Appendix A. Supplemental material for "Higher incentive payments in Medicare Advantage's pay-for-performance program did not improve quality, but did increase plan offerings"

Description of star ratings

The performance measures making up the star ratings come from the Consumer Assessment of Healthcare Providers and Systems (CAHPS), the Health Outcomes Survey (HOS), and the Healthcare Effectiveness Data and Information Set (HEDIS). The vast majority of measures are survey-based. Between 33 and 36 measures contribute to the star rating, depending on the year. Each measure is converted to a measure-specific star rating by establishing (either relative or absolute) thresholds of the quality measure for each star. A contract (a bundle of plans) is then assigned from one to five stars for each measure. These measure-level ratings are then averaged and rounded to the nearest half star to determine the summary rating for the contract. All plans within the contract are assigned the same summary rating.

Because the quality ratings are based on plan performance data from prior years, CMS does not provide quality ratings for new plans. Additionally, small plans and Private-Fee-for-Service plans are unlikely to have a quality rating because they were not required to participate in the patient surveys that are an input to the star ratings (Jacobson et al. 2011).

Description of payments to Medicare Advantage plans

Prior to 2012, MA plan payments were based on a benchmark that varied by county. Benchmarks were set somewhat arbitrarily to approximate average costs in fee-for-service Medicare in each county, with some additional payments to urban and rural "floor" counties to encourage plan entry. To determine payment rates, plans would first submit bids to CMS. If the bid was less than the county benchmark, the plan's premium was set equal to $0 and the plan received a "rebate" equal to a

portion of the difference between the benchmark and the bid. The plan was required to spend the entire rebate on additional benefits for enrollees. If the bid was larger than the county benchmark, the plan's premium was set equal to the difference between the bid and the benchmark. This is shown in Exhibit E1.

The bonuses paid to MA plans via the MA QBP Demonstration are made up of two components, the "benchmark bonus" and the "rebate bonus". First, the county-level benchmark used to determine plan payments is a function of the star-rating. Call this the "benchmark bonus." Let B_c be the baseline benchmark for county c. B_c will be the applicable benchmark for any plan with fewer than 3 stars. For any plan with 3, 4, or 5 stars, the applicable benchmark will be increased by 3%, 4%, or 5%, respectively. Formally, the benchmarks for these plans will be $B_c + 0.03B_c$, $B_c + 0.04B_c$, or $B_c + 0.05B_c$. For example, in 2011, before the MA QBP demonstration began, the benchmark for Autauga County, AL was $814.36 for all plans. In 2012, after implementation of the MA QBP demonstration, the benchmark was $786.42 for a plan with fewer than 3 stars, $810.30 for a 3-star plan, $818.25 for a 4-star plan, and $826.21 for a 5-star plan. The benchmark bonus results in five-star plans that bid above the benchmark having lower premiums because the difference between their bid and the premium will be smaller due to the higher five-star plan benchmark. Plans bidding below the benchmark receive larger rebates, allowing them to provide additional benefits. In both cases, the MA QBP demonstration should result in increased enrollment and profits in higher quality plans, either due to lower premiums or additional benefits. This provides plans with additional incentives to be high quality.[1]

The rebate paid to MA plans under the MA QBP demonstration is also a function of the star-rating. Call this the "rebate bonus." Prior to the implementation of the MA QBP demonstration, a plan's rebate was equal to 75% of the difference between the benchmark and the plan's bid. The quality-based

[1] As discussed in the previous section, new, small, and PFFS plans are likely to be unrated due to missing data. For purposes of the MA QBP demonstration, these plans are treated as 3-star plans.

rebates are phased in between 2012 and 2014. In 2014, the rebate is equal to 70% of this difference for plans with 4.5 stars or more, 65% for plans with 3.5 to 4.5 stars, and 50% for plans with fewer than 3.5 stars. These rebate bonuses grow over time. Thus, under the MA QBP, high quality plans not only have higher benchmarks to bid against, but, if they bid below those benchmarks, they receive a larger share of the difference between their bid and the benchmark in the form of a rebate that they are then required to pass on to their enrollees in the form of additional benefits. These additional benefits will again result in higher enrollment and profits among high quality plans, providing plans with an amplified incentive to boost quality.

The bonuses can be seen in Figures 2 and 3. Figure 2 shows the case where a plan bids above the benchmark. In the figure there are two plans, a 2-star plan and a 5-star plan. The plans have identical bids. The benchmark for a 5-star plan is 5% larger than the benchmark for a 2-star plan. The figure shows that this results in a smaller premium for the 5-star plan. The difference between the benchmarks, and thus the difference between the premiums, is the "bonus" paid to the 5-star plan. In this case, the "rebate bonus" is irrelevant because there is no rebate. Figure 3 shows the more complicated case where a plan bids below the benchmark. Again, the figure shows a 2-star plan and a 5-star plan. Again, the 5-star benchmark is 5% larger than the 2-star benchmark. In this case, however, the plan is paid the bonus via the rebate, which is a portion of the difference between the benchmark and the plan's bid. The larger 5-star benchmark results in a larger rebate for the 5-star plan. This is the "benchmark bonus." Additionally, while the 2-star plan only receives a rebate equal to 50% of the difference between its bid and the benchmark, the 5-star plan receives a rebate equal to 70% of this difference. This is the "rebate bonus." The two bonuses are combined to form the total bonus, shown in Orange.

Our Construction of double-bonus counties

Our dataset provides the benchmark for a plan with each quality rating. We calculate each county's 5-star benchmark bonus by dividing the difference between the 5-star bonus and the 2-star bonus by the 2-star bonus. The 5-star benchmark bonus for non-double bonus counties should be 5% and the bonus for double bonus counties should be 10%. Due to quirks in CMS' formula for the benchmarks, the estimated benchmark bonuses are usually close, but not identical to these values. Additionally, as discussed above, some counties moved in or out of double bonus status during our sample period. To deal with these irregularities we calculated the minimum and maximum 5-star benchmark bonuses over the period of the Demonstration. We then 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 5-star benchmark bonus exceeded 9% and as a non-double-bonus county if its maximum 5-star benchmark bonus was less than 6%. 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% 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.

Details of Matching Procedure

We implemented our matching procedure 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 three years prior to the start of the Demonstration, along with lagged levels of the county benchmark were the only variables used for matching. The matching procedure was implemented in Stata using a user-written command (Leuven and Sianesi 2003). In all analysis, standard errors are clustered at the county-level.

Supplemental Analysis

We performed a series of supplemental analyses to extend the main results reported in the body of the paper. First, we explored whether Medicare Advantage plans were able to expand high quality plans immediately following the passage of the ACA. To do this, we changed the pre-intervention period to include star-rating years 2011-2014 as the pre-intervention period, rather than star rating years 2012-2014 in the main analysis. The results show no evidence that plans responded to the incentives to improve quality in this earlier period (Exhibits E5 and E6). However, we found some evidence that double-bonuses led to plan expansions as early as 2011.

