HSR Health Services Research

DOI: 10.1111/1475-6773.12409
BEST OF 2015 ACADEMYHEALTH ANNUAL RESEARCH MEETING

Higher Incentive Payments in Medicare Advantage's Pay-for-Performance Program Did Not Improve Quality But Did Increase Plan Offerings

Timothy J. Layton and Andrew M. Ryan

Objective. To evaluate the effects of the size of financial bonuses on quality of care and the number of plan offerings in the Medicare Advantage Quality Bonus Payment Demonstration.

Data Sources. Publicly available data from CMS from 2009 to 2014 on Medicare Advantage plan quality ratings, the counties in the service area of each plan, and the benchmarks used to construct plan payments.

Study Design. The Medicare Advantage Quality Bonus Payment Demonstration began in 2012. Under the Demonstration, all Medicare Advantage plans were eligible to receive bonus payments based on plan-level quality scores (star ratings). In some counties, plans were eligible to receive bonus payments that were twice as large as in other counties. We used this variation in incentives to evaluate the effects of bonus size on star ratings and the number of plan offerings in the Demonstration using a differences-in-differences identification strategy. We used matching to create a comparison group of counties that did not receive double bonuses but had similar levels of the preintervention outcomes.

Principal Findings. Results from the difference-in-differences analysis suggest that the receipt of double bonuses was not associated with an increase in star ratings. In the matched sample, the receipt of double bonuses was associated with a statistically insignificant increase of +0.034 (approximately 1 percent) in the average star rating ($p > .10$, 95 percent CI: −0.015, 0.083). In contrast, the receipt of double bonuses was associated with an increase in the number of plans offered. In the matched sample, the receipt of double bonuses was associated with an overall increase of +0.814 plans (approximately 5.8 percent) ($p < .05$, 95 percent CI: 0.078, 1.549). We estimate that the double bonuses increased payments by $3.43 billion over the first 3 years of the Demonstration.

Conclusions. At great expense to Medicare, double bonuses in the Medicare Advantage Quality Bonus Payment Demonstration were not associated with improved quality but were associated with more plan offerings.

Key Words. Pay-for-performance, health insurance, econometrics

In response to widespread concern about the low value of medical spending, pay-for-performance has quickly proliferated throughout the U.S. health care system. The Centers for Medicare and Medicaid Services (CMS) has implemented pay-for-performance for hospitals through Hospital Value-Based Purchasing (Ryan et al. 2014), for physician groups through the Physician Value-Based Payment Modifier (Ryan and Press 2014), and for dialysis providers under the End-Stage Renal Disease Quality Incentive Program. Demonstration programs are under way for nursing homes and home health providers. Numerous private payers have also initiated pay-for-performance programs during the last 15 years (Rosenthal et al. 2004; Alexander et al. 2013). Despite their widespread adoption, evidence that these programs improve quality is mixed (Town et al. 2005; Petersen et al. 2006; Rosenthal et al. 2006; Van Herck et al. 2010; Flodgren et al. 2011; Scott et al. 2011; Houle et al. 2012). In addition, basic questions about pay-for-performance programs, such as the dose–response relationship between the size of financial incentives and quality improvement, remain unanswered.

In 2012, pay-for-performance was extended to Medicare Advantage, the market for private Medicare plans, through the Quality Bonus Payment Demonstration. The Demonstration, occurring between 2012 and 2014, was initiated by the Patient Protection and Affordable Care Act. Bonus payments in the Demonstration were quite large, ranging from 3 to 10 percent of plan payments, much larger than the 1–2 percent of revenue typically at risk in other pay-for-performance programs (Van Herck et al. 2010). The structure of the Demonstration provides a unique opportunity to study some key issues in pay-for-performance. Specifically, the Demonstration used strict criteria to designate counties that were eligible for bonus payments that were twice as large as other counties. In this study, we used variation in the size of the bonus payments from these "double–bonus" counties to evaluate the relationship between the size of bonuses and quality of care in the Demonstration. The existence and size of a "dose–response" with respect to the effect of the size of a quality-based payment is critical for policy makers to consider when designing payment systems both for insurers and providers. Yet there is little current research that can guide these decisions. In addition to the dose–response with

---

Address correspondence to Andrew M. Ryan, Ph.D., Department of Health Management and Policy, School of Public Health, University of Michigan, 1415 Washington Heights, SPH II RM. M3124, Ann Arbor, MI 48109; e-mail: amryan@umich.edu. Timothy J. Layton, Ph.D., is with the Department of Health Policy, Harvard Medical School, Boston, MA.

