual versus expected total allowed charges (referred to as "efficiency" in this context). These performance measures are calculated using Medicaid claims data. Shared savings are allocated based on how physicians score on the performance and efficiency metrics relative to their peers. Physicians ranking in the lowest quartile overall are ineligible for shared savings payments. The first shared savings payments totaling $5.76 million were distributed in April 2007

for performance April 1, 2005 through March 31, 2006. Additional payments of $4.7 million were distributed in 2009.

We selected Georgia as a comparison state for reasons of geographic similarity, demographics, and the existence of a large PCCM Medicaid population. In our analyses, we only use patients enrolled in the PCCM program in Georgia to form a comparison group.

DATA AND EMPIRICAL STRATEGY

Data Sources

The principal data sources for the three intervention sites are the billing systems from the respective Medicaid agencies. For the three comparison states, claims data are taken from CMS's Medicaid Analytic Extracts. The administrative data for each site include Medicaid enrollment data and membership information, which contains demographic variables for enrollees, and inpatient, outpatient, emergency department, and drug claims (fee-for-service and PCCM) or encounter data (managed care).

Study Cohorts

Table 2 describes the panel characteristics of the study cohorts for each of the three state pairs. Because we hypothesized that the greatest impact of the programs would be on adults, we excluded enrollees under the age of 18 or older than 64, although some states extended their efforts to children. We also excluded individuals with less than 6 months of Medicaid eligibility in a given year, enrollees who are eligible for Medicare, and pregnant women. In Pennsylvania and Florida, only individuals diagnosed with any of five chronic conditions were included. In Minnesota and Wisconsin, only diabetics were included.

Study Outcomes

We evaluate the effectiveness of the three programs based on claims-based measures of quality and utilization available in both intervention and comparison cohorts. The key utilization variables we examine for all three programs are the number of ambulatory visits, the number of emergency department visits, and the number of hospital admissions. In addition, we examine hospital and emergency room use for ambulatory care sensitive (ACS) conditions.

Table 2: Panel Characteristics Before and after Propensity Score Matching

	Prematching						*Postmatching*					
	PA	*FL*	*MN*	*WI*	*AL*	*GA*	*PA*	*FL*	*MN*	*WI*	*AL*	*GA*
Variables												
Age	44.3	47.1**	43.1	41.7**	38.0	37.1**	44.3	44.1	43.1	44.5**	38.0	38.0
% male	38.2	35.0**	34.5	32.3	28.9	30.2	38.2	37.4	34.5	34.9	28.9	26.8
% white	88.6	44.0**	–	–	54.1	42.6**	88.6	86.3**	–	–	54.1	54.5
% black	6.3	24.0**	–	–	38.6	49.4**	6.3	8.6**	–	–	38.6	38.6
No. of patients	2.6	9.0**	2.9	3.1	51.9	49.1**	2.6	1.8**	2.9	2.8	51.9	47.7
Hospital admissions	0.06	0.08**	0.05	0.03**	0.03	0.03**	0.06	0.06	0.05	0.05	0.03	0.04
Ambulatory visits	0.57	0.81**	0.81	0.79	0.46	0.49**	0.57	0.55*	0.81	0.79	0.46	0.47
Emergency room visits	0.07	0.11**	0.11	0.08**	0.09	0.10**	0.07	0.08**	0.11	0.08**	0.08	0.08
Emergency ACS visits	0.02	0.03**	0.03	0.02**	0.009	0.01**	0.02	0.02**	0.03	0.02	0.009	0.009
PQI admissions	0.01	0.02**	0.010	0.006**	0.006	0.005**	0.01	0.02*	0.01	0.01	0.006	0.005
% diabetes	46.8	49.0*	100	100	17.1	13.0**	46.8	47.6	100	100	17.1	18.5
% asthma	23.0	22.0	6.4	6.3	5.8	4.7**	23.0	22.6	6.4	6.4	5.84	6.32
% CHF	9.7	15.4**	5.3	4.0	4.4	4.6	9.74	9.39	5.3	5.4	4.38	4.32
% COPD	25.0	34.0**	6.39	4.4*	9.7	7.4**	25.0	25.6	6.4	8.0	9.66	10.6
% CAD	29.0	33.0**	8.86	7.9	6.3	6.8	29.0	30.2	8.9	8.7	6.32	6.28
No. of physicians	2,206	4,416	746	1,254	905	11,590	2,206	2,206	746	746	905	905

Note. $^{**}p < .01$, $^{*}p < .05$.
ACS, ambulatory care sensitive; CAD, coronary artery disease; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; PQI, prevention quality indicator.

