Quality, Quality Payments, and Risk Selection in Private Medicare

Michele Fioretti[*] **Hongming Wang**[†]

August 21, 2018

Abstract

As Medicare payment increasingly rewards value of care, empirical evidence on the effect of value-based payment is still scarce. This paper examines the incidence of quality bonus payments in the Medicare Advantage market. We find that high quality contracts increased bids by the full amount of the bonus payment, leaving enrollee premium and rebate unchanged. Within contract, premium increased (decreased) in high (low) risk counties, and the risk pool improved significantly over low quality contracts. Evidence suggests that the selection arises from a negative correlation between risk scores and patient outcome measures in the quality rating. The selection response calls into question the distributional implication of quality payments, and the risk confounds in measures of quality linked to payment.

JEL classifications: I13, I18, L15
Keywords: Medicare, quality ratings, quality-based payments, risk selection

[*]Department of Economics, University of Southern California, 3620 South Vermont Ave, Los Angeles, CA 90089. e-mail: fioretti@usc.edu
[†]Department of Economics, University of Southern California, 3620 South Vermont Ave, Los Angeles, CA 90089. e-mail: hongminw@usc.edu

1 Introduction

Medicare is the primary source of health insurance for the elderly population in the United States, enrolling more than 45 million beneficiaries in 2015. Spending on Medicare reached $500 billion in 2010, or 20% of total health spending, and continues to grow at a rate of 5% each year (CMS, 2016). To combat cost and promote quality, the 2010 Affordable Care Act (ACA) introduced payment models that reward the value of care in addition to volume.[1]. Evaluating and improving the effectiveness of quality-linked payment is of central policy interest to the Medicare program.

At the same time, Medicare enrollees increasingly receive service from private insurers on the Medicare Advantage (MA) program.[2] Effective cost control in the private Medicare is complicated by potential supply-side capture: every dollar increase in enrollee benefit costs the government more than a dollar in insurer payment. Moreover, insurer selection on market or enrollee characteristics carries unintended social consequences that further complicate the contract design. Because a similar contracting structure is adopted in the ACA Exchange and increasingly among state Medicaid programs, understanding the incentive and incidence of value-based payment in the MA context has implications for a wide range of insurance markets.

In this paper, we provide the first evidence on the effectiveness of value-based payment in private Medicare exploiting the 2010 legislative changes introduced in the ACA and the Quality Bonus Payment (QBP) demonstration. Prior to ACA, MA payment follows a competitive bidding process established in the 2003 Medicare Modernization Act (MMA): plans submit cost estimates, or bids, and rebate a fraction of the cost saving below a benchmark to enrollees as premium reduction or extra benefits. ACA lowered the benchmark to roughly equal the fee-for-service spending.[3] Moreover, moving from volume to value, ACA awarded bonus benchmark and rebate percentage to high-quality insurers according to a star rating system introduced in 2009. The ACA payment formula is set to be effective in 2015; for a three-year period between 2012 and 2014, the Quality Bonus Payment (QBP) demonstration experimented with more generous quality payments than ones stipulated in the ACA.

[1] The goal was to link all payments to private insurers with service quality by 2012, as well as linking 85% of the payments to traditional Medicare (TM) providers by 2016, a figure that was set to increase to 90% by 2018 (Burwell, 2015)

[2] MA market share increased rapidly from 23% of Medicare beneficiaries in 2009 to 31% in 2015. Payment to insurers doubled from 15% of Medicare spending in 2006 to 30% in 2016 (Congressional Budget Office, 2017, Kaiser Family Foundation, 2017).

[3] Before the legislation, the Medicare Payment Advisory Commission (MedPAC) concluded MA payment is 14% higher than FFS cost. Aligning MA benchmark with FFS cost is expected to reduce program spending.

We examine the effect of differential payment to quality over the QBP period, relative to a pre-QBP baseline where payment is indiscriminate of quality. Our difference-in-difference strategy tracks baseline high and low quality contracts. Both quality exhibits similar trend before the reform in terms of pricing, enrollment and cost. After QBP, we find that high quality contracts increase bids by nearly the full size of the bonus payments, and as a result, do not pass on more rebates to enrollees relative to low quality contracts. Correspondingly, the average enrollee premium did not decrease more for high quality contracts, suggesting bonus payments are by and large retained by high-achieving insurers as revenues.

The absence of an effect on average pricing, however, masks substantial pricing variation facing enrollees across service areas. Looking *within* the contract market set, we uncover significant pricing differentials — higher incidence of zero-premium pricing and enrollment in counties with lower fee-for-service (FFS) risk score — for baseline high quality contracts, but not for baseline low quality contracts. The within-contract premium variation explains the risk improvement of high quality contracts despite a null effect on average premium. Consistent with the risk selection mechanism, we find baseline high quality contracts with lower FFS risk score across service areas experience even greater reduction in enrollee risk after QBP, and so do contracts with higher baseline market share in the low risk counties.

We do not find strong evidence for other mechanism that may lead to improved risk profile of high quality contracts. For example, we do not find significant changes in market coverage: high quality contracts did not differentially exit counties with higher FFS risk score, or enter counties with higher baseline or ACA-revised benchmark; nor did they vary the number of plans offered, or the number of counties served. Within contract, we also do not find significant differential in prescription drug deductible across service areas, although we cannot rule out differentials in other aspects of benefit design which we do not observe. Overall, evidence points to the importance of premium variation across markets in the advantageous selection of high quality MA contracts.

To understand *why* quality-linked payments incentivize risk selection, and ultimately *how* the selection can be undone, in the final part of the paper, we investigate the relationship between enrollee risk and contract quality. Conceptually, if lower enrollee risk contributes to higher quality, then the bonus payment can lead insurers to favor enrollments from low-risk regions, with stronger incentive on high quality contracts eligible for larger bonus. This is consistent with the finding that only baseline high quality contracts engage in risk selection after QBP, and almost none of the bonus payment is passed through to enrollees.

We exploit the same difference-in-difference variation to characterize the correlation between risk score and quality, and the associated selection incentive. We show that contracts with higher risk score in the baseline are less likely to perform well in patient outcome measures in the quality rating. Furthermore, high quality contracts with high baseline risk scores experience smaller improvement in outcome ratings relative to low quality contracts with a more favorable risk pool, consistent with the selection incentive for continued quality payments. We further document a negative correlation between contemporaneous risk score and outcome rating as well as the final rating, particularly for baseline high quality contracts. For these contracts, performance in outcome-related domains are the most predictive of the final rating.

The selection suggests an outcome-based quality rating is biased due to baseline differences in health conditions: contracts enrolling patients with more complicated diagnoses perform less well in outcome measures, and are disadvantaged in the quality rating. A better measurement of quality would compare outcomes only across patients with a similar case-mix of diagnoses, or adjust outcome by baseline risk. The adjustment is lacking in the current rating for outcome measures in the "managing chronic conditions" domain. Removing the risk bias in quality rating should lessen the selection incentive associated with quality payments.

The paper contributes to a growing literature on the economic incidence of government payments to health insurers and providers (see for example Dafny, 2005, Clemens and Gottlieb, 2014 and Carey, 2018). In the Medicare Advantage context, a near zero pass-through to enrollees is striking, but not unprecedented. Cabral *et al.* (2018) exploits the payment floor variation in the 2000 Benefits Improvement and Protection Act (BIPA), and finds a pass-through rate of around 50%, with more than 80% of the pass-through in the form of lower premium. Since then, the MA market underwent sweeping changes in terms of risk adjustment, managed competition and the prescription drug program. Looking at year 2007-2011, Duggan *et al.* (2016) finds a near zero pass-through across neighboring urban and rural counties. We continue this line of research by showing the recent quality-linked payments similarly have minimal pass-through to enrollees: bid increased by almost the full amount of bonus payment, leaving rebate to consumers largely unchanged.

The within-contract premium variation we uncover is related to the literature on selection in Medicare Advantage (Brown *et al.*, 2014; Newhouse *et al.*, 2012). A common strategy adopted by researchers is comparing the cost of enrollees who switched from traditional Medicare to MA. Based on the similar intuition, we find high quality contracts differentially lowered premium in counties where potential enrollees likely have lower

cost as indicated by the FFS risk score. The within-contract variation complements existing evidence on cross-contract variation in pricing and benefit design as potential mechanism of advantageous selection in high quality MA plans (Decarolis *et al.*, 2017; Decarolis and Guglielmo, 2017).[4]

The selection result also adds to a nascent literature that recognizes the limitation of standard risk adjustment on all aspects of "cream-skimming" contract design (Einav *et al.*, 2016). For example, drugs receiving low payment relative to cost are more likely placed on higher cost-sharing tier (Geruso *et al.*, 2016), and so are drugs treating diagnosis rendered unprofitable by technology change (Carey, 2017). In this paper, we highlight another aspect of insurance contract, i.e., quality rating, as a potential source of risk selection.In this case, correcting for the risk confound in quality rating can suppress the selection incentive without recourse to ex post risk adjustment.

Taken as a whole, our evidence from private Medicare emphasizes the role of insurer selection on the distributional incidence of quality payments, with mixed implications for welfare. Since high quality service is made less (more) costly to enrollees in low (high) risk counties, the benefit of bonus payment is not evenly felt across beneficiaries[5]. Still, total enrollee surplus may increase if quality improves in both high and low risk regions. The overall welfare effect, however, depends on the economic incidence between enrollee benefits and cost to the government, intermediated by insurer selection and pass-through. Curto *et al.* (2014) estimates that during 2006-2011 two thirds of the payment surplus is in the form of insurer profit and one third goes to enrollees who suffer some disutility from managed care. The benchmark and rebate variation in QBP further complicates the split between insurers and enrollees, and among enrollees, those with high and low potential gain from quality. Hence we view evidence in this paper as constructive inputs to a normative characterization of value-based payment in health insurance markets, which we leave for future work.