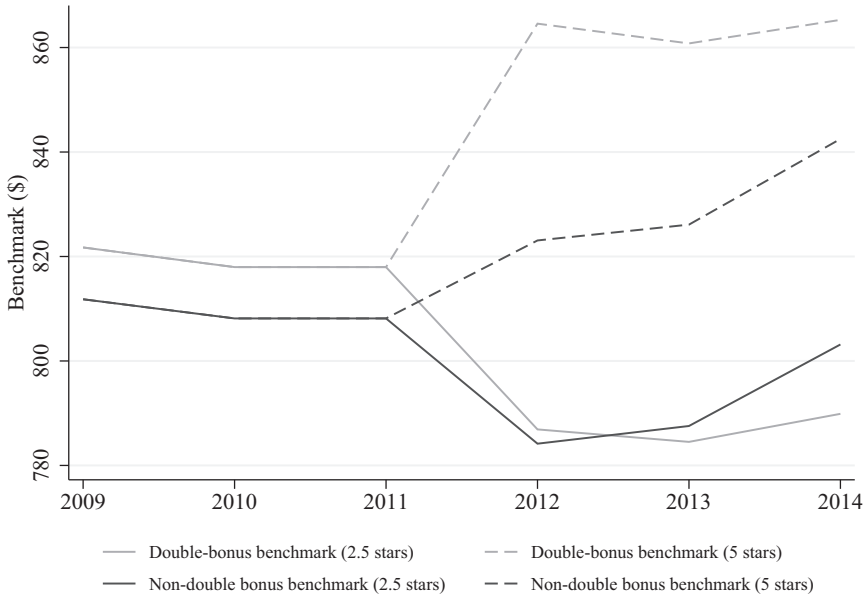
respect to plan quality, we also test whether Medicare Advantage insurers expanded plan offerings in counties that were eligible for double bonuses.

Description of the Medicare Advantage Quality Bonus Payment Demonstration

Under the Demonstration, the architecture for payment to Medicare Advantage plans remained the same, but quality-based bonuses were added. Quality was assessed using a five-star quality rating that was assigned to the plan for the prior year. Health plans received between 1 and 5 stars based on performance on more than 30 measures across five domains: preventive care and staying healthy, management of chronic conditions, health plan responsiveness and care, customer satisfaction, and telephone customer service. Star ratings are prominently listed with the name and cost-sharing information for each plan on CMS's Medicare Options Compare website. Most of the quality measures are related to provider, rather than insurer, behavior, presenting *insurers* with incentives to improve *provider* quality.

Payments to Medicare Advantage plans are based on county-specific benchmarks set by CMS. Insurers submit bids for each plan. If the bid is above the benchmark, consumers pay the difference in the form of a premium. If the bid is below the benchmark, the insurer receives a specified portion of the "shared savings" in the form of a rebate. The quality-based bonuses paid to plans under the Demonstration were made up of two components—the "benchmark bonus" (higher benchmarks for high-quality plans) and the "rebate bonus" (higher portion of shared savings)—which sum to a total bonus (see Appendix A for details). In the Demonstration, some counties are designated as "double-bonus" counties by CMS. For these counties, the benchmark bonus is doubled while the rebate bonus remains the same. A county qualified as a double-bonus county if it had lower than average fee-for-service Medicare costs in 2012, if it had a Medicare Advantage penetration rate of 25 percent or greater as of December 2009, and if it was designated as an urban floor county in 2004. Figure 1 illustrates how the Demonstration affected bonuses differently for plans with 2.5 stars or fewer and plans with five stars in double-bonus and non-double-bonus counties. In all counties, the benchmark for a low-quality (2.5 stars or lower) plan decreased by approximately the same amount due to the Demonstration. While the benchmarks for high-quality (five stars) plans increased for all counties, this increase was much larger for double-bonus counties.

Figure 1: Changes in County Benchmarks for Low-Quality and High-Quality Plans in Double-Bonus and Non-Double-Bonus Counties Due to the Medicare Advantage Quality Bonus Payment Demonstration



Notes: The figure shows changes in county benchmarks over time separately for double-bonus and non-double-bonus counties.

Conceptual Framework

We illustrate the insurer's problem by considering two hypothetical settings. First, consider a setting where the insurer could offer a set of plans with differing levels of quality and cost. The insurer can project expected revenues and costs for each plan in the county and chooses which plans to offer. If the insurer cancels lower quality plans or initiates new higher-quality plans, the overall plan quality in the county increases.

As shown in Figure 1, the Demonstration decreased revenues for low-quality plans and increased revenues for high-quality plans. Assuming plan costs are fixed over time, such a change would make marginal low-quality plans unprofitable while making marginal high-quality plans profitable, resulting in insurers canceling low-quality plans and initiating new high-quality plans. In addition, if we assume that plan costs are similar across county type

(double-bonus vs. non-double-bonus) conditional on quality, we would expect that insurers would (1) cancel low-quality plans at a similar rate in double-bonus counties and non-double-bonus counties and (2) initiate plans at a higher rate in double-bonus counties. This would imply that the overall level of plan quality in double-bonus counties would increase more than in nondouble-bonus counties. If, on the other hand, plan costs differed across county types conditional on quality, it is not clear what the change in plan quality for double-bonus counties relative to non-double-bonus counties would be. For example, if the cost of high-quality plans was greater in double-bonus counties (resulting in fewer "marginal" high-quality plans), larger bonuses may be required to make high-quality plans profitable. This could result in double bonuses having the same effect on overall quality in double-bonus counties as normal bonuses have in non-double-bonus counties.

In addition to the effect on average plan quality, double bonuses could lead insurers to offer more plans out of their portfolios of potential plans. Figure 1 shows that the benchmarks for low-quality plans in the double-bonus and non-double-bonus counties track quite closely before and after the start of the Demonstration. The benchmarks for the high-quality plans, on the other hand, are much higher in the double-bonus counties. This may result in the same number of low-quality plans becoming unprofitable and being eliminated, but a greater number of high-quality plans becoming profitable and being added. This would likely result in an increase in the overall number of plans in the county.