ACS conditions are a set of medical conditions (such as diabetes and asthma) in which appropriate primary care can reduce the likelihood of preventable hospitalizations or emergency room visits (Falik et al. 2001). Our ACS ED measure was developed by NYU's Center for Health and Public Service Research (Billings 2002), and our measure of ACS hospitalizations is prevention quality indicators (PQIs) (Quality 2012).

We use HEDIS specifications to define claims-based quality measures that vary across the intervention. We measure the impact of the Pennsylvania P4P program on the quality of care by examining three process of care measures that are incentivized through the program: beta blockers for CHF, a controller medication for asthma, and LDL screenings for diabetes. We also examine a fourth related measure, HbA1c tests for diabetics.

Under BTE, Minnesota practices are incentivized to control patients' blood sugar and cholesterol, and therefore to closely monitor these levels by performing routine tests and screenings. The key quality measures we use to assess the effect of BTE is the percentage of diabetics who receive HbA1c blood tests and LDL screenings in a given year.

In Alabama, we examine the following process measures of quality that we expect will be influenced by the adoption of "medical home" features: chlamydia screening, cervical screening, appropriate asthma medication, HbA1c testing for diabetes, beta blockers for CHF, and LDL screening for people with diabetes.

Analytic Approach

Our primary analytic approach is to examine dependent variables before and after the intervention alongside a contemporaneous comparison group (a "difference-in-difference" approach). To ensure a valid comparison, we further refine our cohorts so that the intervention and comparison groups are similar along important dimensions using a propensity score approach (Rosenbaum and Rubin 1984; D'Agostino 1998; Rubin 1998). While our analysis is conducted at the patient-level, the intervention targets physician behavior so we undertake propensity score matching of physicians and include characteristics of their patients in the model. We begin by creating datasets for each natural experiment which include the broader groups of patients targeted by the intervention and comparison enrollees and their claims and eligibility data for the first year of analysis. We then aggregate these data to the physician level, creating variables that summarize the characteristics of each physician's panel. We also calculate average utilization at baseline.

We estimate a logistic regression model linking the probability of being in the experimental group with average patient characteristics (age, gender, race, disability status, chronic conditions including AMI, diabetes, asthma, CHF, COPD, and CAD), physician panel characteristics (number of Medicaid patients in panel at baseline), and baseline utilization measures (inpatient, ambulatory visits, ER visits, ACS ER visits, and ACS hospitalizations), and compute the predicted probability of being exposed to pay for performance based on physician panel and average patient characteristics from this model. For each physician that is exposed to pay for performance, we find the physician from the no P4P group that has the closest probability of having been exposed, that is, propensity score matching. The final cohort consists of patients of the treated and matched physicians, followed throughout the pre- and postintervention period. However, we do not restrict our sample to those continuously enrolled and thus some enrollees only contribute to either the pre- or postintervention measurement period.

Because we match physicians according to aggregate patient characteristics and cannot find perfect matches, we expect small residual differences in observed patient and physician panel characteristics to remain in the matched samples. We fit regression models to control for these residual observed differences.

While physicians are the focus in each of these interventions, the relevant outcomes are measured at the patient level. We thus fit models at the patient (or patient-month) level and use a generalized estimating equations approach that takes account of the clustering of patients by physician and the repeated observations at both the patient and provider level, while allowing us to incorporate patient-level covariates as risk adjustors.