Next we evaluated whether the Demonstration had heterogeneous effects across the different incentivized performance domains. We specified different models for each of four separate performance domains related to staying healthy (i.e., screenings, tests and vaccines), managing chronic disease, ratings of health plan responsiveness and care, and health plan members' complaints, appeals, and choosing to leave the health plan. While part of the star-ratings, we do not assess the telephone customer service domain because it is not included because it is not used in all years. In the early years, there is a "getting timely care from providers" domain instead of the telephone customer service domain.

Exhibits E9, E11, E13, and E15 show substantial variation in plan performance among these domains over time. For instance, plans performance generally improved performance for the managing chronic disease, plan responsiveness, and staying healthy domains, but got worse for the managing appeals and complaints domain. Nonetheless, results from our difference-in-differences analysis find little evidence that the timing of the double-bonuses was associated with incremental improvement for any of these domains, particularly in the matched sample (Exhibits E10, E12, E14, and E16).

We then assessed whether the expansion of high quality plans in double-bonus counties led to a greater share of enrollment in higher-quality plans. For this analysis, our dependent variable was the county-level average star rating weighted by share of enrollment in plans of varying quality. We then re-performed the difference-in-differences analysis with this dependent variable. This analysis showed that statistical significance of the effect of double bonuses depended on the specification (Exhibits E17 and E18). In the entire sample, the effect of double bonuses was significant (+0.043, $p < .05$). This was driven by a large effect in 2013 (+0.097, $p < .01$). In the matched sample, a similar pattern of results are observed, but the effects were not statistically significant. Given our result that there was no effect of the double bonuses on unweighted quality, this result provides some, albeit inconclusive, evidence of a shift in enrollment toward higher quality plans due to the larger bonuses those plans received. This is consistent with the idea that those higher quality plans received larger benchmark payments after the start of the Demonstration and competition drove those plans to pass a portion of the additional payments through to consumers in the form of lower premiums or additional benefits (Cabral et al. 2014).

Finally, we evaluated whether double-bonuses were associated with Medicare Advantage enrollment. We performed the main analysis, using total Medicare Advantage enrollment as the dependent variable. We found that enrollment appeared to increase slightly in double-bonus counties, but the effect was not significant (Exhibits E19 and E20). Here, results vary across the full sample and the matched sample. In the full sample, our analysis suggests that double-bonuses increased enrollment, and that the rate of increase increased in the post-intervention period. However, in the matched sample, our estimates indicate that double-bonuses decreased enrollment, although the results were not significant.

Calculation of the size of double bonuses

There were 588 county-year observations from double-bonus counties in our analytic file. To determine the size of the double bonuses we first defined a county as a double-bonus county in a given year if the county's 5-star benchmark bonus was larger than 7.5% during that year (unlike the analysis described in the text, for this back-of-the-envelope calculation we allow counties to be classified as double-bonus counties in some years and not others). We then calculate simulated bonuses for each county based on the 2.5-star benchmark. Following the rules of the Demonstration, for 2012 and 2013, 3-star simulated bonuses were equal to the 2.5-star benchmark multiplied by 0.03, 3.5-star simulated bonuses were equal to the 2.5-star benchmark multiplied by 0.035, 4-star and 4.5-star simulated bonuses were equal to the 2.5-star benchmark multiplied by 0.04, and 5-star simulated bonuses were equal to the 2.5-star benchmark multiplied by 0.05. 2014 bonuses are defined in the same way with the exception of 4-star and 4.5-star benchmarks which are equal to the 2.5 star benchmark multiplied times 0.05, as stipulated by the rules of the Demonstration. We then also simulate the double bonuses for each county by multiplying the simulated normal bonuses by 2. We then determine for each double bonus county the total double bonus payment by multiplying the simulated double bonus for each star rating by the number of individuals enrolled in a plan of that rating in that county. Finally, we sum these total double bonus payments across all double bonus counties to get an estimate of a cost double bonuses of $3.4B over the three years of the demonstration.

Exhibit E1. Description of star ratings and Quality Bonus Program

Star rating and bonus payment rating year	Data collection period	Measures contributing to star rating
2009	January 2007 - June 2008	36
2010	January 2008 - June 2009	33
2011	January 2009 - June 2010	36
2012	January 2010 - June 2011	36
2013	January 2011 - June 2012	37
2014	January 2012 - June 2013	37

Exhibit E2. Illustration of Medicare Advantage payment prior to the Quality Bonus Payment Demonstration

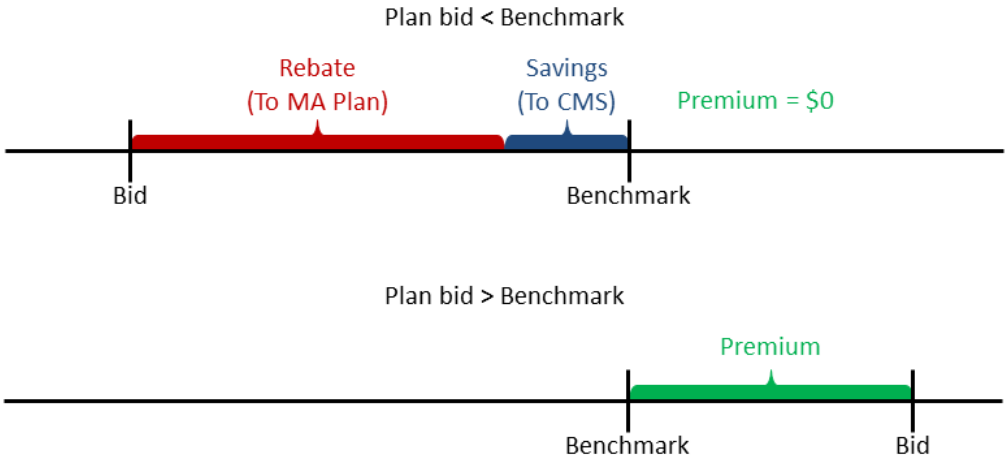

Exhibit E3. Illustration of Medicare Advantage payment in the Quality Bonus Payment Demonstration for bids above the benchmark