Now, consider a setting where each insurer offers just one plan, but the quality of the plan is endogenous. Assume that the benchmarks are always high enough to ensure that all insurers participate. In this setting, an insurer will choose the level of quality such that the marginal cost of an additional unit of quality is equal to the marginal benefit. Now, assume that the cost function with respect to quality is fixed before and after the start of the Demonstration. Then, because the Demonstration increases the revenue (benefit) a plan receives for each additional unit of quality, it should lead insurers to increase the quality of their plans. With respect to the relative change in quality in double-bonus versus non-double-bonus counties due to the Demonstration, if the cost functions with respect to quality are similar across county types, then quality should increase more in double-bonus than in non-double-bonus counties due to the Demonstration. However, if cost functions differ such that the marginal cost of quality at high levels of quality is higher in double-bonus than in non-double-bonus counties, the relative effect of the Demonstration is ambiguous.

Bringing the two settings together, if the cost of plan quality is similar across double-bonus and non-double-bonus counties, we expect double bonuses to cause overall plan quality to increase more in double-bonus counties than in non-double-bonus counties due to the Demonstration. If, however, costs differ, the relative effects of the Demonstration are ambiguous. In addition, the first setting suggests that insurers should increase the number of plans offered in double-bonus counties.

METHODS

We used publicly available data from CMS on Medicare Advantage plan quality ratings, the counties in the service area of each plan, and the county-level (quality-based) benchmarks used to construct plan payments.[1] We used these data for 2009 through 2014, the entire period over which these data are available. This provided us with 3 years of data prior to the implementation of the Demonstration in 2012 and 3 years of Demonstration data.

Because variation in bonus size is at the county level, we constructed a dataset where the unit of observation is a county-year. For each county, we used the Medicare Advantage Landscape Files to determine which plans are available in each county. We focused our analysis only on Health Maintenance Organization (HMO) and local Preferred Provider Organization (PPO) plans. All Cost, Medical Savings Account, and Demonstration plans were excluded. We excluded Private Fee-for-service plans because, while they are eligible for bonuses, their quality ratings are calculated differently from the HMO and PPO plans. We also excluded Regional PPO plans because they are paid based on regional, rather than county, benchmarks. Finally, we excluded counties that did not have star ratings from at least one plan in each year of the sample.

Our measure of quality of care is the county-level average summary star rating among rated plans.[2] Unlike the overall rating on which payment is based, the summary star rating does not incorporate quality measures related to Medicare Part D. We use the summary rating because the overall rating is not reported prior to 2011. Nonetheless, the summary rating and the overall rating are highly correlated, with a correlation coefficient of 0.91 in 2011. To assess whether double bonuses impacted plan offerings in the affected counties, our second outcome is the number of plans offered in a county.

We identified double-bonus counties using county-level data from CMS on the quality-based benchmarks. For each county, we calculated the

minimum and maximum five-star benchmark bonuses over the time period. We used these minimum and maximum bonuses to divide counties into *de facto* double-bonus and non-double-bonus counties. We defined a county as a *de facto* double-bonus county if its minimum five-star benchmark bonus exceeded 9 percent and as a non-double-bonus county if its maximum five-star benchmark bonus was less than 6 percent. In our analysis, we excluded all counties not classified as double-bonus or non-double-bonus according to these rules. This resulted in around 15 percent of counties in our dataset being excluded from our analyses. We used these *de facto* definitions rather than identifying double-bonus counties using the criteria outlined in the previous section because, while largely consistent, the classifications are not identical. We expect that insurers would respond to bonuses built into plan payments (*de facto*) rather than to the stated definition of a double-bonus county.

We used a differences-in-differences strategy to evaluate the effects of the Demonstration on quality and plan offerings between counties with normal bonuses and those with double bonuses. To implement this strategy, we estimated the following regression for county c at time t:

$$Y_{ct} = \beta_0 + \beta_1 \text{DoubleBonus}_c * \text{POST}_t + \gamma_c + \delta_t + \varepsilon_{ct} \quad (1)$$

Y_{ct} represents our study outcomes (the average star rating or the number of plans offered), DoubleBonus_c is a dummy variable equal to 1 for counties meeting our definition of a double-bonus county, Post_t is a dummy variable equal to 1 for 2012 and later, γ_c is a set of county fixed effects, δ_t is a set of year fixed effects, and ε_{ct} is the idiosyncratic error term. The coefficient of interest is β_1, the difference-in-differences estimator.

The timing of both the collection of data and the implementation of the quality-based payments is critical for our analysis. Each plan's star rating and bonuses for a given year are based on performance from up to 2 years prior (Centers for Medicare and Medicaid Services 2012) (Appendix Exhibit E1). While the Affordable Care Act was signed into law on March 23, 2010, Medicare Advantage plans did not begin to receive quality-based bonuses until 2012. It is not likely that plans could have adjusted quality in response to the law prior to the law being written, so we would not expect plans to be able to improve quality prior to 2013. However, insurers could have responded to the Demonstration in other ways that do not require such a long lag. For example, an insurer could have taken all of its low-quality plans off the market starting in 2012. Insurers could have also expanded the service area of high-quality plans to include the double-bonus counties. Due to inertia in Medicare