Let i index physician; j index the Medicaid enrollee; and p denote the measurement period and y_{ipj} represent a patient-level outcome measure. For example, y_{ipj} may be total Medicaid hospital admissions in month p for patient j treated by physician i. We estimate models with the following general structure:

$$h(e(y_{ipj})) = \alpha + \beta_1 Treated + \beta_2 post + \beta_3 Treated * post + \beta_4 x_{ipj} + \beta_5 z_i$$

where $h()$ is a link function, $\mathbf{x}_{ipj}$ a set of (centered) patient characteristics (age, gender, race, dummies for five chronic conditions [diabetes, asthma, CHF, COPD, CAD], Medicaid basis of eligibility, Medicaid maintenance assistance status), and $\mathbf{z}_i$ is a set of physician panel characteristics (number of Medicaid patients, average age of Medicaid panel, average sex of Medicaid panel, average race of Medicaid panel). The primary estimate of interest is the coefficient β_3, the difference-in-difference estimator. In the case of binary patient mea-

sures such as whether or not the patient received HbA1c testing, $h()$ is a logit link and we assume y follows a binomial distribution. For utilization measures, we use a linear model with physician fixed effects to remove the influence of unobservable but time-invariant physician characteristics that may bias our estimates. In nonlinear models, we calculate the difference-in-difference estimate on the original scale, setting all covariate values at their mean (Karaca-Mandic, Norton and Dowd 2012).

Sensitivity Analyses

We perform a number of sensitivity and robustness checks to our results to ensure that our results are robust to different estimation and matching specification. We fit ordinary least squares (OLS) models without physician fixed effects because of concern about statistical power and the validity of the implicit assumption that only within-physician variation is relevant to causal inference. For utilization variables that are count variables, we estimate negative binomial regressions as an alternative to the OLS models. Because the events modeled in the negative binomial regressions are relatively rare, we were unable to include physician fixed effects in these models. For the HEDIS quality regressions, we run linear probability models with and without physician fixed effects as alternative specifications. Finally, we experiment with two alternative versions of matching: (1) matching on preintervention trends (rather than levels) in quality and utilization and (2) coarsened exact matching (Iacus, King, and Porro 2009).

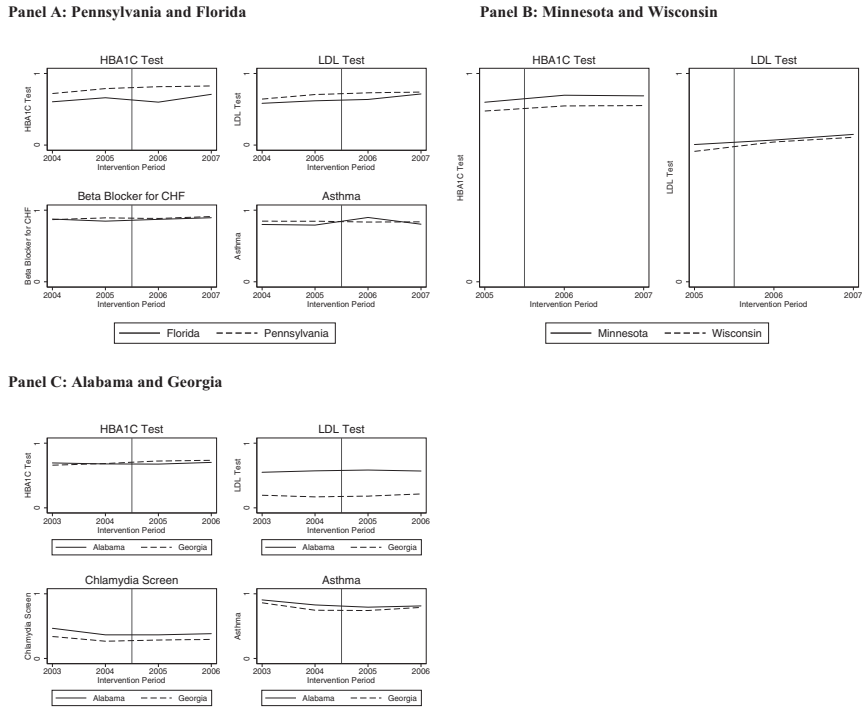
All results from these alternative specifications are presented in Appendix SA2–SA7.