[4]Risk selection incentive changed for a handful of 5-star contracts in 2012 due to a special enrollment provision that allows enrollees to switch to a 5-star plan at any time in a year. Despite concerns of adverse selection, Decarolis *et al.* (2017) and Decarolis and Guglielmo (2017) find 5-star contracts advantageously selected low risk enrollees with lower premium and generosity, and risk pool did not worsen relative to 4 and 4.5-star contracts.

[5]Echoing the disparity concern, there is evidence that plans serving enrollees with disability (MedPAC, 2015), in low socio-economic status (SES) (NQF, 2014) and certain geography (Soria-Saucedo *et al.*, 2016) are disfavored in the quality rating. In response, a categorical adjustment index (CAI) is applied to selected measures to adjust for SES factors in 2017.

2 Medicare Advantage

2.1 Plan bidding

Payment from government represents the largest source of revenue for MA insurers (Newhouse and McGuire, 2014). The competitive bidding model is introduced in the Medicare Modernization Act (MMA) of 2003. Under this model, the government sets statutory benchmarks capitation rates for each county from historic FFS costs, and MA insurers submit *bids* b to the government. CMS assigns a benchmark B to each plan as a weighted average of the county benchmarks in its service area. The bid reflects the projected cost of MA enrollees plus an administrative load. If the bid is below the benchmark, then the government pays off the bid, and in addition returns a fraction of cost saving below the benchmark – in MMA it is 75% of $B - b$ – to insurers as *rebate*. The rebate is then passed on to enrollees in the form of premium reduction or extra benefits. The vast majority of plans submit a bid below their benchmark, providing enrollees with more generous coverage than traditional Medicare at little extra cost above Part B premium.[6]

When the bid exceeds the benchmark, the insurer is paid the benchmark from the government but receives no rebate. The excess cost $b - B$ is passed on to enrollees as extra premium. Hence government spending per enrollee is capped at the benchmark, and plans with more cost saving offer more generous coverage. In practice, bid, benchmark, and rebate are multiplied by the enrollee risk score to calculate the final payment. The risk adjustment is intended to make different risk types equally profitable for plan payment (Brown *et al.*, 2014; Newhouse *et al.*, 2012). For MA enrollees with an average FFS risk score, the rebate formula is as follows

$$rebate_i^{MMA} = \begin{cases} 0 & \text{if } b_i \geq B_i, \\ 0.75 \cdot (B_i - b_i) & \text{if } b_i < B_i. \end{cases}$$

2.2 Quality rating

To better inform plan choice, quality rating is reported to consumers on a scale of 1 to 5 stars in the "star rating program." While more disaggregated quality information was previously available to potential enrollees in the "Medicare and You" handbook (CMS, 2008), 2009 is the first year when performances on multiple domains are aggregated in

[6] In our data 41% of the plans charge zero premium (above standard Part B premium), and 84% require no deductible for prescription drug (see Table 2).

a single star rating reported to consumers.[7] Previous research has found modest effect of quality rating on enrollment in 2009, but not significant effect in 2010 (Darden and McCarthy, 2015). The weaker demand effect in later years is partly due to supply side pricing response (McCarthy and Darden, 2017) – a mechanism we show has intensified when payment is later linked to quality – or due to consumer inattention to quality information after the policy phase-in.[8]

The star rating summarizes overall plan performance across eight domains, five concerning Part C coverage and three concerning prescription drug coverage. To calculate the domain rating and the final star rating, plans receive scores on specific quality measures within domains. Measure scores are based on performance data from a number of Centers for Medicare and Medicaid Services (CMS) administrative datasets. [9] Depending on the percentile rank, each measure is assigned a star rating. The measure ratings are then averaged to generate the final rating, or within domain to generate the domain rating.

The composition of measures in the final rating changed from year to year. Some measures become obsolete when new measures are introduced. Continuing measures are subject to higher quality standards.[10] Broadly speaking, in 2011-2014, quality measures are divided into the following eight domains (the data source is in parenthesis):

1. *staying healthy*: screening, vaccine, BMI (HEDIS), and self-reported physical and mental health (HOS)

2. *managing chronic conditions*: percent of enrollees diagnosed with diabetes, high blood sugar and cholesterol, etc., who have the condition controlled (HEDIS)

3. *plan responsiveness*: ease of getting needed care, setting up appointment, etc (CAHPS)

[7] This initiative was supported by theoretical analysis showing that a functioning quality reporting system is as important as risk adjustment in correcting market inefficiencies (Glazer and McGuire, 2006).

[8] For example, a 2011 poll by Kaiser Permanente shows that almost 60% of Medicare eligible seniors are unaware of the 5 Star Ratings (Harris Interactive, 2011).

[9] Sources of performance data include the Healthcare Effectiveness Data and Information Set (HEDIS), the Consumer Assessment of Healthcare Providers and Systems (CAHPS), the Health Outcomes Survey (HOS), the Complaints Tracking Module (CTM), the Independent Review Entity (IRE), the Medicare Beneficiary Database Suite of Systems (MBDSS), the Call Center, the Medicare Advantage and Prescription Drug System (MARx), the Prescription Drug Event (PDE), among others. Most measures reflect previous year plan quality, with most of the health outcome measures further lagged by two years. For a detailed list of measures in the 2013 rating, the source of each measure and the time frame of measurement, see Appendix Table A1 and A2.

[10] For example, final rating in 2011 is based on 51 measure stars, 50 measure stars in 2012, and 49 measure stars in 2013. The measure "access to primary care doctor", in particular, is dropped in 2013 because nearly all MA plans meet high quality status (above 85% for 4.0 star and 95% for 5.0 star), and the measure "plan all-cause (30-day) re-admission" revised the threshold for 5.0 rating from below 5% in 2012 to below 3% in 2013; only a handful of local coordinated care plans (CCP) with very small enrollments ever obtained 5.0 rating on this measure, and average readmission is 15% for MA enrollees (or 20% in FFS).

4. *consumer complaint*: rate of complaint received, number of enrollees leaving the plan and reported difficulty in care access (CTM)

5. *timely service*: call center availability and timely response and satisfactory resolution on consumer appeal (IRE)

6. *Part D timely service*: call center availability, timely response and satisfactory resolution on consumer appeal (IRE) and timely enrollment in drug plan (MARx)

7. *Part D experience*: ease of getting information on drug coverage and the needed drug from plan and member rating on plan (CAHPS)

8. *Part D safety and adherence*: percent choosing high-risk drug instead of a safer option and percent taking drugs as directed (PDE)

Measure stars are aggregated to the final rating through a weighting procedure that assigns higher weights to outcome measures and penalizes high variance across measures.[11] Possible weights are 1.0, 1.5 and 3.0. Measures of patient satisfaction and access typically receive the 1.5 weight, and measures of medical process such as screening, testing and vaccination, receive the 1.0 weight. Patient outcome measures, on the other hand, receive the highest weight (3.0), and are important predictors of the final rating.[12]

Due to the data collection effort, the quality rating for enrollment period t, released in the fall of $t-1$, is based on performance measured over $t-2$, especially for outcome measures in "staying healthy" and "managing chronic conditions". These measures capture the improvement in health conditions relative to a baseline.[13] Intended as a measure of quality, patient outcome is biased due to differences in baseline risk. While low-risk enrollees with milder, more manageable conditions tend to speak to high quality of care, high-risk enrollees with more complicated conditions may see smaller improvement in outcomes despite the high quality care applied in the treatment.

For an alternative measure, consider the case-mix index (CMI) intensely applied in the payment adjustment to hospital discharges. Patients are categorized by their "severity-

[11] The weighting procedure started in 2012, although nearly all measures are already present in 2011.

[12] For instance, in 2013 only outcome measures received the 3.0 weight. These measures are "improving physical health" and "improving mental health" in the"staying healthy" domain, management of blood sugar, blood pressure, and cholesterol, and all-cause re-admission measure in the "managing chronic condition" domain, and all the drug safety and adherence measures in the "Part D safety and adherence" domain.

[13] For example, measures of "maintaining and improving physical (mental) health" in the "staying healthy" domain come from respondent self-reports in the Health Outcomes Survey (HOS), adjusted by socio-demographic factors. Outcome improvement measures in the "managing chronic conditions" domain come from clinical records in the Healthcare Effectiveness Data and Information Set (HEDIS), unadjusted by baseline case-mix or socio-demographic factors.

adjusted diagnosis-related group (DRG)", which considers up to eight additional co-morbidities in additional to the principal diagnosis, up to six procedures performed in hospital, and adjusts for socio-demographic factors such as age and sex. The case-mix adjustment for baseline severity is nearly non-existent in clinical outcome measures in the MA quality rating.[14]

2.3 Quality bonus payment

The 2010 ACA modified the payment formula in the Medicare Advantage in multiple ways. First, it gradually reduced the benchmark faced by MA contracts to a level closer to FFS spending. The new benchmark ranges from 95% of FFS cost in counties in the top quartile of FFS cost, to 115% in those in the lowest quartile. In addition, rebate percentage varied by quality rating, whereas it was held constant at 75% before ACA.[15] In this paper, we refer to *bonus payments* as the sum of extra payments from either the bonus benchmark or the bonus rebate to higher quality plans.