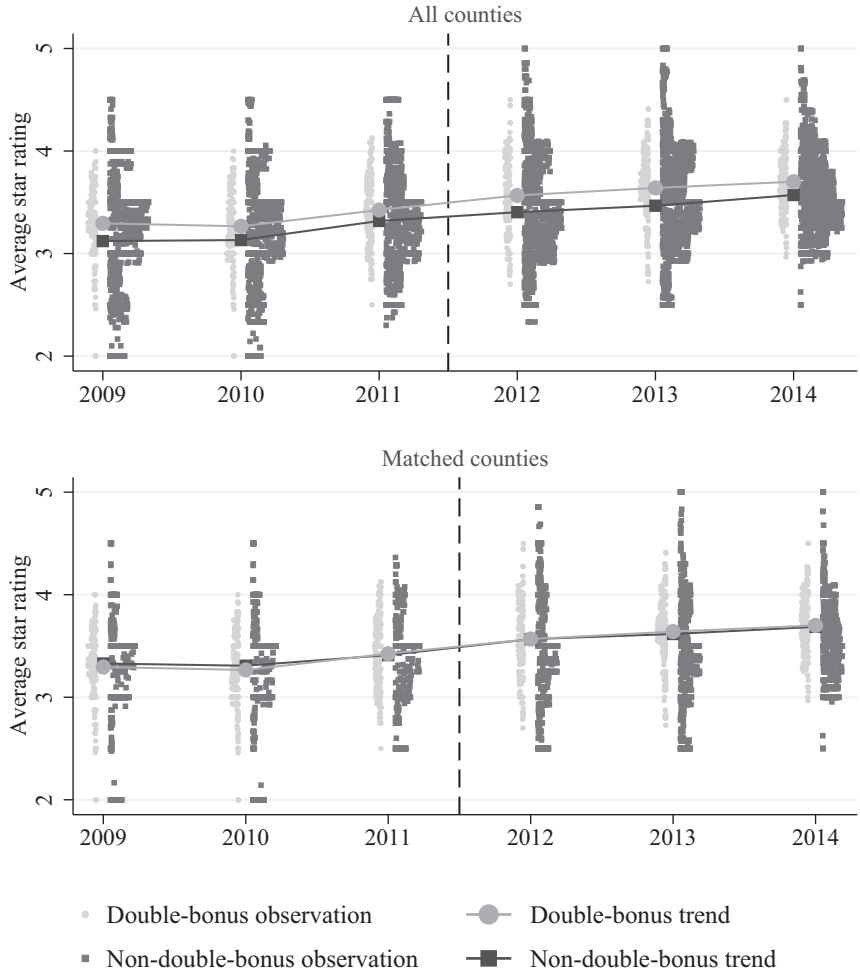
Advantage plan choice, it may be optimal for insurers to expand these high-quality plans early to attract more enrollees (Sinaiko, Afendulis, and Frank 2013). Because these strategies imply different "start dates" for the policy, we examined the impact of double bonuses across a range of the postintervention periods. Sensitivity analysis further explored whether Medicare Advantage plans expanded plan offerings immediately following the passage of the ACA (Appendix A, Exhibits E4-E7).

To accurately identify the impact of programs, difference-in-differences analysis relies on the parallel trends assumption, indicating that the outcomes for the treatment and comparison groups follow the same trajectory prior to the start of the program (Angrist and Pischke 2008). Figures 2 and 3 show that the study outcomes followed a similar trend for counties that received and did not receive the double bonuses prior to the Demonstration. However, tests of parallel trends were rejected for both outcomes. It is also important to note that to interpret β_1 as the causal effect of paying a double bonus instead of a normal bonus (i.e., a dose–response effect), it is necessary to assume that the causal effect of a normal bonus is the same across double-bonus and non-double-bonus counties (i.e., no treatment effect heterogeneity). In the Conceptual Framework above, we point out that this is equivalent to assuming that the cost of quality is similar across double-bonus and non-double-bonus counties. The intuition behind the necessity of this identification assumption is that the coefficient β_1 is a function of two parameters, γ_1 and γ_2, where γ_1 represents the causal effect of the larger bonus and γ_2 represents treatment effect heterogeneity with respect to the baseline bonus available in all counties. Because we are interested in γ_1, the causal effect of a larger bonus, we assume that $\gamma_2 = 0$, or that the cost of quality is similar across double-bonus and non-double-bonus counties.

To increase confidence that both the parallel trends and "similar cost of quality" assumptions are satisfied, we do three things. First, we graphically illustrate the preperiod levels and trends in quality in the double-bonus and non-double-bonus counties. Figure 1 shows that the levels of the benchmark payments are similar on average across these two sets of counties. Because the levels of quality are similar across these two sets of counties that have similar benchmark payments, it is unlikely that the cost of quality differed substantially.