RESULTS

Matching

Prior to matching, physicians in Florida care for larger numbers of Medicaid patients who are older and more likely to be racial or ethnic minorities than those in Pennsylvania. Baseline utilization rates are also higher in Florida relative to Pennsylvania, and Florida physicians care for more Medicaid patients (Table 2). Medicaid patients with diabetes were also slightly older, and baseline utilization rates were higher in Minnesota relative to Wisconsin. Physicians in Alabama cared for substantially larger numbers of Medicaid patients relative to those in Georgia. All of these differences are substantially dimin-

Figure 1: Comparison of HEDIS Quality Measures. (A) Pennsylvania and Florida. (B) Minnesota and Wisconsin. (C) Alabama and Georgia



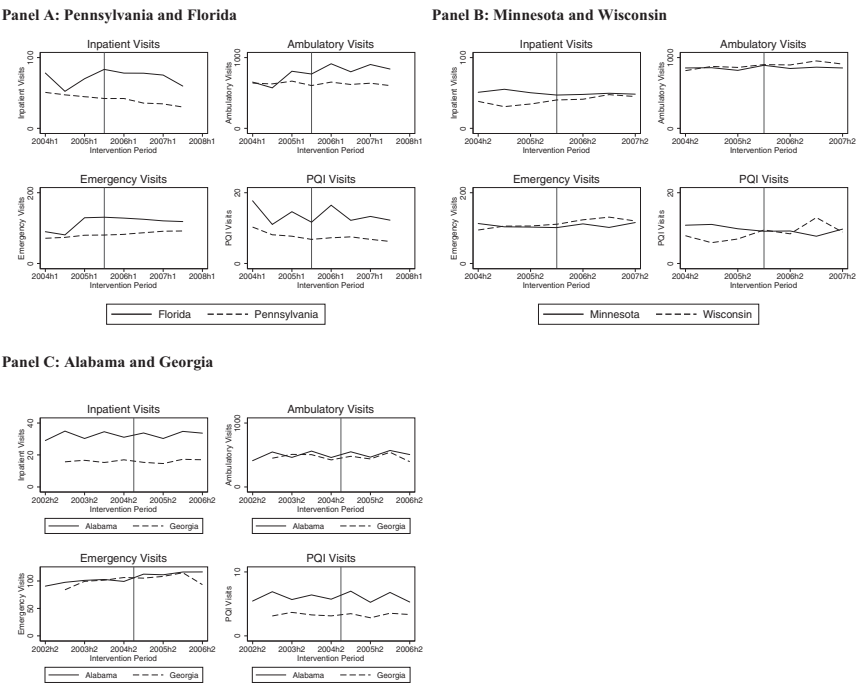
ished in the matched samples. However, even after matching, Minnesota's diabetic Medicaid population is slightly younger at baseline and has more emergency visits and PQI hospital admissions compared to Wisconsin.

Figures 1 and 2 display trends in quality and utilization in the matched samples. Preintervention trends were similar in intervention and control states for most of the outcomes.

Pennsylvania and Florida

Table 3A displays the difference-in-difference regression results of the effect of P4P on adherence to HEDIS process of care measures. Compared with Florida, Pennsylvania's P4P program is not associated with a larger increase in adherence in the HBA1C, LDL, or beta blocker measure. For the asthma measure, Pennsylvania shows a statistically significant

Figure 2: Comparison of Utilization Measures. (A) Pennsylvania and Florida. (B) Minnesota and Wisconsin. (C) Alabama and Georgia



decrease in adherence of 6 percentage points compared with Florida. This is a surprising result, considering the main focus of Pennsylvania's P4P program was management of chronic diseases. However, as Figure 1B shows, Pennsylvania's asthma adherence was already at a very high level before the P4P program (85 percent adherence) and remained roughly flat over the time period, while Florida's adherence was at a slightly lower level and increased during the time period.