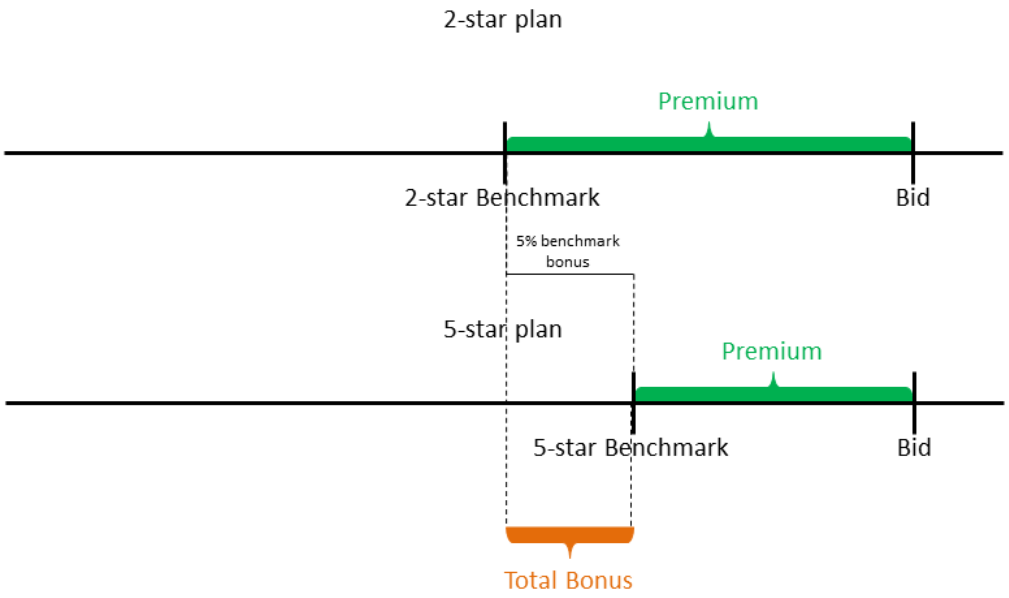

Exhibit E4. Illustration of Medicare Advantage payment in the Quality Bonus Payment Demonstration for bids below the benchmark

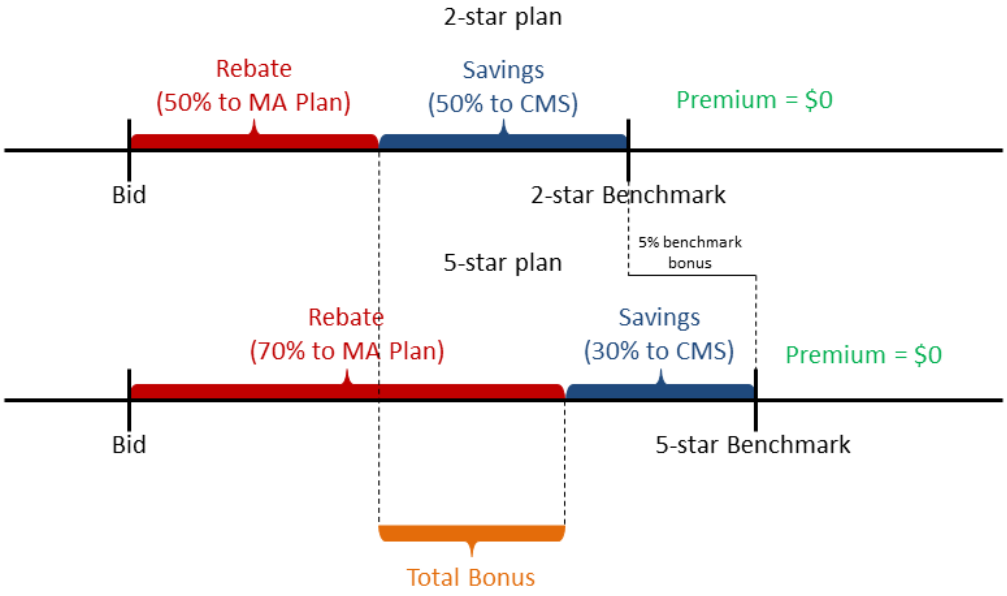

Exhibit E5. Sensitivity analysis of quality of care among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration, assuming effects started in 2011