Second, we created a comparison group of counties that did not receive double bonuses but had similar levels and trends of the preintervention outcomes. Counties that have similar levels of quality during the preintervention period are less likely to have different costs of quality. To do this, we

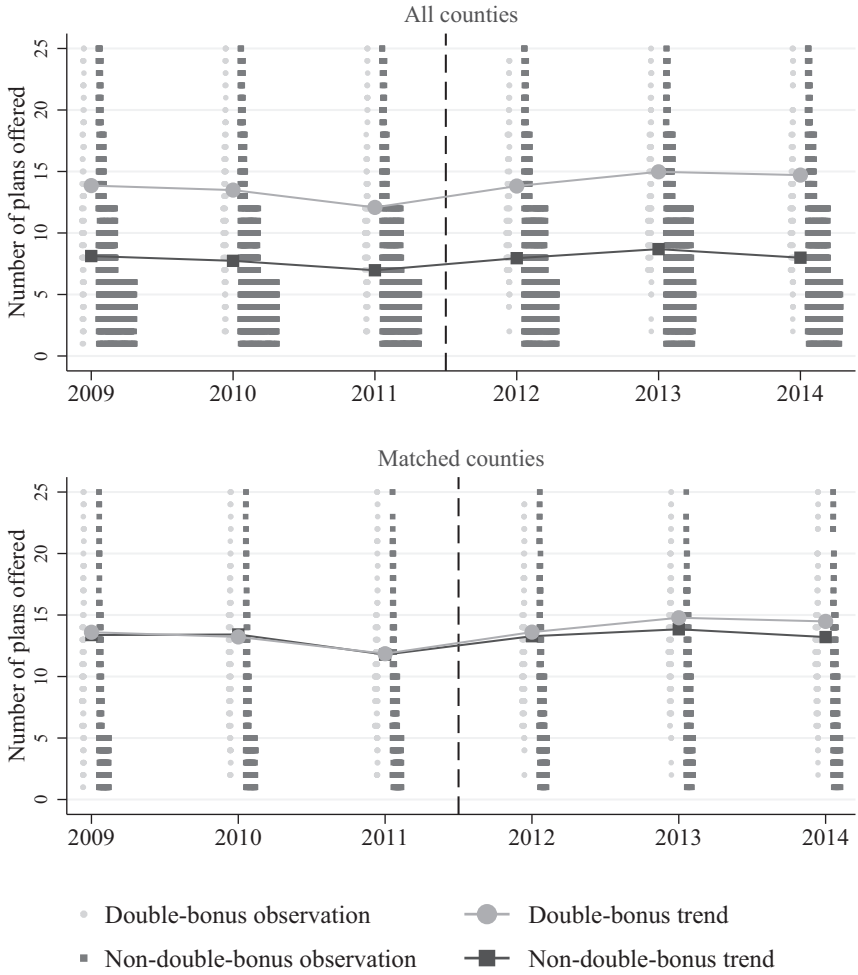
Figure 2: Quality of Care among Counties Receiving and Not Receiving Double Bonuses in the Medicare Advantage Quality Bonus Payment Demonstration



Note: The dashed line denotes the first year preceding the bonus payments in the Demonstration.

implemented a matching strategy using propensity scores, performing one-to-one matching with replacement, calipers of .01, and enforcing common support. Matching was performed separately for each outcome. Lagged levels of the outcome for each of the 3 years prior to the start of the Demonstration were the only variables used for matching. In cases where multiple compar-

Figure 3: Number of Plans Offered among Counties Receiving and Not Receiving Double Bonuses in the Medicare Advantage Quality Bonus Payment Demonstration



Note: The dashed line denotes the first year preceding the bonus payments in the Demonstration.

ison units had the exact same propensity score as a treatment unit, comparison unit n received a weight of 1/n. Consequently, our matched comparison groups included more counties than were included in our treatment group. The matching procedure was implemented in Stata using a user-written command (Leuven and Sianesi 2003). Recent research suggests that matching

can result in more accurate estimates from difference-in-differences analysis (Ryan, Burgess, and Dimick 2014).

Third, we repeated the analysis with an additional comparison group of counties that are similar to the treatment counties with respect to the criteria for double-bonus status. To do this, we isolated counties that were "close" to qualifying for double-bonus status. Specifically, we created a comparison group consisting of (1) non-double-bonus counties that were urban floor counties and had Medicare Advantage penetration between 15 and 25 percent; (2) non-double-bonus counties that were not urban floor counties and that had Medicare Advantage penetration between 25 and 35 percent. The intuition behind this robustness check is similar to the intuition behind a regression discontinuity design. It ensures that the treatment and control counties are as similar as possible with respect to the treatment group criteria, limiting concerns about our estimates of the effect of double bonuses being contaminated by treatment effect heterogeneity with respect to the baseline bonuses.[3]

To determine whether the results were driven by the elimination or introduction of small plans with little enrollment, supplemental analysis examined the effects of double bonuses on enrollment-weighted average quality and total Medicare Advantage enrollment. We also analyzed the effects of double bonuses on average performance for the separate domains used to construct the star measures in the Demonstration to determine whether the effects of double bonuses were different for some measures than for others.

RESULTS

Enrollment in Medicare Advantage plans grew substantially between 2009 and 2014, both among plans in counties that received and did not receive double bonuses (Table 1). During this period the proportion of plans with unassigned stars and plans with two or three stars decreased in all counties, while the proportion of plans with four or five stars increased. Table 1 also shows that while there were differences in preintervention quality in the treatment and control counties, the treatment and matched control counties have similar preintervention quality.