Table 3B shows results from the difference-in-difference utilization regressions. Pennsylvania experienced a statistically significant decrease of 88 ambulatory visits per 1,000 enrollee months compared with Florida, a roughly 15 percent decrease. Once again this seems to be driven by physicians in Florida. Figure 1B shows that Pennsylvania's ambulatory visits are roughly flat over the period, while Florida's are increasing. No other utilization measures are statistically significant.

Table 3: Results from HEDIS Quality and Utilization Regressions

	Pennsylvania & Florida	*Minnesota & Wisconsin*	*Alabama & Georgia*
(A) HEDIS quality			
HBA1C for diabetes	0.194 (0.112)	0.0575 (0.144)	−0.232 (0.141)
Marginal effect	3.45	0.69	−4.91
LDL for diabetes	−0.0981 (0.114)	−0.00205 (0.0971)	0.0187 (0.156)
Marginal effect	−2.01	−0.0004	0.39
Beta blockers for CHF	−0.0628 (0.160)		−0.0175 (0.110)
Marginal effect	−0.65		−0.18
Preferred asthma treatment	−0.455* (0.181)		−0.143 (0.132)
Marginal effect	−6.11		−2.12
Cervical cancer screening			−0.0832 (0.0728)
Marginal effect			−1.73
Chlamydia screening			−0.0818 (0.177)
Marginal effect			−1.76
(B) Utilization			
Inpatient admissions	−6.77 (8.09)	−7.24* (3.51)	−1.61* (0.651)
Ambulatory visits	−88.0** (25.7)	−32.8 (26.1)	59.1** (6.27)
Emergency visits	−6.18 (9.13)	−9.45 (8.34)	−3.74 (2.35)
ACS ER visits	4.77 (3.63)	−5.24 (3.22)	8.17e-02 (0.409)
PQI admissions	1.19 (2.22)	−2.34 (1.58)	−0.224 (0.299)
PQI diabetes visits		−1.58 (1.12)	

Note. $^{**}p < .01$, $^{*}p < .05$.

All quality results are derived from logistic regressions at the enrollee year level, restricted to individuals with propensity score-matched physicians. Log odds, standard errors, and marginal effects are reported. Standard errors are clustered at the physician level. The "Marginal Treatment Effect" represents the estimated effect of the intervention on the dependent variable, calculated at the means of the covariates; the units are percentage points. Covariates include patient and physician characteristics and linear time trends. Dependent variables are quality measures based on HEDIS specifications with adaptations as described in the text. "HBA1C" denotes whether an individual identified as having diabetes received an HbA1c test during the measurement year; "LDL" marks whether an individual identified as having diabetes received an LDL test during the measurement year; "Beta CHF" identifies whether an individual diagnosed with congestive heart failure has taken beta blockers during the measurement year; "Asthma" denotes whether individuals diagnosed with asthma are taking a "preferred" asthma medication during the measurement year; "Cervical" denotes whether women 21–64 have been screened for cervical cancer in the measurement year; "Chlamydia" indicates whether women ages 15–25 have been screened for chlamydia in the measurement year.

All utilization results are OLS with physician fixed effect regressions at the enrollee month level, restricted to individuals with propensity score-matched physicians. Standard errors are clustered at the physician level. The units are "visits" per member month. Covariates include patient and physician characteristics and a linear time trend. Hospital visits, ambulatory visits, and emergency room visits are measured using HEDIS specifications (NCQA 2012). Hospital admissions do not include mental health or pregnancy related diagnoses. Ambulatory visits include office and home visits by a doctor or nurse. Emergency ACS (or "Ambulatory Care Sensitive") visits are a quality measure developed by NYU's Center for Health and Public Service Research (Billing 2002) and identify emergency room visits that are potentially preventable through better primary care management. "PQI Total" are hospitalizations that are identified as preventable through better primary care management (AHRQ 2012). "PQI Diabetes" are hospitalizations with diabetes diagnoses that are identified as preventable through better primary care management (AHRQ 2012).