The Quality Bonus Payment (QBP) demonstration, signed into law in November 2010, revised the ACA bonus payments in the demonstration period from 2012 to 2014. Benchmark and rebate bonus is extended to contracts with 3.0 and 3.5 star ratings. Total bonus payment is more generous under the demonstration, and is scheduled to phase into ACA levels in 2015. Moreover, QBP revised ACA benchmark bonus by designating a set of double bonus counties — contracts eligible for a 5% benchmark bonus receive a 10% bonus in these counties.[16] While our main focus is on the bonus payment differential at the contract level, we examine if high quality contracts differentially entered double bonus counties after QBP, and control for both the ACA and the QBP benchmark in cross-county analysis.

Hence the QBP rebate to an insurer of quality q_i serving enrollees comparable to the FFS risk pool is given by

[14] Girotti *et al.* (2013) presents a case study of how adjusting for complication severity can meaningfully alter the quality ranking in the context of vascular surgeries.

[15] In ACA, bonus benchmark and rebate percentage are awarded to high-performing contracts with at least a 4.0 star rating. The highest performing, 5.0-star plans receive a 5% bonus above the ACA benchmark, and a 70% rebate to enrollees if bid is below the bonus-inclusive benchmark. Bonus benchmark and rebate are more generous in 2012-2014, when the Quality Bonus Payment demonstration is effective.

[16] Extra bonus is awarded to counties based on a number of criteria, including population size, MA penetration rate, and FFS costs relative to the national average. Most counties receive some extra bonus, if not a 100% top-off. Layton and Ryan (2015) defines a double bonus county as having at least a 80% top-off: a plan eligible for 5% benchmark bonus receives more than 9% bonus in this county. They find double bonus counties are not associated with higher quality, but with a greater number of plans offered.

$$rebate_i^{QBP} = \begin{cases} 0 & \text{if } b_i \geq \alpha(q_i) \cdot B_i \\ \tau(q_i) \cdot (\alpha(q_i) \cdot B_i - b_i) & \text{if } b_i < \alpha(q_i) \cdot B_i \end{cases}$$

where B_i is the average ACA benchmark rate across service area, and $\alpha(q_i) \cdot B_i$ is the quality-adjusted benchmark relevant for plan bidding. $\alpha(q_i)$ adjusts for the benchmark bonus from plan quality and the top-off bonus from double bonus counties. Abstracting from the top-off, Table 1 shows the variation of bonus benchmark $\alpha(q_i)$ and rebate $\tau(q_i)$ over quality rating. Although both benchmarks and rebates are lower than the 2009-2011 period, the reduction is smaller for higher quality contracts due to more generous quality bonus payments.

Table 1: Bonus and rebates by quality scores for the period 2009-2014

	Star Rating					
Year	1 - 2.5	3	3.5	4	4.5	5
	Benchmark Bonus $\alpha(q_j) = 1 + \%$					
2009/11	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2012	0.0%	3.0%	3.5%	4.0%	4.0%	5.0%
2013	0.0%	3.0%	3.5%	4.0%	4.0%	5.0%
2014	0.0%	3.0%	3.5%	5.0%	5.0%	5.0%
	Rebate Percentage $\tau(q_j)$					
2009/11	75.0%	75.0%	75.0%	75.0%	75.0%	75.0%
2012	66.7%	66.7%	71.7%	71.7%	73.3%	73.3%
2013	58.3%	58.3%	68.3%	68.3%	71.7%	71.7%
2014	50.0%	50.0%	65.0%	65.0%	70.0%	70.0%

3 Data summary

Our data come from administrative CMS registry of all MA-PD plans offered over 2009-2014 (the Landscape File). The data contains detailed insurer (or contract) information, such as quality ratings, and within-insurer plan availability across service areas (counties). Moreover, within each county, plan characteristics, such as premium and drug deductible are observed. A separate enrollment file contains plan-county-month enrollment counts, which have been aggregated into plan-county-yearly counts.

Because the quality rating varies at the level of contracts, to understand how quality-

linked payment affects pricing and product design, we aggregate plan-county characteristics to the contract level, weighted by enrollment share. Plans in counties with fewer than 10 enrollees are dropped, since CMS masks the exact enrollment count in these cases. We further restrict contracts to those with at least a 3.0 rating in the baseline (2009-2010). Because contracts failing to obtain a 3.0 or above rating for three consecutive years are subject to suspension, low performing contracts face additional incentive of risk selection not generalizable to higher performing contracts.

Table 2 summarizes key contract-year statistics in Panel A for contracts with non-missing quality rating in the previous year.[17] The treated group is defined as high quality contracts with at least a 4.0 rating in both 2009 and 2010, and the control group as low quality contracts with at most a 3.5 (but no less than 3.0) rating in 2009-2010. Column (1)-(2) pool over both treated and control contracts: an average MA-PD contract has 3 plans serving over 25 counties. Baseline high quality contracts are more likely to remain high quality (star$\geq$ 4.0) in the sample period, bid closer to the benchmark, and receive smaller rebates.[18] They are also more likely to charge higher premium, and less likely to offer zero-premium plans. Differences in drug deductible, on the other hand, are small and not significant.

Panel B shows more disaggregated variation at the level of contract, year, and location (county). Because contracts can design plan characteristics differentially across service areas, the within-contract cross-location variation is one margin of selection overlooked in cross-contract comparisons. We cluster standard errors two-way at the level of contract and county in Panel B. We therefore allow a given contract to be arbitrarily correlated over time within county, and a given county arbitrarily correlated over time within contracts. As in Panel A, high quality contracts charge higher premium and somewhat lower drug deductible, and the difference in deductible is not statistically significant.

4 Contract-level evidence

We start the empirical analysis with a contract-level difference-in-difference model. Baseline high quality contracts are more likely to experience higher bonus payments after the reform, and they form the treated group. Because the ACA was signed into law in April

[17] For the main analysis we restrict attention to continuing contracts, since bonus payment for enrollment year t is determined by year $t-1$ quality rating. New contracts, or contracts with missing previous rating, are eligible for bonus payments according to a different rule. In robustness analysis we examine how the entry of new contracts may respond to region characteristics (FFS risk and benchmark rate, for example) after QBP. We do not find differential entry response at the contract level.

[18] Standard errors are clustered at the level of contract linked over time.

Table 2: Summary statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	full sample		low quality		high quality		(5)-(3)
	mean	s.e.	mean	s.e.	mean	s.e.	p-value
			Panel A: contract-year observations				
	$N = 1,122$		$N = 775$		$N = 347$		
risk score	0.97	0.0075	0.97	0.0093	0.96	0.012	0.55
star≥ 4.0 (%)	0.35	0.025	0.17	0.018	0.75	0.032	0.00
star score	3.55	0.031	3.35	0.027	3.99	0.041	0.00
# county	25.09	5.40	28.19	7.74	18.18	2.21	0.22
# plan	3.40	0.23	3.53	0.31	3.12	0.28	0.33
enrollment (k)	334.75	34.95	328.35	39.19	349.06	71.56	0.80
benchmark	874.10	5.72	883.08	6.52	854.06	10.87	0.023
bid	763.38	6.28	763.65	7.58	762.80	11.25	0.95
benchmark-bid	110.72	5.55	119.43	6.90	91.27	8.68	0.012
rebate	78.37	3.73	83.55	4.68	66.80	5.74	0.025
premium	49.07	3.38	35.25	3.59	79.93	5.70	0.00
zero premium (%)	0.41	0.029	0.51	0.035	0.19	0.035	0.00
drug deduc	32.62	4.42	32.85	5.72	32.11	6.40	0.93
zero drug deduc (%)	0.84	0.019	0.85	0.024	0.83	0.031	0.68
			Panel B: contract-year-location observations				
	$N = 20,472$		$N = 14,861$		$N = 5,611$		
enrollment (k)	18.25	2.35	17.00	2.48	21.57	4.64	0.35
# plan	1.76	0.073	1.59	0.088	2.22	0.093	0.00
premium	52.69	3.78	42.93	4.21	78.55	6.71	0.00
zero premium (%)	0.33	0.036	0.39	0.047	0.16	0.039	0.00
drug deduc	28.65	5.88	30.05	7.69	24.92	6.23	0.60
zero drug deduc (%)	0.85	0.030	0.84	0.040	0.87	0.031	0.44

Notes: Table shows summary statistics for the full sample (column 1-2) and the treated (baseline high quality, column 5-6) contracts and control (baseline low quality, column 3-4) contracts. Plan characteristics are aggregated to the contract-year level in Panel A, and to contract-year-county level in Panel B, both weighted by enrollment. Standard errors are clustered at the level of contracts in Panel A, and two-way clustered at the level of contract and county in Panel B. Details are in the text.

2010, and MA contracts do not submit bid and benefit design for 2011 enrollment until June 2010, insurer response to quality payment incentives may already be detectable in 2011 contract design and enrollment outcomes. We hence define the variable $post = 1$ for year 2011 and after, and inspect the timing of the effect in detail in event studies below. The difference-in-difference model is given by

$$y_{ct} = \beta \cdot high_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}$$

where we compare contracts (c) with differential baseline quality ratings over time (t).