Quality of care increased at a similar rate over the study period between counties that received double bonuses and those that did not (Figure 2). The start of the Demonstration did not appear to increase quality more among those counties receiving double bonuses. However, the number of plans

Table 1: Characteristics of Counties Receiving and Not Receiving Double Bonuses

	2009			*2014*		
	Treatment	*All Control*	*Matched Control*	*Treatment*	*All Control*	*Matched Control*
N	197	1,125	532	197	1,125	532
Medicare Advantage enrollment	10,192	3,853	1,403	16,152	6,318	2,878
No. of plans	13.91	8.83	6.82	14.77	8.54	7.05
Average star rating	3.29	3.12	3.12	3.70	3.57	3.51
% 5-star plans	2.1	2.8	2.1	14.3	11.5	7.5
% 4-star plans	42.1	34.7	36.7	57.8	52.5	52.4
% 3-star plans	32.7	32.2	30.6	18.5	28.6	33.5
% 2-star plans	3.1	12.7	15.3	0.0	0.1	0.0
% plans without rating	20.0	17.6	15.3	9.4	7.2	6.6

Notes: The "control" group is all non-double-bonus counties with at least one HMO/PPO plan available for purchase. The treatment group is all double-bonus counties. The "matched control" group consists of those double-bonus counties that received a non-zero weight in the propensity score matching procedure. Data are shown for the analytic sample for the "average star rating" outcome.

offered increased in the double-bonus counties, relative to the counties not receiving double bonuses, after the start of the Demonstration (Figure 3). This relationship is apparent among both the entire sample and the matched sample.

Results from the difference-in-differences analysis suggest that the receipt of double bonuses was not associated with an increase in quality (Table 2). In the matched sample, the receipt of double bonuses was associated with a statistically insignificant increase of +0.034 (approximately 1 percent) in the average star rating ($p > .10$). Estimates for the separate postintervention periods are also small and nonsignificant. Similar effects were observed for the entire sample. In the appendix, we present very similar results for the third control group of counties that are "close" to qualifying for double-bonus status (Appendix A, Exhibit E21). Supplemental analysis found that double bonuses were not associated with higher quality for any of the separate domains that were used in the star ratings (Appendix A, Exhibits E9–E16), nor did we find evidence that double bonuses had larger effects in states with greater HMO penetration (Appendix A, Exhibit E22). Additional analysis also found that in the matched sample double bonuses were also not consistently associated with improved *enrollment-weighted* quality (Appendix A, Exhibits E17–E18).

Table 2: Estimates of the Effects of the Medicare Advantage Quality Bonus Payment Demonstration from Difference-in-Differences Models

	Average Star Rating				*Number of Plans Offered*			
	All Counties		*Matched Counties*		*All Counties*		*Matched Counties*	
	Overall Estimate	*Estimate for Each Period*	*Overall Estimate*	*Estimate for Each Period*	*Overall Estimate*	*Estimate for Each Period*	*Overall Estimate*	*Estimate for Each Period*
Overall estimate (2012–2014)	0.018 [−0.021, 0.057]	–	0.034 [−0.015, 0.083]		0.758** [0.285, 1.231]	–	0.814* [0.078, 1.549]	–
Estimate for 2012	–	0.026 [−0.013, 0.065]	–	0.022 [−0.028, 0.071]	–	0.335 [−0.075, 0.746]	–	0.298 [−0.380, 0.977]
Estimate for 2013	–	0.036 [−0.015, 0.087]	–	0.044 [−0.025, 0.114]	–	0.754** [0.188, 1.320]	–	0.905* [0.063, 1.747]
Estimate for 2014	–	−0.008 [−0.056, 0.040]	–	0.035 [−0.027, 0.097]	–	1.185** [0.561, 1.810]	–	1.238** [0.305, 2.172]
N observations	7,932	7,932	4,374	4,374	8,880	8,880	3,222	3,222
N treatment counties	197	197	197	197	198	198	195	195
N comparison counties	1,125	1,125	532	532	1,282	1,282	342	342
N total counties	1,322	1,322	729	729	1,480	1,480	537	537

Notes: 95% confidence intervals shown in brackets. Standard errors are robust to county-level clustering. Test of difference in preintervention trends between treatment and all control counties (parallel trends) rejected at $p < .05$ for the average star rating outcome and rejected at $p < .10$ for the number of plans offered outcome. Tests of parallel trends for treated and matched counties were not rejected for either outcome.
*$p < .05$, **$p < .01$.

In contrast, difference-in-differences estimates indicate that the receipt of double bonuses was associated with an increase in the number of plans offered. In the matched sample, the receipt of double bonuses was associated with an overall increase of +0.814 plans (approximately 5.8 percent) ($p < .05$). Estimates for the separate postintervention periods indicate that these effects grew over time (+0.298 plans in 2012 [$p > .10$]; +0.905 plans in 2013 [$p < .05$]; +1.238 plans in 2014 [$p < .05$]). These estimates are almost identical to those from the entire sample of counties. The results are somewhat attenuated in the analyses using the control group of counties that are "close" to qualifying for double-bonus status, though the overall estimate is still positive and the estimates for the separate postintervention periods are still increasing over time, becoming significant in 2014 (Appendix A, Exhibit 21). Despite the increase in plan offerings, supplemental analysis found that double bonuses were not consistently associated with increased enrollment in affected counties (Appendix A, Exhibits E19–E20).

DISCUSSION

This is the first study to estimate the effect of the size of incentives in pay-for-performance on quality of care. This is also the first evaluation of the effects of pay-for-performance incentives targeted toward insurance plans. While numerous pay-for-performance programs have been initiated by states to improve quality for managed care plans in Medicaid (Kuhmerker and Hartman 2007), we are not aware of published evaluations of these programs.