Minnesota and Wisconsin

We observe no significant differences in trends in quality between intervention and control sites (Table 3A).

Table 3B displays utilization results from the physician fixed effect regressions, our preferred specification. The intervention is associated with a statistically significant decrease in hospital utilization, with Minnesota showing a pre-post decrease of 7.2 visits per thousand enrollee months compared to Wisconsin (a 15 percent decrease over baseline). No other coefficients are statistically significant.

Alabama and Georgia

Table 3A shows results from the difference-in-difference regressions on the quality of care. For all six quality measures, the intervention shows no statistically significant effect. Table 3B shows results from the utilization regressions. Alabama had an increase in ambulatory visits of 59 per thousand enrollee months, a 13 percent increase, and a decrease in inpatient utilization of 1.6 visits per thousand member months, a decline of 4.6 percent, relative to Georgia.

Sensitivity Analyses

Detailed results from all but our main specification are included in the Appendix SA2–SA7. In summary, we find that when matching on trend as opposed to the levels of our covariates, we do not find a decrease in inpatient admission in Minnesota; other conclusions remain, while coarsened exact matching decreases power and thus significance due to the reduced number of eligible matches. Removal of fixed effects (using our preferred matching approach) has some impact on the detailed findings for utilization measures. In the utilization models without fixed effects, the decrease in inpatient utilization in Minnesota is not significant, but we do find a significant decrease in PQI admissions (in a negative binomial model). In Alabama, the increase in inpatient admissions is insignificant in the model without fixed effects. There are no significant improvements in the quality measures in any of the alternate models.

DISCUSSION

Medicaid is arguably the most important health care payer in the United States, covering more people than any other program and insuring some

of our most vulnerable citizens. Historically, Medicaid agencies have responded to budgetary pressures by keeping fees low or carving out contracts to managed care organizations. Over the last decade, however, a number of states have begun to move toward more active value-based purchasing in their provider relationships. Our analysis paints a somewhat mixed picture of the impact of these nascent programs on related utilization measures, with little measurable gain in process measures of quality but modest and potentially important reductions in inpatient use in Minnesota and Alabama, where chronic disease management were important targets of the incentive program.

Pay for performance on the whole has posed significant implementation challenges and, perhaps as a result, many impact evaluations have yielded negative findings. With regard to the process measures of quality, our findings are consistent with previous studies of pay for performance across a wide variety of payers and delivery system contexts. However, both Minnesota and Alabama rewarded a broad set of performance measures including intermediate health outcomes and for those states the reduction in hospital use may be an important signal that these efforts are paying off in fewer preventable exacerbations of chronic illness. Our data are too limited to examine the cost-effectiveness of these efforts, but this is an important topic for future research.

Medicaid may be particularly challenged to make effective use of payment incentives due to the low baseline payment rates and the likelihood that safety net providers have less capacity to put in place performance improvement initiatives. Moreover, many state Medicaid agencies are themselves underresourced and face regulatory hurdles. Designing effective programs, educating providers about their goals and requirements, and distributing bonus payments in a timely manner may be challenging. In Alabama, for example, the state could not make bonus payments until the end of each 2-year waiver period, which almost surely dampened the potential impact. Given these constraints, increased technical and financial assistance from the Federal government through the Centers for Medicare and Medicaid Innovation and other mechanisms should be directed at the states during this critical era of Medicaid expansion.

The introduction of pay for performance is associated with a decrease in inpatient utilization in both Minnesota, whose program mainly targeted chronic illness management, and Alabama, which targeted broader measures of cost and utilization. This is suggestive evidence that both quality and cost incentive measures can be effective. It is interesting that we did not find P4P

programs led to higher quality of care, even with significant incentives in Minnesota and Pennsylvania. One possibility is due to the difficulty of measuring the quality of care in claims data, which is imperfectly captured by our HEDIS variables. Our results are robust to different regression specifications as well as matching techniques.