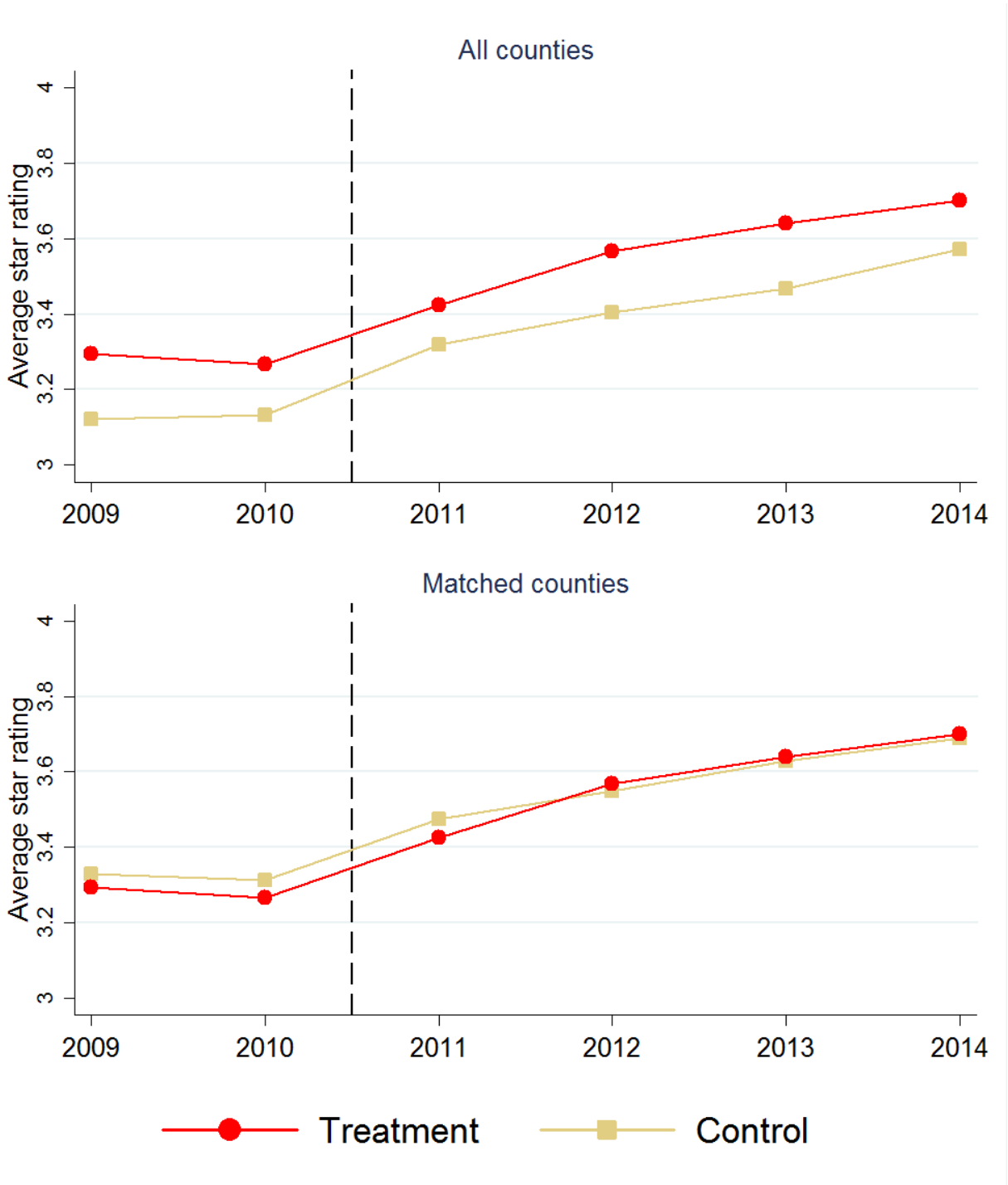

Exhibit E6. Sensitivity estimates of the effects of the Medicare Demonstration Quality Payment Demonstration from difference-in-differences models, assuming effects started in 2011

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	-0.010		0.038	
	(0.022)		(0.026)	
treat2011		-0.047*		-0.010
		(0.022)		(0.031)
treat2012		0.010		0.059*
		(0.024)		(0.028)
treat2013		0.020		0.051
		(0.029)		(0.035)
treat2014		-0.024		0.052+
		(0.028)		(0.031)
N	7932	7932	5604	5604

Standard errors in parentheses

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E7. Sensitivity analysis of number of plans offered among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration, assuming effects started in 2011

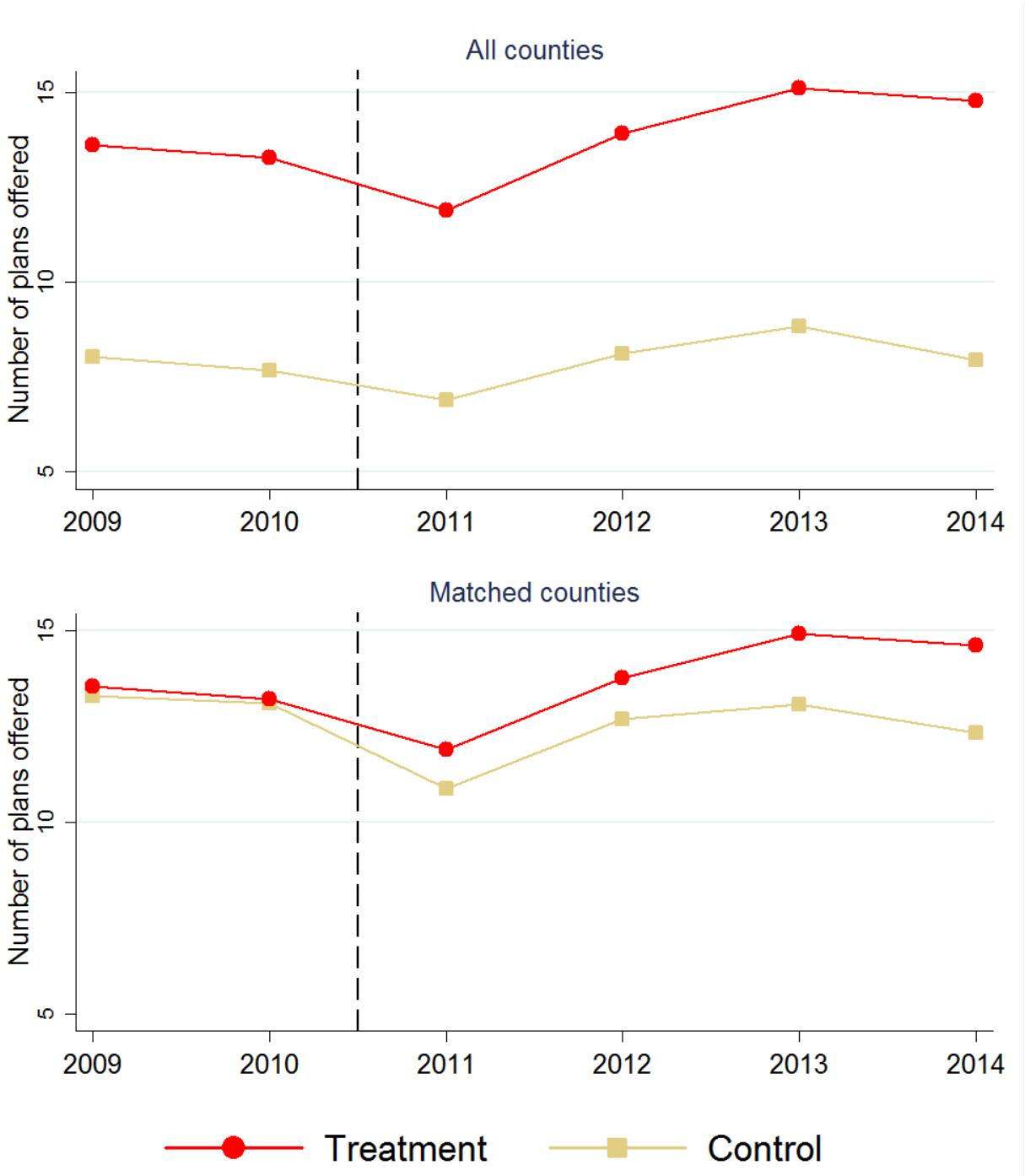

Exhibit E8. Sensitivity estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on the number of plans offered from difference-in-differences models, assuming effects started in 2011

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	0.254		1.329**	
	(0.263)		(0.385)	
treat2011		-0.630*		0.836*
		(0.257)		(0.386)
treat2012		0.125		0.875*
		(0.263)		(0.370)
treat2013		0.544+		1.617**
		(0.326)		(0.450)
treat2014		0.975**		1.986**
		(0.358)		(0.527)
N	8880	8880	5226	5226

Standard errors in parentheses
+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E9. Quality of care for the managing chronic disease domain among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