We assume that the trending of high and low contracts is parallel absent the policy. To sharpen the identification of trends, we include dummies of longitudinal contract id's (α_c), and use within-contract variation over time to isolate the effect of bonus payments. The contract fixed effects importantly sweep out baseline heterogeneity across contracts, such as differences in service area, enrollee characteristics and provider networks, among others. However, to the extent that confounding factors may vary around the same time as the reform, the difference-in-difference estimate β will be biased. The absence of time-varying confounds is not directly testable. As with most difference-in-difference analyses, we rely on visual inspection of parallel trends before the reform to assess the validity of the model. We then apply the model to study the effect of bonus payment on risk score, market characteristics and pricing.

4.1 Risk score

Table 3 shows the effect on risk score following the reform in 2011. We show the robustness of the result to different controls and level of analysis. Columns (1)-(2) measure risk score aggregated at the contract level using plan enrollment weights. Columns (3)-(4) measure risk score as unweighted average of plan risk scores. Columns (5)-(6) measure risk score at the raw plan level. For each measure, we study the robustness of results with and without contract or plan level fixed effects.

INSERT TABLE 3 APPROXIMATELY HERE

We find significant improvement in the risk pool of higher quality contracts, relative to the control. The preferred specification in column (1) shows a 2.6 percentage point reduction in risk score by high quality contracts, and the effect is similar if not larger under alternative measures.

Figure 1 examines the validity of the difference-in-difference design and the timing of the effect. Before the proposal of quality bonus payments became law in 2010, risk scores

for both treatment and control groups stay parallel. The trend departs visibly in 2011: for the first time high quality contracts have *lower* risk score than low quality contracts, and the gap widens in later years. Similarly, in the event study, although the effect in 2011 is not statistically significant, the effect becomes stronger and more significant over time.

INSERT FIGURE 1 APPROXIMATELY HERE

We then turn to the potential mechanism of risk selection by high quality contracts. Although risk pool on average improves, contracts facing different market structure and enrollee base may differ in their ability to risk-select. If the improvement in risk pool is concentrated among the set of contracts with certain characteristics, then the heterogeneous effect is indicative of the mechanism of risk selection. We consider two dimensions of heterogeneity based on baseline characteristics: risk composition across service areas, and market competition measured by the Herfindahl-Hirschman index (HHI).

To proceed, we first define the market set for each contract as the union of counties served in baseline 2009-2010. For each county in the market set, we attach the average FFS risk score over the baseline as a measure of potential gain from risk selection in this county: lower FFS risk score implies new enrollments into private Medicare likely have lower risk, making the county more advantageous for risk selection. We then average over counties to derive a risk selection measure at the contract level. The median across contracts is 0.99, and the 15th (85th) percentile is 0.90 (1.07).

We hence posit that high quality contracts with service area risk score below the median are more likely to succeed in risk selection, relative to low quality contracts with service area risk score above the median. Furthermore, we expected to see even stronger effects at the 15% tails, where we compare high quality contracts below the 15th percentile of service area risk with low quality contracts above the 85th percentile.[19]

INSERT TABLE 4 APPROXIMATELY HERE

Columns (1)-(2) in Table 4 show significant reduction in the risk score of high quality contracts more advantageous for risk selection: those in the lower 15% of service area risk distribution decreased risk score by 4.5 percentage points, relative to low quality contracts in the upper 15% of the distribution. Figure 2 shows that the effect is visible starting in year 2011, and becomes stronger and more significant in later years.

[19] Bauhoff (2012) conducted a field experiment to test supply-side selection among private insurers in Germany. He finds that plans are less likely to follow-up on applications from high-risk regions.

INSERT FIGURE 2 APPROXIMATELY HERE

Alternatively, one might believe contracts with greater market power are better able to risk select. To derive a measure of market power at the contract level, we first compute the HHI for each county over the baseline, and then average over the market set for contracts. The median HHI in the baseline is 0.44, and the 15th (85th) percentile is 0.31 (0.61).

Columns (3)-(4) in Table 4 show the same effect by market competition. In the full sample, baseline high quality contracts with high market power reduced their risk score by 1.8 percentage points relative to low quality contracts with low market power, although the effect is not significant. The result is more tenuous when comparing the 15% tails. Figure 3 shows the corresponding raw trends and event study estimates.[20] Overall, unlike service area risk score, there is no clear evidence that market competition has strong bearing on risk selection in this context. Therefore, in the within-contract triple-difference analysis below, we utilize the risk score variation across service areas to detect any differential pricing response that may have contributed to changes in the risk pool.

INSERT FIGURE 3 APPROXIMATELY HERE

4.2 Market characteristics

Another possible explanation to the increasing risk selection that emerged in the previous section concerns contracts altering the characteristics of the service areas following the reform. For example, high quality contracts may have expanded their service areas to include more counties with low risk scores, or may have increased the number of plans offered in these counties. The exact mechanism of risk selection has implications for the empirical strategy suitable for the analysis. In particular, if characteristics of service area responded endogenously to bonus payment incentives, then within-contract cross-location variation should not be interpreted as exogenous.

To detect changes in market characteristics attributable to changes in the market set, we replace yearly county characteristics with values in 2012, and then average over the market set for contract-year observations. The resulting variable captures the effect of market composition on contract characteristics, rather than the temporal variation in these characteristics. For example, if the service area FFS risk score improved, then there is reason to believe contracts may have entered low risk counties or exited high risk counties, depending on the change in market size.

[20] Comparing high quality contracts with HHI below the 15th percentile with low quality contracts with HHI above the 85th percentile also renders insignificant estimate.

Table 5 shows little change in market set characteristics. There is some evidence of high quality contracts expanding their market set over time, but the effect is not significant. The number of plans offered also did not change (column 5). Importantly, service area risk (column 2) barely changed for high quality contracts after the reform, suggesting contracts did not differentially enter or exit counties based on the baseline risk. While counties with low FFS risk score are more advantageous for selecting low cost enrollees, there is no evidence of extensive margin selection whereby high quality contracts expand their market set to include more of these low risk counties.

In addition, since the QBP also varied the county benchmark rate by quality rating, we check if contracts differentially select into counties with higher ACA benchmark (column 3) unadjusted by quality, or double-bonus counties (column 4) under the QBP where benchmark bonus to 5-star contracts is over 8%, or more than a 60% top-off. We see no evidence of differential selection by high quality contracts along these margins.

INSERT TABLE 5 APPROXIMATELY HERE

4.3 Bid, rebate, and pricing

Since market characteristics did not change significantly to explain the risk selection result, we now turn to assess how pricing responded to quality bonus payments. Under the law, higher quality contracts face higher benchmark. In principle, contracts can increase the bid to receive higher payments without reducing the rebate to enrollees. Furthermore, the rebate bonus to quality allows the bid to increase more than the benchmark for a fixed amount of rebate. Of course, rebates need not stay constant, if part of the bonus payment is passed on to enrollees in the form of lower premium or cost-sharing.

Table 6 studies the effect of QBP on bids and rebates. As a result of the bonuses introduced by QBP, benchmarks for high quality contracts increased by $27.84 (cf Table 1). In response, high quality contracts raised their bids by $37.01, resulting in a net narrowing of the benchmark-bid gap by $9.17. However, adjusting for the rebate bonus in the QBP, the final rebate accruing to enrollees did not change significantly. At the contract level, enrollees in high quality contracts received $0.40 more in rebate, but this effect is not significant.[21]

INSERT TABLE 6 APPROXIMATELY HERE

[21] Since the QBP reform is essentially a supply-side shock directly affecting revenues rather than marginal costs, the estimates in Table 6 suggest that bids do not *only* depend on marginal costs (Song *et al.*, 2012, 2013) – a finding highlighted in Curto *et al.* (2014) as well.

Absent large changes in the rebate, changes in premium and cost-sharing are also small. Table 7 shows noisy effects on average premiums and the offering of zero-premium contracts. For cost-sharing, we focus on drug deductible. Although the vast majority of contracts have zero deductible, the mass at zero did not change after the reform. Average drug deductible decreased by half. On the raw trend, however, the difference appears to be driven by an early increase in deductible by low quality contracts in 2011: starting in 2012 both high and low quality contracts increased deductible at a similar rate (Figure 4). The effect, moreover, is only marginally significant. Therefore aggregated at the level of contracts, there is no clear mapping from pricing variation to risk selection through quality. In particular, high quality contracts did not increase premiums or drug deductibles, measures commonly used to select low risk enrollees, nor did they decrease them, which is consistent with a null effect on rebate.

INSERT TABLE 7 APPROXIMATELY HERE

INSERT FIGURE 4 APPROXIMATELY HERE

The contract-level comparison, however, does not capture the within-contract cross-location pricing and benefit design. The within-contract selection can potentially alter the risk pool composition without revealing any marked changes in contract-level pricing, if, for example, price increases in higher risk areas are offset by decreases in lower risk areas. To further probe this possibility, in the analysis below, we investigate how pricing varies within contract across counties above and below the median risk county in the market set.

5 Within-contract cross-county evidence

This section looks inside the market set, and examines if contracts are more likely to deploy high-premium, high-deductible plans in service areas less advantageous for risk selection. For identification, we assume that absent the policy, the distribution of insurance pricing across risk regions is parallel for both high and low quality contracts. As with most difference-in-difference analysis, we assess the plausibly of the identifying assumption by inspecting the pre-trend commonality for high and low risk regions within and between high and low quality contracts.

To measure average prices at the contract-year-location level, we weight plan premiums and deductibles by enrollment. To measure a county's relative standing in the overall risk composition across service areas, we measure the distance of county risk to the median

risk in the market set, and use the distance to median as the driving variation in the within-contract analysis.