We find little evidence that larger bonuses in the Medicare Advantage Quality Bonus Payment Demonstration led to greater improvements in quality of care. Specifically, in specifications in which the comparison group consisted of matched counties that had preintervention quality that was similar to the double-bonus counties, the receipt of double bonuses was not associated with improved quality. This is consistent with much of the recent research showing little evidence that financial incentives have improved quality of care (Van Herck et al. 2010; Flodgren et al. 2011; Scott et al. 2011; Houle et al. 2012; Ryan et al. 2014). We are not able to make general inferences about whether the Demonstration improved quality for all plans. While doubling the bonuses in the Demonstration may not have been sufficient to motivate plans to improve quality incrementally, it is possible that standard bonuses led to quality improvements among all plans (compared to a counterfactual of no bonuses). Table 1 suggests a shift in the distribution of

quality ratings away from two- and three-star plans toward four- and five-star plans, though it is difficult to attribute this shift to the Demonstration without a valid control group of health plans that were not exposed to the Demonstration.

We did, however, find evidence that insurers increased plan offerings in counties that were eligible for double bonuses. This increase in plan offerings in double-bonus counties increased each year following the start of the Demonstration, suggesting that plans adjusted gradually to the new incentives. Thus, evidence from this study suggests that double bonuses simply acted as transfer payments to high-quality plans in double-bonus counties, not as a stronger quality incentive as initially intended. We note, however, that if the new plans entering double-bonus counties are part of new Medicare Advantage contracts, they will not receive star ratings for 2–3 years. It is possible that these new plans will eventually receive high quality ratings, increasing plan quality in double-bonus counties.

There are three main reasons why larger bonuses for quality may not have led to high-quality among health plans. First, the incremental increase in revenue for high-quality plans may not have been sufficient to make any additional "marginal" high-quality plans profitable. In addition, if plans face substantial fixed costs for quality improvement, even larger bonuses may not be sufficient to induce higher quality. For instance, if plans attempted to improve quality through changing their physician networks, substantial time and resources may be needed to identify high-performing physicians in the markets where the plan is operating. Investment in quality improvement programs, potentially involving hiring nurse care managers or deploying health information technology, could also involve substantial fixed costs. These fixed costs are likely to vary across the submeasures used to compute the summary measures on which payment are based. However, we find no effect of the double bonuses on any of the domains of measures.

Second, even if plans wanted to improve quality in response to the larger bonuses, they may face short-term constraints in doing so. In both of the settings outlined in the conceptual framework, the response of insurers to the quality-based bonuses may take some time. For example, if insurers respond by initiating new high-quality plans, they may need to construct new provider networks and negotiate new contracts with hospitals. In addition, new plans will not immediately receive a quality rating. If, on the other hand, insurers choose to respond by improving the quality of existing plans, they may need to restructure contracts with their physicians to include quality-based incentives, a process that may be encumbered by the multiyear nature of physician–insurer contracts.

Third, unlike other pay-for-performance programs that compensate providers or provider groups directly for the quality of care they provide, the Demonstration attempted to pass *provider* performance incentives through private Medicare Advantage insurers. This additional complexity of the Demonstration's incentive scheme may have diluted its impact. Plans may also have been limited to initiating broad, across-the-board efforts to improve quality. They may not have been able to target new initiatives specifically toward providers in double-bonus counties. While Medicare Advantage insurers have been engaged in efforts to improve provider quality since the start of HEDIS quality measurement and reporting in the 1990s, this is the first time Medicare Advantage insurers have been financially incentivized to do so.

Our study has a number of limitations. First, we used county-level, rather than plan-level data. While plan data were available, the churn of individual plans makes it challenging to longitudinally assess quality for specific plans. In addition, the use of the county as the unit of analysis corresponds to the policy in question. In the Demonstration, bonus size only varies at the county level, so the question of whether bonus size matters should be asked empirically at the county level as well. Insurers' choice of where to offer plans also takes place at the county level, given that an insurer can choose to offer a plan in one county and not in a similar contiguous county that has a lower benchmark or bonus.

Our study was limited to studying whether a doubling of the bonuses in the Demonstration, which started at between 3 and 5 percent of revenue, affected our study outcomes. Because quality-based bonuses were made available in all counties at the same time, we were not able to test the effects of a 3 percent bonus compared to a counterfactual of no bonuses. Thus, while our study is an evaluation of the effects of the double bonuses in the Demonstration, it is not an evaluation of the Demonstration itself. In addition, because all counties are effectively treated by the Demonstration and double-bonus status is not randomly assigned across counties, we are limited in our ability to disentangle the causal effect of the larger bonus from any treatment effect heterogeneity with respect to the implementation of the Demonstration (i.e., the normal bonuses).

CONCLUSION

"Pay enough, or don't pay at all" was the message from a classic study which found that large financial incentives motivated higher performance on a standardized test, while small incentives led to lower performance

than no incentives at all (Gneezy and Rustichini 2000). Unlike the incentives in other programs (Ryan and Blustein 2011; Ryan, Blustein, and Casalino 2012; Ryan 2013), it would be hard to call the double bonuses in the Medicare Advantage Quality Bonus Payment Demonstration "small." We estimate that the double bonuses increased payments by $3.43 billion over the first 3 years of the Demonstration (see Appendix A for a calculation). It is therefore worrying that the higher incentives in the program were not associated with higher quality. Future research should continue to evaluate how different components of pay-for-performance programs—including the magnitude of payments and the organizational level at which incentives are targeted—affect the outcomes of these programs. Research should also assess whether the long-term benefits of pay-for-performance programs exceed their costs.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Andrew Ryan was supported by a career development award from the Agency for Healthcare Research and Quality (K01 HS18546). Timothy Layton gratefully acknowledges support from the National Institute of Mental Health (T32-019733).