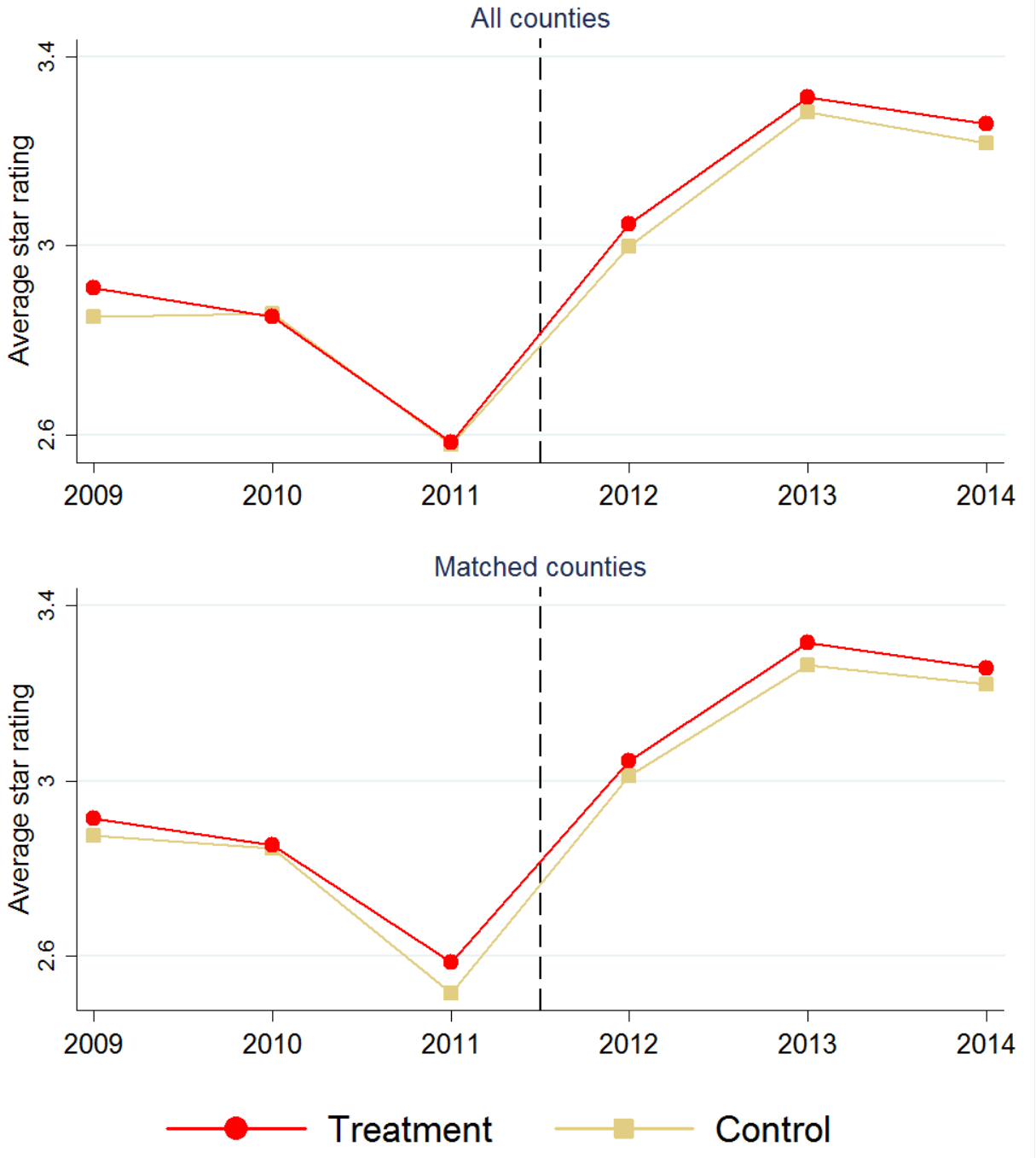

Exhibit E10. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on the managing chronic disease domain from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	0.029		0.013	
	(0.023)		(0.034)	
treat2012		0.031		0.000
		(0.023)		(0.031)
treat2013		0.021		0.022
		(0.037)		(0.051)
treat2014		0.035		0.017
		(0.025)		(0.037)
N	5172	5172	3300	3300

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E11. Quality of care for the ratings of health plan responsiveness and care domain among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