Our study has a number of important limitations. First and foremost, we examine natural experiments using a quasi-experimental approach that is subject to bias if our comparison groups are not well matched. For some measures and state pairs, the graphical presentation of trends (see Figures 1 and 2) suggests that the assumption the states were on the same trend before the intervention may not be correct. In some cases changes in comparison states drove the effects in ways that seem less likely to be a function of the success or failure of pay for performance. Each state health care market and Medicaid program differs in many important ways and while we selected the states that we considered to be the best possible match, we acknowledge that differences between intervention and control states remain. In addition, there are many sources of potentially confounding policy and market changes that could have occurred contemporaneously with the P4P programs in our three states, and we cannot be sure such changes are not influencing our results. We know, for example, that at the same time the physician P4P was implemented, Pennsylvania allowed hospitals to compete for grants for quality improvement. It is possible that these quality improvements spilled over to our results, in which case we have identified the combined effect of both programs. In Minnesota, many private employers also participated in the BTE program. If there are spillovers in physician practice styles between patients, then the results show the combined effect of Medicaid and commercial P4P incentives.

Second, we have only claims data to assess impact. Both Pennsylvania and Minnesota relied on clinical data for some aspects of their incentive program, but we are unable to examine changes in clinical measures (e.g., blood pressure control) because we do not have access to these data for either the states that implemented pay for performance or the comparison states. Claims data are, however, a very reliable source of information about utilization and we are able to examine reliably whether the efforts to improve quality in those states had a measurable impact on acute care utilization. In Minnesota and Wisconsin, we rely on encounter rather than claims data, and these may cause underestimation of utilization and process measures of quality because encounter data are not relied upon for payment.

A third limitation to our analysis is the well-known churning of Medicaid eligibility. Because we have a fragmented picture of patient care as patients enter and exit the program, our results may poorly capture both quality and cost consequences of improved performance.

Finally, no sample of P4P programs in Medicaid can be fully generalizable to all other states given the many dimensions on which Medicaid and the health care system vary. Nonetheless, this group of states and programs offers diversity by region, Medicaid program (PCCM vs. managed care), and payment model. Many of these characteristics overlap with other state Medicaid programs. For example, of the 25 state Medicaid programs with physician incentives in 2006, 20 of them used HEDIS-like measures. Five states incorporated measures of efficiency, while there were also a number of states that gave incentives for disease management and using electronic records (Kuhmerker 2007).

Now more than ever, Medicaid programs need to find ways to obtain the greatest value for increasingly constrained resources and increased demand. Lessons learned from state efforts to improve quality through pay for performance will provide a valuable foundation for current and future initiatives that now typically involve more holistic payment reform along with new models of care delivery in the form of medical homes and Accountable Care Organizations. The challenges associated with successful implementation of these reform efforts are greater than the relatively narrow P4P programs we studied and likely to exceed the capacity of both the state agencies and providers involved. For this reason and the fact that it has more of its own resources at stake in the expansion population, the Federal government should increase technical and financial assistance to states through the Centers for Medicare and Medicaid Innovation and other mechanisms.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Financial support for this research was provided by the Agency for Healthcare Research and Quality. Dr. Rosenthal was the Principal Investigator on the grant and acquired the data for the study. Drs. Rosenthal, Landrum, and Schneider participated in the design of the study, interpretation, and presentation of the findings. Mr. Robbins was the data analyst and assisting with drafting and revising the manuscript.

Disclosures: None.

Disclaimers: None.

REFERENCES

Agency for Healthcare Research and Quality (AHRQ). 2012. *Prevention Quality Indicators.* Rockville, MD: Agency for Research and Quality.

Billings, J. 2002. *ACS Conditions.* New York: NYU Wagner School of Public Service.

Borck, R., V. Cheh, and L. Lucy. 2011. *Recent Patterns in Children's Medicaid Enrollment: A National View.* Princeton, NJ: Mathematica Policy Research.

Chien, A., Z. Li, and M.B. Rosenthal. 2010. "Improving Timely Childhood Immunizations through Pay for Performance in Medicaid-Managed Care." *Health Services Research* 45 (6): 1934–47.