In particular, we use baseline market set characteristics in 2009-2010 to derive the distance measure. We rank all counties served by a contract in the baseline by their baseline FFS risk score. Comparing counties with the median county gives the distance-to-median measure. Note that this measure is fixed for a given contract-location pair, and does not vary over time.

We estimate the following triple-difference design

$$y_{clt} = \beta_0 \cdot risk_{cl} \cdot high_c \cdot post_t + \beta_1 \cdot risk_{cl} \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt},$$

where the unit of observation is at the level of contract c, location (county) l, and year t. $risk_{cl}$ is the distance-to-median measure. We include year indicators τ_t to control for common temporal shocks, and contract-county indicators α_{cl} to absorb unobserved heterogeneity across contracts and service areas, and the baseline selection between the two. The fixed effects would not be adequate to address time-varying selection response if contracts are shown to enter or exit service areas based on local risk factors after the reform. This is not the case, however, as we showed in Section 4.2.

In addition, we control for time-varying location-specific factors in X_{lt}. Most notably, since the raw county benchmark is time-varying, and since QBP increased the bonus benchmark for some counties (commonly known as "double-bonus counties"), the set of which is also time varying, we include the raw benchmark, the bonus payment rate, and their interaction.[22] We hence allow for separate pricing response to benchmark variation in local markets.

We cluster standard errors two-way at the level of county and contract: contracts observed in different counties are correlated, and so are counties entering different contracts' service areas; within contracts (counties), counties (contracts) are assumed to be independent. Clustering at the intersection of county and contract gives similar standard errors.

INSERT TABLE 8 APPROXIMATELY HERE

INSERT FIGURE 5 APPROXIMATELY HERE

Table 8 displays the premium response to QBP across risk regions for baseline low quality (column 1) and high quality (column 2) contracts, and the differential response by

[22] The interaction measures the maximum benchmark faced by 5-star contracts serving the county in a given year.

high quality contracts (column 3). Low quality contracts increased premium in lower risk counties, whereas high quality contracts increased premium in higher risk counties. Both effects are visually perceptible on the raw trend (Figure 5, Panel a), where the market set of each contract is divided into high (above median) and low (below median) risk regions. Hence the results are not driven by the parametric assumption that effects are linear in the deviation from median.

The triple-difference estimate suggests high quality contracts increased premium by $0.31 per one percentage point increase in risk above the median. Event study shows parallel pre-trend: high and low quality contracts charged premium similarly across risk regions. After the passage of QBP, high quality contracts significantly increased premium in higher risk counties, whereas response by low quality contracts is small and not significant in most years (Figure 5, Panel b). The same pattern holds when we only include counties in the lower or upper 15% of the risk distribution given contract (columns 4-6).

We then investigate any differential pricing in drug deductible. The contract-level analysis suggests that both low and high quality contracts increased deductibles after QBP (see Table 7 and Figure 4, Panel c). Looking within contract across risk regions, Table 9 shows that both contracts raised deductibles more in regions less advantageous for risk selection, and that this expansion is not significantly larger for high quality contracts (see also the raw trend in Figure 6). We similarly do not detect any significant pricing differential by quality when looking at the 15% tails of the market set.

INSERT TABLE 9 APPROXIMATELY HERE

INSERT FIGURE 6 APPROXIMATELY HERE

Put together, we find evidence for *within*-contract pricing adjustments across risk regions, which might explain the risk pool improvement for high quality contracts absent any significant change in average prices (including rebate) at the contract level. The adjustment mostly affects premiums, not drug deductibles, although it may also affect other contract characteristics outside our sample. The premium adjustment, in particular, illustrates one potential mechanism of favorable selection into Medicare Advantage: premium is higher in markets where new enrollees have higher risk, and lower in markets with lower risk. Because premium did not vary at the contract level but varied across service regions, the selection response raises questions on the distributional incidence of quality payments, and its implication for welfare.

6 Why does QBP induce risk selection?

So far we have shown high quality contracts significantly improved risk pool after QBP, and linked the improvement to differential premium pricing across risk regions. That is, we have suggested mechanisms of *how* risk selection is accomplished, but have been silent on *why* risk selection is incentivized in the first place: what are the design features in QBP that made it profitable for high quality contracts to select low risk enrollees?

This section suggests the incentive may lie with the design of the quality rating system, and is activated by the financial reward to high quality introduced in QBP. When high quality contracts are able to keep most of the quality bonus payment as profit rather than rebate to consumers, risk selection becomes profitable if lower enrollee risk contributes to higher quality and hence continued bonus payment. We highlight the linkage between risk, quality and insurer profit in a stylized model below.

We then empirically characterize the correlation between risk and quality. We exploit the same difference-in-difference variation as in Section 4 to show that contracts with higher risk score in the baseline are less likely to perform well in patient outcome measures in the quality rating. Furthermore, high quality contracts with high baseline risk scores experience smaller improvement in outcome ratings relative to low quality contracts with a more favorable risk pool. We further document a negative correlation between risk score and the Star score, particularly for baseline high quality contracts. For these contracts, performance in outcome-related domains are the most predictive of the final rating.

6.1 A model of risk, quality, and insurer profit

To illustrate how the risk-quality linkage can affect insurer pricing, we build a simple 2-period model where the risk pool in the first period affects quality rating and payment in the second period.

We focus on the decision of baseline high quality contracts. Revenues in each period is the sum of a benchmark and a premium charged to the enrollees.[23] The benchmark B is fixed in the first period. In the second period, high quality contracts receive higher benchmark $B_h > B_l$.

Contract quality is not constant: with probability $\lambda(h;r)$, baseline high quality contracts remain high quality in the second period. We model the risk-quality linkage by allowing the transition probability to depend on the risk score r from period one. If low risk score contributes to future high quality, then there is incentive to attract low risk enrollees in

[23] Alternatively, one can think of firms submitting a bid, which then translates into consumer premium according to a known regulatory formula.

the current period. We assume risk adjustment is perfect, so that risk selection would have no bearing on insurer profit, absent the dynamic linkage on quality and quality payments.

A high quality contract chooses premium (p, p') in both periods to maximize profits

$$\Pi(p, p') = (p - c + B_h)\, s_h(p) + \sum_{j \in \{l,h\}} \lambda(j; r)\Big(p' - c + B_j\Big) s_j(p'),$$

where the insurer faces constant marginal cost c, and demand is allowed to differ by quality in $s_j(p)$, $j \in \{l, h\}$. The optimal premium set in the first period is

$$p^* = (c - B) \cdot \left[1 - \frac{1}{|\epsilon_h|} \cdot \Big(1 + \frac{\Delta\pi'^{\star}}{s_h(p)} \cdot \overbrace{\frac{d\lambda}{dr}}^{\text{policy}} \cdot \underbrace{\frac{dr}{dp}}_{\text{selection}} \Big)\right]^{-1}, \tag{1}$$

where $\Delta\pi'^{\star} > 0$ is the optimized profit difference between high and low quality in the second period,[24] and $\epsilon_h < 0$ is the premium elasticity of demand for high quality contracts.

Absent the risk-quality linkage, we have the standard result that optimal premium equals marginal cost plus a mark-up inverse to demand elasticity. When $\frac{d\lambda}{dr} \neq 0$, however, optimal premium responds to the selection term $\frac{dr}{dp}$. Specifically, a market is more advantageous to risk selection, if lower premium attracts enrollees below the average risk of the contract, or $\frac{dr}{dp} > 0$. In these markets, when enrollee risk lowers contract quality, the term $\frac{\Delta\pi'^{\star}}{s_h(p)} \cdot \frac{d\lambda}{dr} \cdot \frac{dr}{dp}$ is signed negative, pushing premium *below* the standard level where $\frac{d\lambda}{dr} = 0$.[25]

Hence observed premium variation is consistent with insurer risk selection, if lower enrollee risk contributes to higher quality rating. Under this configuration, we should expect lower (higher) premium in lower (higher) risk regions, relative to the baseline period where $\frac{d\lambda}{dr} = 0$. We then examine the risk-quality linkage in detail, signing $\frac{d\lambda}{dr}$ empirically.

[24] That is, $\Delta\pi'^{\star} = (p'_h - c + B_h)s_h(p'_h) - (p'_l - c + B_l)s_l(p'_l) > 0$, where $\{p'_h, p'_l\}$ is the vector of optimal premiums to be charged in the second period. In addition, we assume that high quality firms have higher profits than low quality ones.

[25] Alternatively, in markets less advantageous to risk selection with $\frac{dr}{dp} < 0$, premium is set higher than the benchmark level.

6.2 Difference-in-difference evidence on the risk-quality mechanism

Before empirically characterizing the risk-quality correlation, we present difference-in-difference evidence on the nature of the correlation using similar variation as in previous sections. One challenge is that, because the rating algorithm underwent substantial revision in 2011, the same year quality bonus payment was introduced, differential trending in the quality rating after the reform may reflect mechanical differences in the rating computation, rather than insurer risk selection response to payments.

On the other hand, the selection response implies that low risk enrollees with fewer diagnoses are associated with higher quality, possibly through improvements in outcome measures in the quality rating. We hence focus on outcome measures, and document the relationship between ratings in these measures and insurer risk profiles in the baseline. One particular advantage of this strategy is that outcome measures are relatively stable over the sample period, allowing for a difference-in-difference characterization of the risk-outcome channel unaffected by changes in rating measurement or computation.