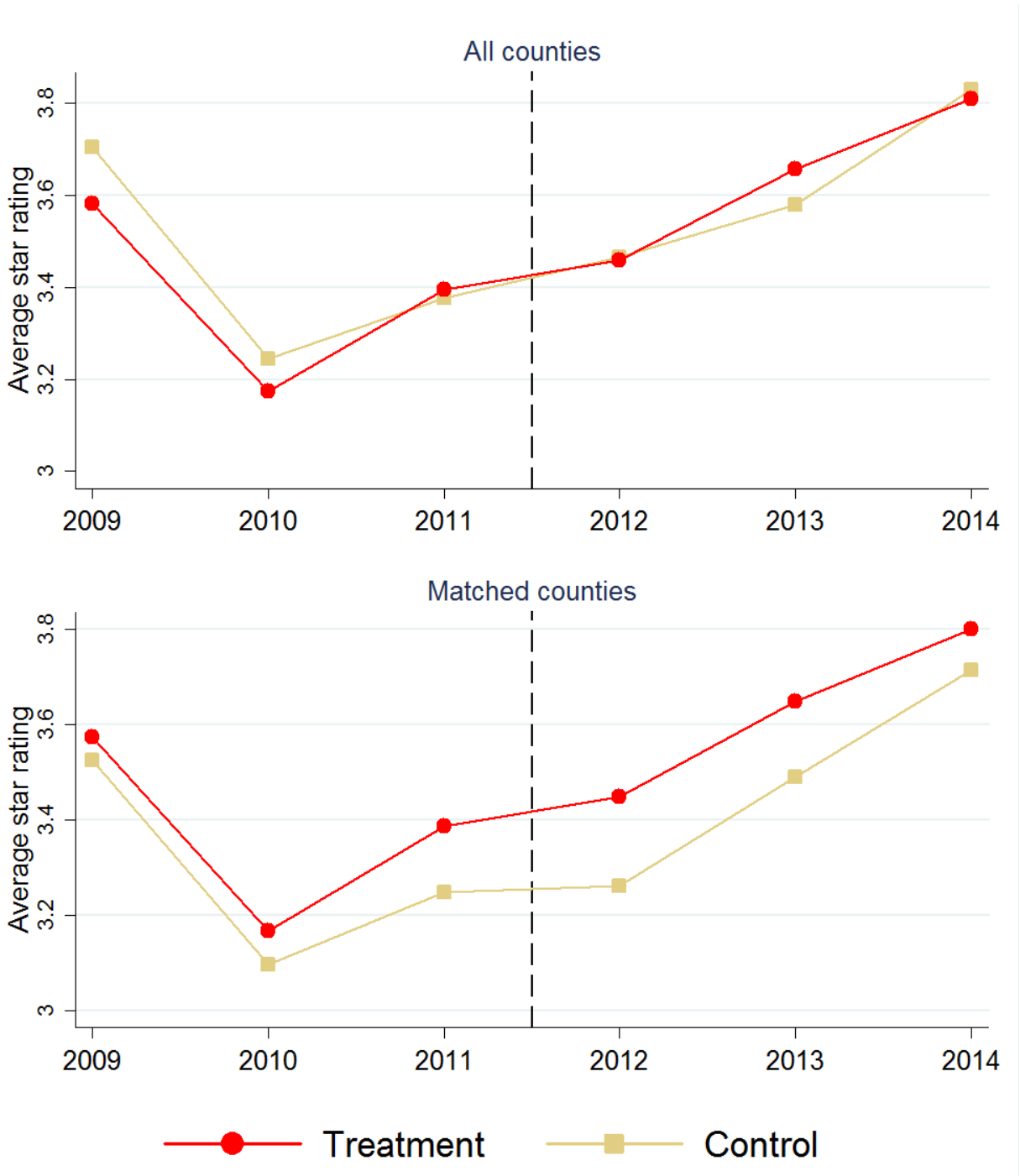

Exhibit E12. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on the ratings of health plan responsiveness and care domain from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	0.113**		0.086	
	(0.030)		(0.061)	
treat2012		0.070+		0.112+
		(0.037)		(0.067)
treat2013		0.170**		0.098
		(0.042)		(0.076)
treat2014		0.100**		0.047
		(0.031)		(0.064)
N	5172	5172	3132	3132

Standard errors in parentheses

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E13. Quality of care for the health plan members' complaints domain among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

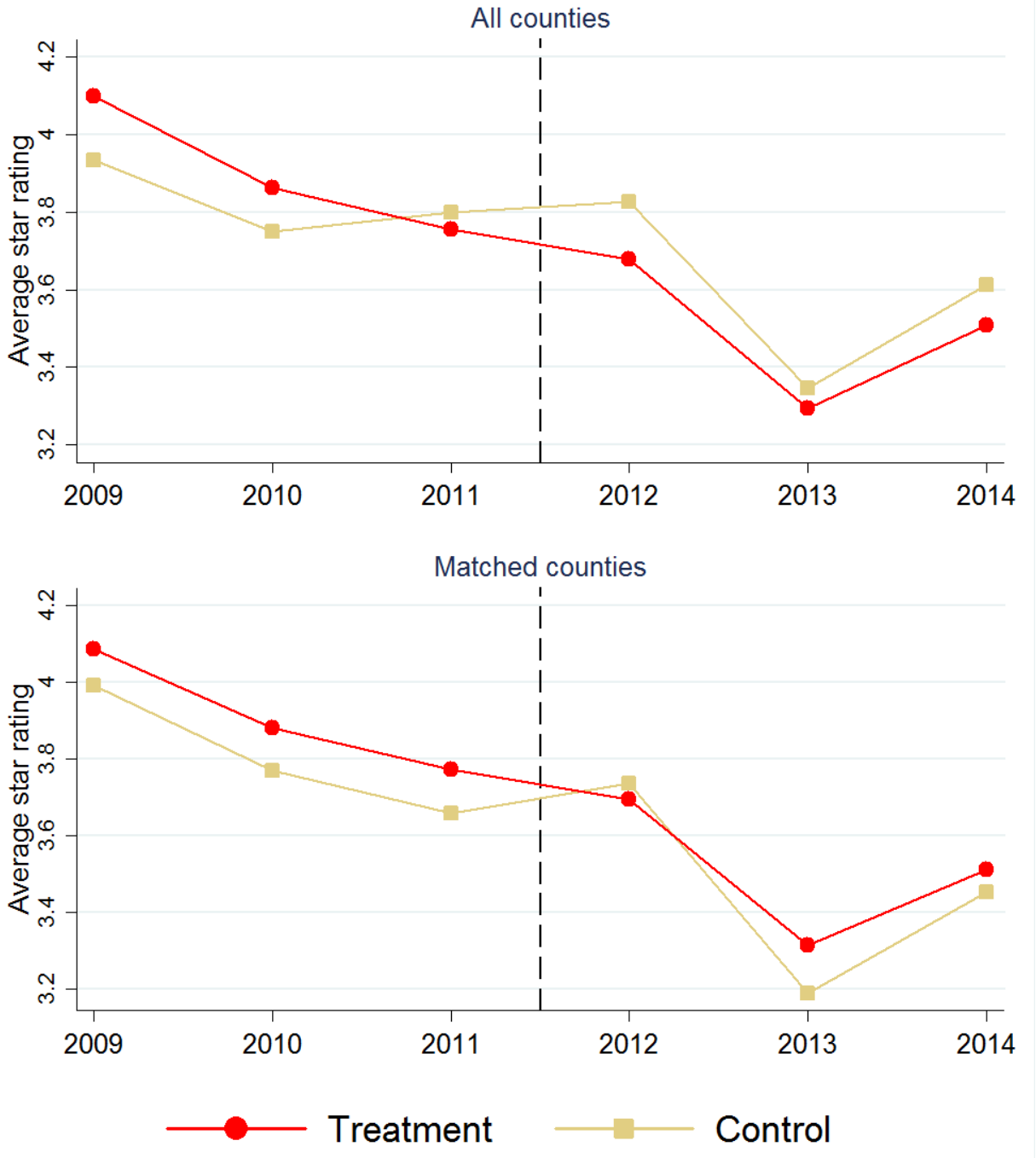

Exhibit E14. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on the health plan members' complaints domain from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	-0.160**		-0.045	
	(0.040)		(0.061)	
treat2012		-0.218**		-0.143+
		(0.047)		(0.073)
treat2013		-0.112*		0.031
		(0.050)		(0.075)
treat2014		-0.150**		-0.023
		(0.051)		(0.074)
N	5172	5172	2916	2916

Standard errors in parentheses

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E15. Quality of care for the staying healthy domain among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