D'Agostino, R. 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Nonrandomized Control Group." *Statistics in Medicine* 17: 2265–81.

Dudley, R.A., A. Frolich, D.L. Robinowitz, J.A. Talavera, P. Broadhead, and H.S. Luft. 2004. *Strategies to Support Quality-based Purchasing: A Review of the Evidence.* Rockville, MD: Agency for Healthcare Research and Quality.

Falik, M., J. Needleman, B.L. Wells, and J. Korb. 2001. "Ambulatory Care Sensitive Hospitalizations and Emergency Visits: Experiences of Medicaid Patients Using Federally Qualified Health Centers." *Medical Care* 39 (6): 551–61.

Federal Register. 2012. "Medicare Program; Revisions to Payment Policies under the Physician Fee Schedule." 74 FR. C. f. M. M. Services. Washington, DC. 44721: 44721 -45061.

Felt-Lisk, S., and F. Smieliauskas. 2006. *Evaluation of the Local Initiative Rewarding Results Collaborative Demonstrations: Final Report.* Washington, D.C.: Mathematic Policy Research.

Highsmith, N., and J. Rothstein. 2006. *Rewarding Performance in Medicaid Managed Care.* Hamilton, NJ: Center for Health Care Strategies.

Hillman, A.L., K. Ripley, N. Goldfarb, I. Nuamah, J. Weiner, and E. Lusk. 1998. "Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care." *American Journal of Public Health* 88 (11): 1699–701.

Hillman, A., K. Ripley, N. Goldfarb, J. Weiner, I. Nuamah, and E. Lusk. 1999. "The Use of Physician Financial Incentives and Feedback to Improve Pediatric Preventive Care in Medicaid Managed Care." *American Journal of Public Health* 104 (4): 931–5.

Iacus, Stefano.M., G. King, and G. Porro. 2009. "CEM: Software for Coarsened Exact Matching." *Journal of Statistical Software* 30: 1–27.

Kaiser Commission on Medicaid and the Uninsured. 2013. *Medicaid: A Primer.* Washington, D.C.: Kaiser Family Foundation.

Karaca-Mandic, P., E. C. Norton, and B. Dowd. 2012. "Interaction Terms in Nonlinear Models." *Health Services Research* 47(1pt1): 255–274.

Kuhmerker, K. 2007. *Pay-for-Performance in State Medicaid Programs: A Survey of State Medicaid Directors and Programs.* New York: T. C. Fund.

Merrill, A., and M. Rosenbach. 2006. "SCHIP and Medicaid: Working Together to Keep Low-Income Children Insured." Final Report Submitted to the Centers

for Medicare & Medicaid Services. Cambridge, MA: Mathematica Policy Research.
National Committee on Quality Assurance. 2012. *HEDIS 2012 Measures.* Washington, D.C.
Petersen, L., L. Woodward, T. Urech, C. Daw, and S. Sookanan. 2006. "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine* 145 (4): 265–72.
Rau, J. (2012, October 1) "Medicare Begins Latest Pay-for-Performance Effort; Hospital Readmissions Among Program Targets." *Kaiser Health News.*
Rosenbaum, P., and D. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of American Statistical Association* 97 (387): 516–24.
Rosenthal, M.B., and R. Frank. 2006. "What is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review: MCRR* 63 (2): 135–57.
Rubin, D. 1998. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8S): 757–63.
Scott, A., P. Sivey, D. Ait Ouakrim, L. Willenberg, L. Naccarella, J. Furler, and D. Young. 2011. "The Effect of Financial Incentives on the Quality of Health Care Provided by Primary Care Physicians." *Cochrane Database of Systematic Reviews* (9): CD008451.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.
Appendix SA2: Alternate Quality Analysis Specifications.
Appendix SA3: Alternate Utilization Analysis Specifications.
Appendix SA4: Quality Regressions Matching on Trend.
Appendix SA5: Utilization Regressions—Matching on Trend.
Appendix SA6: Quality Regressions—Coarsened Exact Matching.
Appendix SA7: Utilization Regressions—Coarsened Exact Matching.