Specifically, we focus on outcome measures that are consistently measured from 2009 to 2014. They are improving physical health, improving mental health, diabetes controlled–blood sugar, diabetes controlled–cholesterol, and blood pressure controlled from Part C.[26] We average over these measures to derive a summary star rating of patient outcome. When we regress the final rating on the constructed outcome rating, the coefficient before the outcome rating mechanically increases after 2012 (Table 10). Although the estimate may also reflect changes in the rating computation other than the weighting, the importance of outcome in the final rating generally increased over the 2012-2014 period (Figure 7).

INSERT TABLE 10 APPROXIMATELY HERE

INSERT FIGURE 7 APPROXIMATELY HERE

We then examine the correlation between risk and outcome measures using the difference-in-difference variation, comparing the outcome rating of contracts serving low versus high-risk enrollees in the baseline. Table 11 shows a ten percentage point increase in baseline risk score in 2009-2010 reduces outcome rating by 12.2 percentage points **(Can we replicate this result using the discrete Star score? We could have a multinomial logit specification, and the parameter will tell us the probability of being of a certain score)**. When we group outcomes by health improvement measures in

[26] Part D outcome measures in the "drug safety and adherence" domain are not consistently present over the sample period. In particular, three new outcome measures (medication adherence for diabetes, hypertension and high cholesterol, respectively) are added in 2012.

the Health Outcome Survey (column 2) and chronic condition measures in the Healthcare Effectiveness Data and Information Set (column 3), the risk-outcome correlation appears entirely driven by the chronic condition measures, with significant advantage to contracts serving low-risk enrollees in the baseline (Figure 8).

INSERT TABLE 11 APPROXIMATELY HERE

INSERT FIGURE 8 APPROXIMATELY HERE

We further inspect the risk-outcome correlation by baseline quality status. Specifically, the treated group is the set of baseline high quality contracts where enrollee risk score over 2009-2010 is above the median of all contracts, and the control is baseline low quality contracts with risk scores below the median.[27] If high risk enrollees are associated with worse outcomes measured in the quality rating, then a more favorable risk pool may help baseline low quality contracts obtain high quality standing. High quality contracts serving high risk enrollees, on the other hand, risk losing bonus payments if outscored in outcome measures by control contracts with low risk enrollees.

Table 12 suggests that the risk-outcome correlation disadvantages high quality contracts with high risk enrollees, where loss of high quality standing to low-risk low-quality contracts is more likely. In odd columns, we compare high and low quality without interacting with baseline risk. Overall, high quality contracts are less likely to retain high outcome rating as low quality contracts are likely to obtain it, in particular due to the difficulty in consistently managing chronic conditions related to blood pressure and diabetes. The difference is more striking, once we compare baseline high-risk high-quality contracts with low-risk low-quality contracts in even columns: falling in the upper half of the risk score distribution exposed high quality contracts to an additional 20 percentage point slippage in outcome rating, relative to low-quality competitors in the more favorable half of the risk distribution (Figure 9).

INSERT TABLE 12 APPROXIMATELY HERE

INSERT FIGURE 9 APPROXIMATELY HERE

6.3 Characterizing the risk-outcome correlation

The difference-in-difference evidence suggests a strong negative correlation between enrollee risk score and outcome rating, in particular, rating in managing chronic conditions

[27] Baseline risk score ranges from 0.74 to 1.46 in 2009-2010 for treated and control contracts, and the median is 0.97.

such as diabetes and blood pressure measured in HEDIS. Instead of using baseline enrollee risk, in this section, we directly characterize the empirical correlation between contemporaneous risk score and outcome rating. To do this, we note that outcome measures relevant for year t quality rating, announced in the fall of year $t-1$, are collected from patients enrolled in the contract two years prior in $t-2$. Due to the timing, we expect a strong negative correlation only between year t rating and year $t-2$ risk score, but not across other lag or lead periods.

Table 13 shows the correlation between year t outcome rating in diabetes and blood pressure control and risk scores from multiple periods. That is, in addition to contemporaneous correlation, we also examine correlation with risk score one year in lag ($riskscore_{t-3}$) up to two years in lead ($riskscore_t$). We further stratify the exercise by baseline contract quality. The risk-outcome correlation is weak among low quality contracts (baseline 3.0-3.5 stars), but becomes more negative and significant as we restrict the sample to higher quality contracts. For the set of contracts achieving at least a 4.5 star rating in the baseline, a ten percentage point increase in enrollee risk score is associated with 33 percentage point decrease in chronic outcome rating. There is, however, no clear and consistent pattern between risk score and outcome for different lag and lead periods. Similar results hold for the correlation between risk score and final star rating (Table 14).

INSERT TABLE 13 APPROXIMATELY HERE

INSERT TABLE 14 APPROXIMATELY HERE

7 Discussion

Central to our finding is the premium variation *within* contract across risk regions. Specifically, high quality contracts differentially attract low-risk enrollees with lower premium in low-risk counties, and improve risk score relative to low quality contracts. Absent changes in average prices, selection implies that within contract, rebates are transferred from high risk to low risk enrollees. The insurer selection thus calls into question the distributional incidence of quality payment, with immediate policy and welfare implications.

First, the benefit of quality improvement may be weighed down by the social cost of unequal access to quality, as high-performing contracts disproportionately serve low-risk enrollees in low-risk counties. To the extent that enrollees in worse health conditions benefit more from quality care, return to care is higher if enrollment increased more in areas with higher baseline risk. In this context, disparity is worsened by a negative

correlation between risk and quality, which tends to penalize contracts serving high-risk enrollees. Alternatively, a positive risk-quality correlation may improve equity by encouraging more high-quality entry in high risk counties, although it is not clear why a quality measure should respond to the risk composition.

More fundamentally, one may question the role of baseline health in measures of plan quality – if quality reflects the "value-added" of health care on health, with higher quality improving health more, then baseline health should be swept out of quality ratings as a fixed effect. In that sense, the empirical correlation with pre-enrollment risk measures should be tenuous. However, using health improvement as indicator of quality may still fall short, if baseline health affects treatment efficacy interactively. This would be the case if, for example, patients with milder conditions recover sooner than those with more severe conditions, and the difference narrows but does not disappear at higher quality.

Therefore to effectively take out the risk confound in quality measures, the current outcome measures need to be adjusted by the severity of baseline health conditions. The idea is that, *conditional* on a similar case-mix of diagnoses and potential interactions, differential improvements in patient outcome are more plausibly attributable to health care quality rather than baseline health. A ready way to control for the risk score in the Star computation, as it predicts health care costs based on past diagnoses. Other adjustments may include socio-demographic factors used in HOS, and the intensively coded case-mix index for hospital payments.

In our analysis, we find the risk-quality correlation operates through the outcome measures in the "managing chronic conditions" domain. Most of these measures treat health improvement, i.e., having chronic conditions controlled, as indicator of quality. For example, a measure of blood pressure control is given by the fraction of baseline hypertension patients (denominator) with blood pressure below 140/90 (numerator) in the measurement period. However, these measures do not adjust for the severity of baseline conditions. This allows baseline risk score, a crude measure of severity, to correlate significantly and negatively with performance in chronic outcome measures. By contrast, self-reported health improvement in HOS is adjusted by respondent socio-demographic characteristics, and we find little residual correlation between risk score and health improvement measures in the "staying healthy" domain of the quality rating.

Hence our results suggest the risk-quality correlation, as well as its perverse incentive on risk selection, is largely suppressed if chronic outcome measures in HEDIS are appropriately adjusted for baseline severity of conditions. Currently, any such adjustment is lacking for these outcomes. An alternative approach is to adjust for enrollee risk score ex-post at the stage of quality payment. In practice, it would require policy makers form

the right belief as to the bias in quality rating ($\frac{d\lambda}{dr}$ in Equation 1) and the magnitude of the selection response ($\frac{dr}{dp}$) to effectively offset the selection incentive. We thereby argue that adjusting for ex-ante risk in the quality rating is simple to implement, recovers a less biased measure of quality, and can go a long way in reducing the selection incentives associated with quality payments.

8 Conclusion

We examine the insurer selection response to quality bonus payments in the Medicare Advantage (MA) market. The 2012 onset of the Quality Bonus Payment (QBP) demonstration varied payment generosity by quality rating, which we exploit in the difference-in-difference analysis. We find that pass-through of the bonus payments to enrollees is minimal: high quality contracts eligible for higher bonus payments increased bids by nearly the full amount of the bonus, leaving rebate to enrollee unchanged. Correspondingly, premium did not differentially decrease, or generosity increase, for enrollees in high quality contracts.

Within contract, however, we uncover significant premium variation across high and low cost counties for high quality contracts, but not for low quality contracts. Coupled with a higher baseline concentration in low risk counties, risk pool improved significantly for high quality contracts after the payment reform. We provide suggestive evidence that a negative correlation between enrollee risk score and patient health outcome can explain the selection response. In this case, low risk enrollees contribute to continued high quality rating and bonus payments.

These results have important normative implications for quality bonus payments in the MA market and similar value-based models elsewhere. One fundamental issue is the measurement of quality. To take out the risk confound, the current outcome measures need to be adjusted for the case-mix and severity of baseline conditions, and ideally be made conditional on risk scores. Failure to do so raises complicated distributional issues. For example, high quality care is made less accessible to the low-income, high-risk population, who potentially benefit more from quality. The equity concern among inframarginal enrollees, and a near zero pass-through on average, illustrate the subtle but critical role of insurer selection in the welfare incidence of value-based payments.