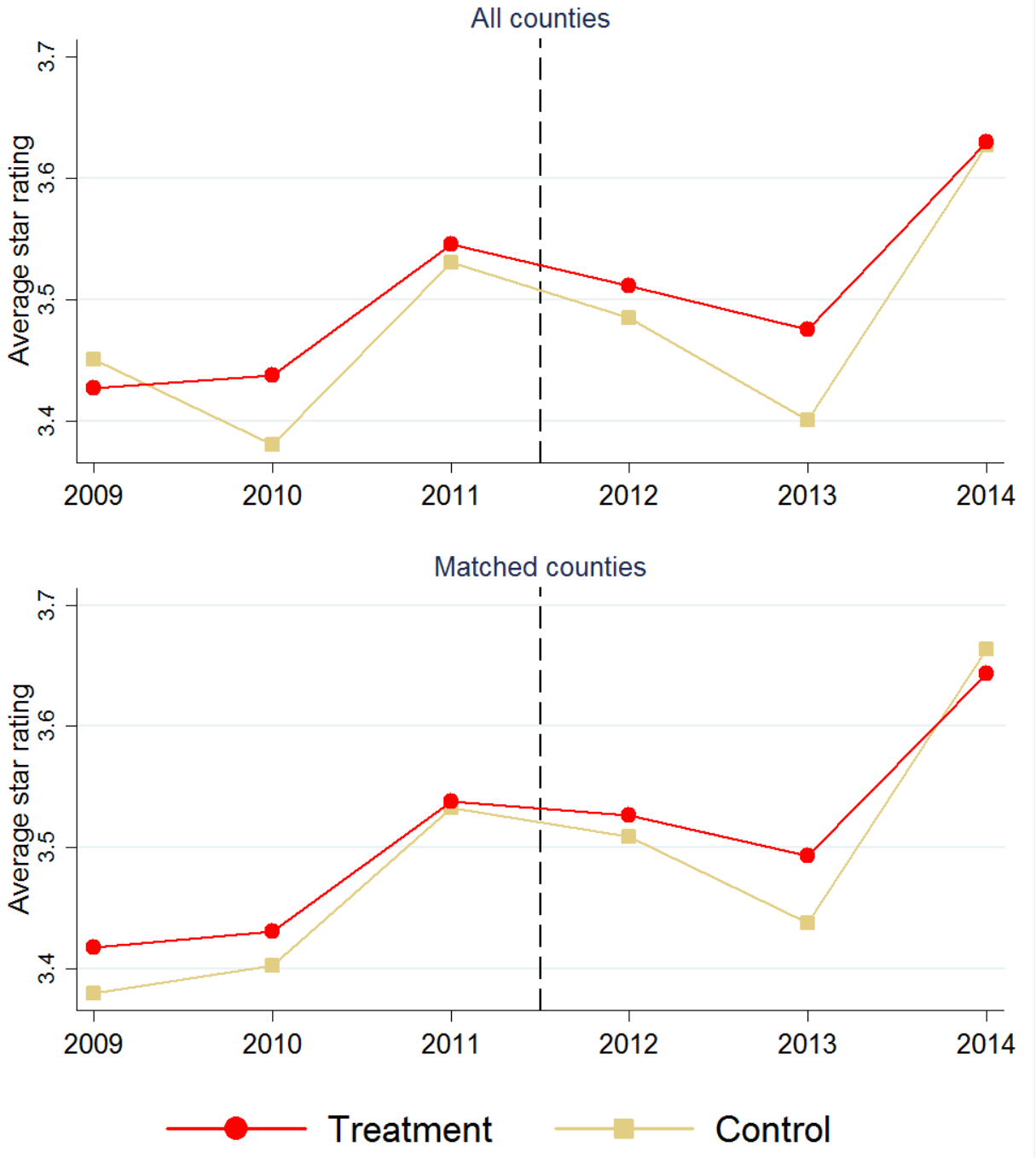

Exhibit E16. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on the staying healthy domain from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	0.041		0.007	
	(0.027)		(0.040)	
treat2012		0.022		0.000
		(0.027)		(0.044)
treat2013		0.079*		0.044
		(0.033)		(0.048)
treat2014		0.023		-0.023
		(0.036)		(0.052)
N	5172	5172	3366	3366

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E17. Enrollment-weighted quality of care among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

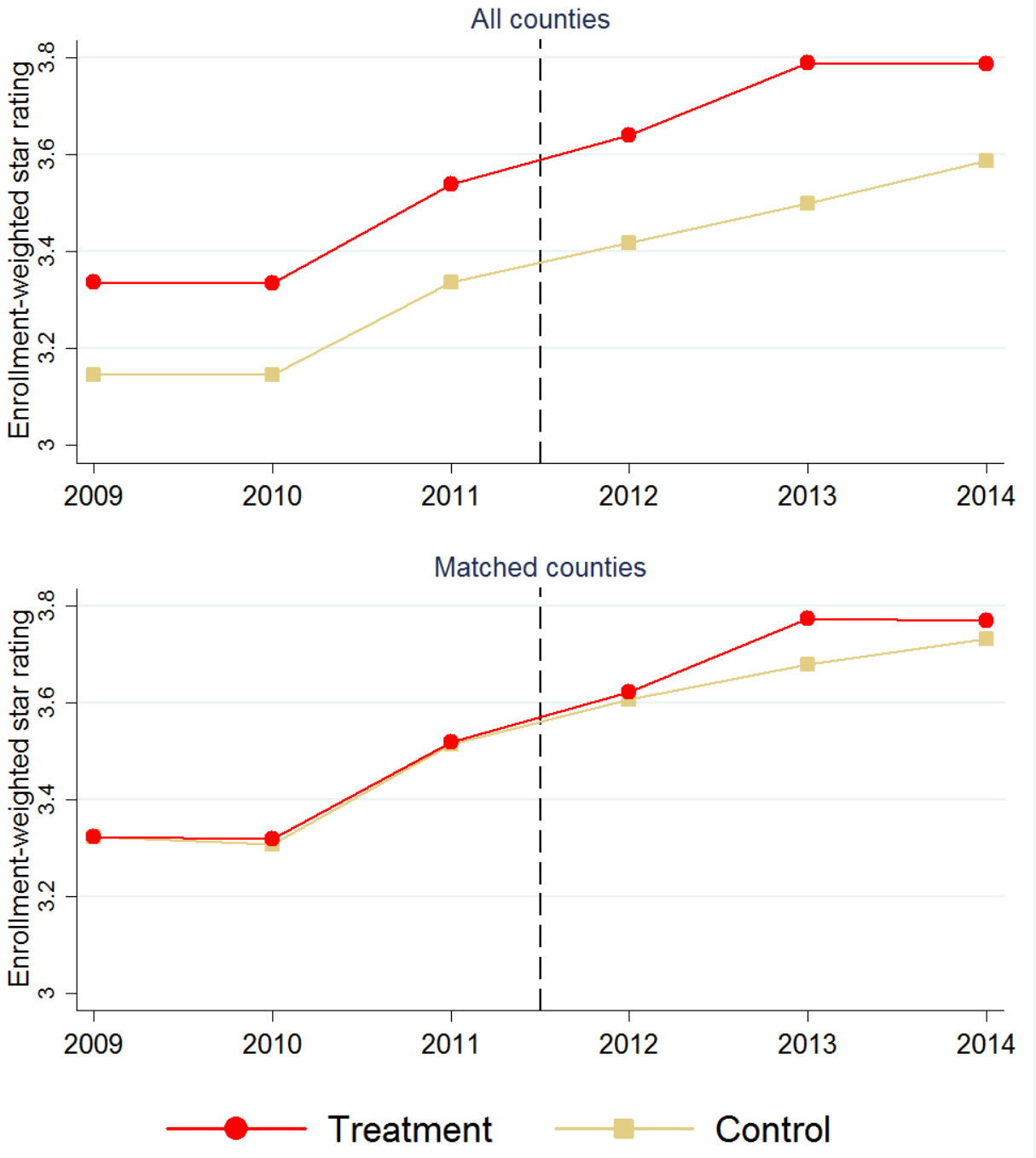

Exhibit E18. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on enrollment-weighted quality from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	0.043*		0.049	
	(0.018)		(0.048)	
treat2012		0.027+		0.014
		(0.016)		(0.053)
treat2013		0.097**		0.095+
		(0.025)		(0.056)
treat2014		0.005		0.039
		(0.024)		(0.044)
N	7482	7482	3876	3876