Disclosures: None.

Disclaimers: None.

NOTES

1. Quality ratings are assigned at the contract level, not the plan level. Contracts typically consist of a bundle of plans. We do all analysis at the plan level because we are interested in the effect of double bonuses on the quality of the MA plan options available to Medicare beneficiaries. Dropping plans within a low-quality contract or adding plans to a high quality contract would provide important improvements to the quality of Medicare Advantage plan options. In addition, improving the quality of a contract that includes five plans would be a much larger quality improvement than improving the quality of a contract that includes only one plan.
2. For the main analysis, the outcomes are simple averages across all plans offered in the county. In a supplementary analysis, we use enrollment-weighted averages of the outcomes.
3. We also attempted to do a standard regression discontinuity analysis, but there are too few counties around the 25 percent MA penetration cut off to allow for any clear conclusions.

REFERENCES

Alexander, J. A., D. Maeng, L. P. Casalino, and D. Rittenhouse. 2013. "Use of Care Management Practices in Small- and Medium-Sized Physician Groups: Do Public Reporting of Physician Quality and Financial Incentives Matter?" *Health Services Research* 38 (2): 376–97.

Angrist, J. D., and J. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Centers for Medicare and Medicaid Services. 2012. "Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies and Final Call Letter" [accessed on March 26, 2015]. Available at https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/Announcement2012.pdf

Flodgren, G., M. P. Eccles, S. Shepperd, A. Scott, E. Parmelli, and F. R. Beyer. 2011. "An Overview of Reviews Evaluating the Effectiveness of Financial Incentives in Changing Healthcare Professional Behaviours and Patient Outcomes." *Cochrane Database of Systematic Reviews* 7: CD009255.

Gneezy, U., and A. Rustichini. 2000. "Pay Enough or Don't Pay at All." *The Quarterly Journal of Economics* 115 (3): 791–810.

Houle, S., F. McAlistair, C. Jackevicius, A. Check, and R. Tsuyuki. 2012. "Does Performance-Based Remuneration for Individual Health Care Practitioners Affect Patient Care? A Systematic Review." *Annals of Internal Medicine* 157: 889–99.

Kuhmerker, K., and T. Hartman. 2007. "Pay-for-Performance in State Medicaid Programs a Survey of State Medicaid Directors and Programs." *The Commonwealth Fund* [accessed on March 26, 2015]. Available at http://www.providersedge.com/ehdocs/ehr_articles/Pay-for-Performance_in_State_Medicaid_Programs.pdf

Leuven, E., and B. Sianesi. 2003. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing" [accessed on March 26, 2015]. Available at http://ideas.repec.org/c/boc/bocode/s432001.html

Petersen, L. A., L. D. Woodard, 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.

Rosenthal, M. B., R. Fernandopulle, H. R. Song, and B. E. Landon. 2004. "Paying for Quality: Providers' Incentives for Quality Improvement." *Health Affairs* 23 (2): 127–41.

Rosenthal, M. B., B. E. Landon, S. L. Normand, R. G. Frank, and A. M. Epstein. 2006. "Pay for Performance in Commercial HMOs." *The New England Journal of Medicine* 355 (18): 1895–902.

Ryan, A. M. 2013. "Will Value-Based Purchasing Increase Disparities in Care?" *New England Journal of Medicine* 369: 2472–4.

Ryan, A. M., and J. Blustein. 2011. "The Effect of the MassHealth Hospital Pay-for-Performance Program on Quality." *Health Services Research* 46 (3): 712–28.

Ryan, A. M., J. Blustein, and L. P. Casalino. 2012. "Medicare's Flagship Test Of Pay-For-Performance Did Not Spur More Rapid Quality Improvement among Low-Performing Hospitals." *Health Affairs* 31 (4): 797–805.

Ryan, A. M., J. Burgess, and J. B. Dimick. 2014. "Why We Shouldn't Be Indifferent to Specification in Difference-in-Differences Analysis." *Health Services Research*. In press.

Ryan, A. M., and M. J. Press. 2014. "Value-Based Payment for Physicians in Medicare: Small Step or Giant Leap?" *Annals of Internal Medicine* 160 (8): 565–6.

Ryan, A. M., J. F. Burgess Jr, M. F. Pesko, W. B. Borden, and J. B. Dimick. 2014. "The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program." *Health Services Research*. doi:10.1111/1475-6773.12206.

Scott, A., P. Sivey, D. A. 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.

Sinaiko, A. D., C. C. Afendulis, and R. G. Frank. 2013. "Enrollment in Medicare Advantage Plans in Miami-Dade County: Evidence of Status Quo Bias?" *Inquiry* 50 (3): 202–15.

Town, R., R. Kane, P. Johnson, and M. Butler. 2005. "Economic Incentives and Physicians' Delivery of Preventive Care—A Systematic Review." *American Journal of Preventive Medicine* 28 (2): 234–40.

Van Herck, P., D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus. 2010. "Systematic Review: Effects, Design Choices, and Context of Pay-for-Performance in Health Care." *BMC Health Services Research* 10: 247.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.
Data S1. Supplemental Material.