References

Bauhoff, S. (2012). Do health plans risk-select? an audit study on germany's social health insurance. *Journal of Public Economics*, **96** (9), 750–759.

Brown, J., Duggan, M., Kuziemko, I. and Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. *The American Economic Review*, **104** (10), 3335–3364.

Burwell, S. M. (2015). Setting value-based payment goals — HHS efforts to improve US health care. *New England Journal of Medicine*, **372** (10), 897–899.

Cabral, M., Geruso, M. and Mahoney, N. (2018). Do larger health insurance subsidies benefit patients or producers? Evidence from Medicare Advantage. *American Economic Review*, **108** (8), 2048—-2087.

Carey, C. (2017). Technological change and risk adjustment: Benefit design incentives in Medicare Part D. *American Economic Journal: Economic Policy*, **9** (1), 38–73.

— (2018). Sharing the burden of subsidization: Evidence on pass-through from a payment revision in medicare part d, Mimeo, Cornell University.

Clemens, J. and Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, **104** (4), 1320–49.

CMS (2008). *Medicare and you 2008*. 10050, DIANE Publishing.

CMS (2016). Nhe fact sheet. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html, accessed: 2018-06-08.

Congressional Budget Office (2017). Medicare - congressional budget office's june 2017 baseline. https://www.cbo.gov/sites/default/files/recurringdata/51302-2017-06-medicare.pdf, accessed: 2018-03-30.

Curto, V., Einav, L., Levin, J. and Bhattacharya, J. (2014). Can health insurance competition work? Evidence from medicare advantage, National Bureau of Economic Research, No. w20818.

Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, **95** (5), 1525–1547.

Darden, M. and McCarthy, I. M. (2015). The star treatment: Estimating the impact of star ratings on medicare advantage enrollments. *Journal of Human Resources*, **50** (4), 980–1008.

Decarolis, F. and Guglielmo, A. (2017). Insurers response to selection risk: Evidence from medicare enrollment reforms. *Journal of Health Economics*, **forthcoming**.

—, — and Luscombe, C. (2017). Open enrollment periods and plan choices, national Bureau of Economic Research.

Duggan, M., Starc, A. and Vabson, B. (2016). Who benefits when the government pays more? pass-through in the medicare advantage program. *Journal of Public Economics*, **141**, 50–67.

Einav, L., Finkelstein, A., Kluender, R. and Schrimpf, P. (2016). Beyond statistics: the economic content of risk scores. *American Economic Journal: Applied Economics*, **8** (2), 195–224.

Geruso, M., Layton, T. J. and Prinz, D. (2016). Screening in contract design: Evidence from the ACA Health Insurance Exchanges. *Mimeo, National Bureau of Economic Research*.

Girotti, M. E., Ko, C. Y. and Dimick, J. B. (2013). Hospital morbidity rankings and complication severity in vascular surgery. *Journal of vascular surgery*, **57** (1), 158–164.

Glazer, J. and McGuire, T. G. (2006). Optimal quality reporting in markets for health plans. *Journal of Health Economics*, **25** (2), 295–310.

Harris Interactive (2011). Medicare star quality rating system study: Key findings. xnet.kp.org/newscenter/pressreleases/nat/2011/downloads/101011medicarerankingsHarrisSurveyInfo.pdf, accessed: 2018-03-22.

Kaiser Family Foundation (2017). The facts on Medicare spending and financing. https://www.kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/, accessed: 2018-03-30.

Layton, T. J. and Ryan, A. M. (2015). Higher incentive payments in Medicare Advantage's pay-for-performance program did not improve quality but did increase plan offerings. *Health services research*, **50** (6), 1810–1828.

McCarthy, I. M. and Darden, M. (2017). Supply-side responses to public quality ratings: Evidence from Medicare Advantage. *American Journal of Health Economics*.

MedPAC (2015). Medicare payment advisory commission, report to the congress: Medicare payment policy.

Newhouse, J. P. and McGuire, T. G. (2014). How successful is Medicare Advantage. *Milbank Quarterly*, **92** (2), 351–394.

—, Price, M., Huang, J., McWilliams, J. M. and Hsu, J. (2012). Steps to reduce favorable risk selection in Medicare Advantage largely succeeded, boding well for health insurance exchanges. *Health Affairs*, **31** (12), 2618–2628.

NQF (2014). Risk adjustment for socioeconomic status or other sociodemographic factors. *National Quality Forum*.

Song, Z., Landrum, M. B. and Chernew, M. E. (2012). Competitive bidding in medicare: who benefits from competition? *The American Journal of Managed Care*, **18** (9), 546.

—, — and — (2013). Competitive bidding in Medicare Advantage: Effect of benchmark changes on plan bids. *Journal of Health Economics*, **32** (6), 1301–1312.

Soria-Saucedo, R., Xu, P., Newsom, J., Cabral, H. and Kazis, L. E. (2016). The role of geography in the assessment of quality: Evidence from the medicare advantage program. *PloS one*, **11** (1), e0145656.

Appendix

A Tables

Table 3: Effect of QBP on the risk score

	(1)	(2)	(3)	(4)	(5)	(6)
	risk score (contract)		risk score (contract)		risk score (plan)	
$high \cdot post$	-0.026***	-0.041***	-0.035***	-0.042***	-0.020***	-0.045***
	(0.0082)	(0.014)	(0.012)	(0.015)	(0.0074)	(0.015)
weights	plan enrollment		equal weights		unweighted	
y mean	0.97	0.97	0.97	0.97	0.96	0.96
fixed effects	contract		contract		plan	
R^2	0.86	0.0068	0.76	0.012	0.79	0.0089
N	1,122	1,127	1,122	1,122	4,549	4,549

Notes: Table shows difference-in-difference estimates on risk score, aggregated at contract level using enrollment weights in column (1)-(2), using equal weights in column (3)-(4), and measured at the raw plan level in column (5)-(6). For each measure we show results with and without fixed effects. Standard errors clustered at the contract level (column 1-4) or plan level (column 5-6) in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 4: Effect of QBP on risk score, by service area risk and competition

	(1)	(2)	(3)	(4)
treat · post	-0.036***	-0.045**	-0.018	0.0094
	(0.011)	(0.019)	(0.012)	(0.018)
treat	low risk, high quality		high hhi, high quality	
control	high risk, low quality		low hhi, low quality	
sample	+/-median	15% tails	+/-median	15% tails
y mean	1.00	1.02	0.98	0.99
R^2	0.88	0.90	0.85	0.89
N	534	211	506	191

Notes: Table shows difference-in-difference estimates on risk score, aggregated at contract level using enrollment weights. In column (1), treated group is baseline high quality contracts with service area risk below the median, and the control group is baseline low quality contracts with service area risk above the median. In column (3), treated group is baseline high quality contracts with HHI below the median, and the control group is baseline low quality contracts with HHI above the median. Column (2) and (4) further limits the sample to 15% tails. Construction of baseline characteristics measures is described in the main text. All regressions include contract level fixed effects. Standard errors clustered at the contract level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 5: Effect of QBP on market characteristics

	(1)	(2)	(3)	(4)	(5)
	# counties	risk	benchmark	high-bonus county	# plans
$high \cdot post$	8.70	0.0024	1.80	-0.020	-0.17
	(8.39)	(0.0024)	(2.94)	(0.021)	(0.23)
y mean	25.09	0.99	799.15	0.72	3.40
R^2	0.73	0.98	0.96	0.90	0.87
N	1,122	1,122	1,122	1,122	1,122

Notes: Table shows difference-in-difference estimates on market size and characteristics. Outcome is at the contract-year level. Numbers of counties and plans are counted within contract-year. *risk* and *benchmark* are contract-year averages of 2012 characteristics over the market set, and hence reflect differential county entry or exit by these characteristics. If a county later on receives an above 8% bonus benchmark for 5-star contracts, it is assigned the high-bonus status. Column (4) looks at if contracts cover more of these counties after the reform. All regressions include contract level fixed effects. Standard errors clustered at the contract level in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 6: Effect of QBP on bidding and rebate

	(1)	(2)	(3)	(4)
	benchmark	bid	benchmark-bid	rebate
$high \cdot post$	27.84***	37.01***	-9.17	0.39
	(7.10)	(7.50)	(6.07)	(3.68)
y mean	874.10	763.38	110.72	78.37
R^2	0.83	0.84	0.83	0.87
N	1,122	1,122	1,122	1,122

Notes: Table shows difference-in-difference estimates on benchmark, bid and rebate. Outcome is at the contract-year level. We aggregate plan level benchmark (an enrollment-weighted average of county benchmarks, higher for higher quality contracts after QBP), bid, and rebate (inclusive of rebate bonus after QBP) to the contract level using enrollment weights. All regressions include contract level fixed effects. Standard errors clustered at the contract level in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 7: Effect of QBP on premium and drug deductible

	(1)	(2)	(3)	(4)
	premium	zero premium	drug deduc	zero deduc
$high \cdot post$	3.14	0.032	-16.98*	0.051
	(3.56)	(0.025)	(8.98)	(0.045)
y mean	49.07	0.41	32.62	0.84
R^2	0.91	0.88	0.69	0.63
N	1,122	1,122	1,122	1,122

Notes: Table shows difference-in-difference estimates on premium and drug deductible. Outcome is at the contract-year level. We aggregate plan level data to the contract level using enrollment weights. All regressions include contract level fixed effects. Standard errors clustered at the contract level in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 8: Effect of QBP on premium, within-contract cross-county variation