Standard errors in parentheses
+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E19. Enrollment in Medicare Advantage plans among counties receiving and not receiving double-bonuses in the Medicare Demonstration Quality Payment Demonstration

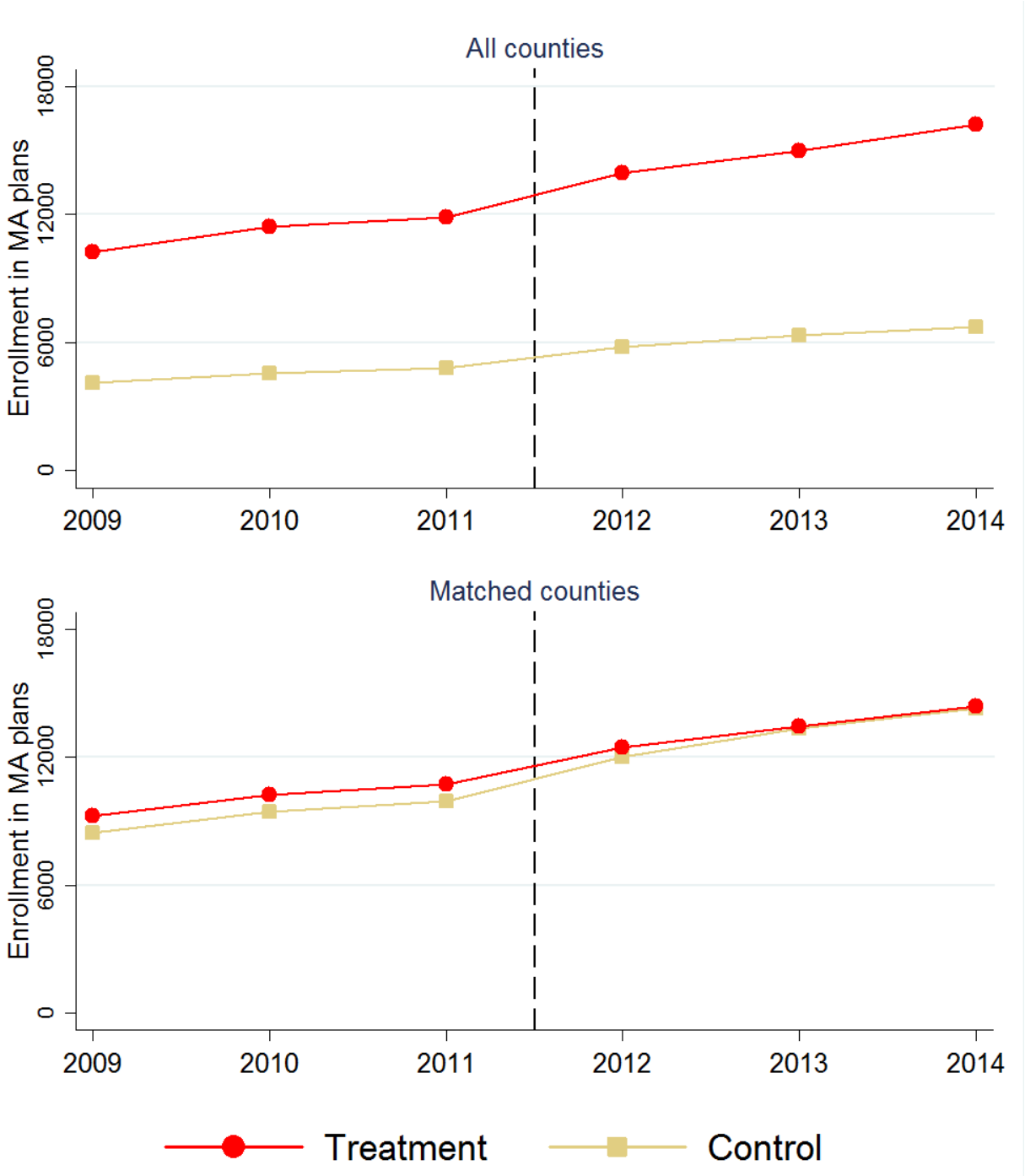

Exhibit E20. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on total enrollment in Medicare Advantage plans from difference-in-differences models

	(1) All counties	(2) All counties	(3) Matched counties	(4) Matched counties
treat_post	2063.157**		225.698	
	(363.628)		(2829.711)	
treat2012		1453.835**		454.674
		(268.155)		(2656.105)
treat2013		1952.111**		107.462
		(361.924)		(2856.399)
treat2014		2783.526**		114.957
		(475.875)		(2992.257)
N	7482	7482	1950	1950

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Exhibit E21. Estimates of the effects of the Medicare Demonstration Quality Payment Demonstration on plan quality and number of plans from difference-in-differences models using only counties "close" to qualifying for double bonuses

	(1) All plans, Average star rating	(2) All plans, Average star rating	(3) All plans, Number of plans offered	(4) All plans, Number of plans offered
treat_post	0.015		0.302	
	(0.025)		(0.295)	
treat2012		0.013		-0.317
		(0.025)		(0.260)
treat2013		0.023		0.386
		(0.031)		(0.351)
treat2014		0.008		0.836*
		(0.031)		(0.382)
N observations	3060	3060	3186	3186

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Note: Specifications estimated for all treatment counties and the set of control counties with urban floor status and 2009 MA penetration between 15% and 25% or non-urban floor status and MA penetration between 25% and 35%

Exhibit E22. Sensitivity analysis for HMO penetration

	Average star rating				Number of plans			
	(1) Quartile 1 penetration, Average star rating	(2) Quartile 2 penetration, Average star rating	(3) Quartile 3 penetration, Average star rating	(4) Quartile 4 penetration, Average star rating	(5) Quartile 1 penetration, # plans	(6) Quartile 2 penetration, # plans	(7) Quartile 3 penetration, # plans	(8) Quartile 4 penetration, # plans
treat_post	0.067	0.017	0.083	-0.095	1.062	2.229	0.030	0.455
	(0.134)	(0.050)	(0.124)	(0.089)	(1.091)	(1.439)	(1.111)	(0.614)
N observations	1716	2022	1812	2382	1986	2454	1998	2442

Standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$