	(1)	(2)	(3)	(4)	(5)	(6)
$risk \cdot high \cdot post$			30.51**			35.61**
			(12.54)			(14.26)
$risk \cdot post$	-14.99*	20.99*	-14.75*	-13.49	21.64	-14.66*
	(8.73)	(11.20)	(8.66)	(8.87)	(13.37)	(8.60)
$high \cdot post$			-1.25			-1.16
			(4.86)			(4.83)
counties		all			15% tails	
sample	low	high	full	low	high	full
y mean	42.93	78.55	52.69	42.44	75.47	51.42
R^2	0.85	0.85	0.87	0.85	0.86	0.87
N	14,861	5,611	20,472	4,393	1,641	6,034

Notes: Table shows the within-contract premium variation across risk regions. Column 1-2 show the difference-in-difference estimates on premium across risk regions for baseline low quality (column 1) and high quality (column 2) contracts. Column 3 shows the triple-difference estimate, which gives the differential pricing response by high quality contracts in higher risk counties. Column 4-6 repeat the analysis, but only include counties in the lower and upper 15% of the risk distribution within a contract's market set. All regressions include contract-county fixed effects. Standard errors clustered two-way at the contract and county level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 9: Effect of QBP on drug deductible, within-contract cross-county variation

	(1)	(2)	(3)	(4)	(5)	(6)
$risk \cdot high \cdot post$			-7.06			-22.48
			(46.09)			(52.21)
$risk \cdot post$	36.54*	48.45	39.25**	33.62**	29.33	35.50**
	(18.78)	(46.03)	(19.46)	(16.87)	(51.01)	(17.26)
$high \cdot post$			-12.10			-13.42
			(10.02)			(9.52)
counties		all			15% tails	
sample	low	high	full	low	high	full
y mean	30.05	24.92	28.65	28.62	24.91	27.61
R^2	0.70	0.61	0.67	0.68	0.67	0.68
N	14,861	5,611	20,472	4,393	1,641	6,034

Notes: Table shows the within-contract variation in drug deductible across risk regions. Column 1-2 show the difference-in-difference estimates on zero-premium pricing across risk regions for baseline low quality (column 1) and high quality (column 2) contracts. Column 3 shows the triple-difference estimate, which gives the differential pricing response by high quality contracts in higher risk counties. Column 4-6 repeat the analysis, but only include counties in the lower and upper 15% of the risk distribution within a contract's market set. All regressions include contract-county fixed effects. Standard errors clustered two-way at the contract and county level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 10: Weight increase of outcome measures in quality rating

	(1)	(2)	(3)	(4)	(5)	(6)
	star rating		≥ 4.0 star		≥ 4.5 star	
$outcome \cdot post$	0.18***	0.24***	0.14***	0.19***	0.18***	0.20***
	(0.030)	(0.031)	(0.036)	(0.033)	(0.024)	(0.030)
y mean	3.59	3.59	0.35	0.35	0.16	0.16
$post$	≥ 2011	≥ 2012	≥ 2011	≥ 2012	≥ 2011	≥ 2012
R^2	0.77	0.78	0.57	0.58	0.57	0.57
N	1,080	1,080	1,080	1,080	1,080	1,080

Notes: Table shows the difference-in-difference estimate of average outcome measure rating (*outcome*) in the final star rating before and after the reform year in 2011 (odd columns) or in 2012 (even columns). All outcome measures averaged in *outcome* are present in the rating and measured consistently throughout 2009-2014. Starting 2012, these measures receive a 3.0 weight in the computation of the final rating. Column 1-2 estimates the effect of weight increase on the final star rating, which ranges from 1.5 star to 5.0 star at 0.5 star increments. Column 3-4 estimates the effect on the binary outcome of having at least 4.0 stars. Column 5-6 estimates the effect on having at least 4.5 stars. Unlike other difference-in-difference analysis in the paper, *outcome* is measured at the same period as outcome, and is not fixed at baseline (2009-2010) value. The purpose is to confirm the mechanic weight increase in the rating computation. All regressions include contract and year fixed effects. Standard errors clustered at the contract level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 11: Risk score and outcome rating

	(1)	(2)	(3)
	outcome mean	health improved	diabetes and blood pressure
$risk \cdot post$	-1.22**	-0.11	-1.37**
	(0.48)	(0.27)	(0.58)
y mean	3.45	3.28	3.60
R^2	0.63	0.22	0.69
N	997	888	991

Notes: Table shows the difference-in-difference estimates on outcome rating, across contracts with different enrollee risk scores in the baseline (2009-2010). Column 1 looks at the effect of baseline risk on the average rating of outcome measures. Column 2 and 3 divide the outcome measures by data source. Self-reported improvement in physical and mental health, collected from HOS, is the focus of column 2. HEDIS measures of having diabetic conditions and high blood pressure controlled are the focus of column 3. All regressions include contract and year fixed effects. Standard errors clustered at the contract level in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 12: Quality, risk and outcome rating

	(1)	(2)	(3)	(4)	(5)	(6)
	outcome mean		health improved		diabetes and blood pressure	
high · post	-0.26***	-0.44***	0.026	0.043	-0.42***	-0.62***
	(0.074)	(0.10)	(0.044)	(0.070)	(0.098)	(0.13)
y mean	3.45	3.38	3.28	3.30	3.59	3.49
treated	high quality (+ high risk)		high quality (+ high risk)		high quality (+ high risk)	
control	low quality (+ low risk)		low quality (+ low risk)		low quality (+ low risk)	
R^2	0.63	0.65	0.23	0.20	0.70	0.71
N	1,089	525	952	456	1,083	522

Notes: Table shows the difference-in-difference estimates on outcome rating, across contracts with different baseline quality in odd columns, and further across baseline enrollee risk scores in even columns. Specifically, in even columns, treated contracts are baseline high quality with enrollee risk score higher than median (0.97), and control contracts are baseline low quality with risk score below the median. Column 1-2 look at the quality and risk differential on average outcome ratings. Column 3-4 look at the effect on self-reported improvement in physical and mental health collected from HOS. Column 5-6 look at the effect on diabetes and high blood pressure control measures from HEDIS. All regressions include contract and year fixed effects. Standard errors clustered at the contract level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 13: Risk-outcome correlation across periods

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$riskscore_{t-3}$	0.90	1.56*	0.97									
	(1.14)	(0.90)	(1.20)									
$riskscore_{t-2}$				0.39	-1.07	-3.34**						
				(0.75)	(1.09)	(1.44)						
$riskscore_{t-1}$							1.30*	-0.67	0.68			
							(0.69)	(0.73)	(1.08)			
$riskscore_t$										1.53***	-0.47	-1.54
										(0.53)	(0.71)	(1.32)
baseline star	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5
R^2	0.67	0.85	0.66	0.65	0.84	0.84	0.63	0.78	0.80	0.63	0.75	0.74
N	336	146	46	472	210	70	611	269	91	760	340	126

Notes: Table shows OLS-estimated correlation between outcome rating in diabetes and blood pressure management and enrollee risk score across multiple periods. Contemporaneous correlation occurs between year t outcome rating and year $t-2$ risk score. Correlation with lag risk score in $t-3$ and lead risk score up to year t is also examined. For each time pair, table shows separate correlation for baseline low quality (3.0-3.5 stars), high quality (4.0 stars and above) and very high quality (4.5 stars and above) contracts. All regressions include contract and year fixed effects. Standard errors clustered at the contract level in the parenthesis.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

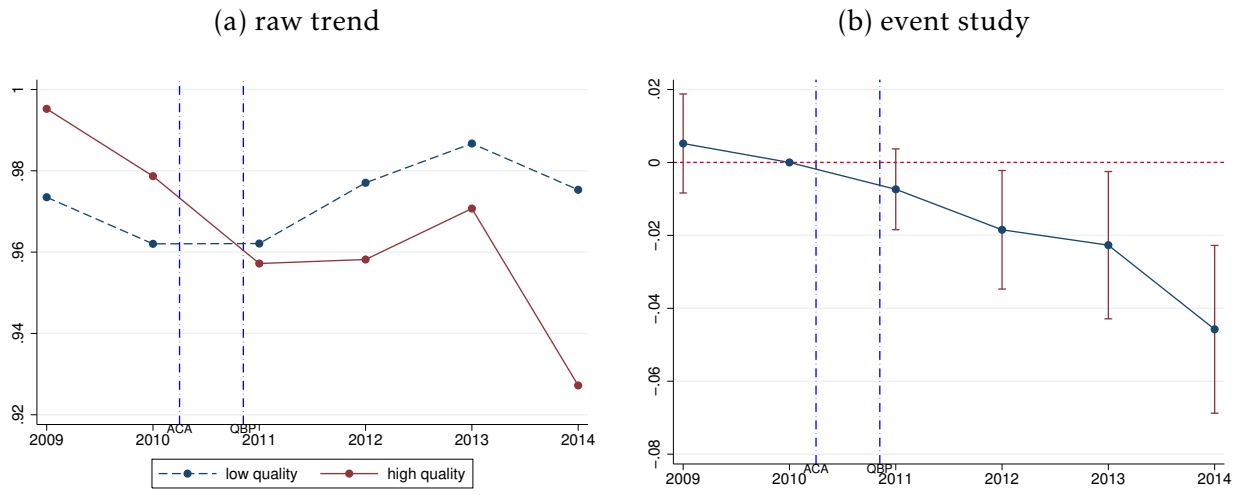
Table 14: Risk-quality correlation across periods

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$riskscore_{t-3}$	0.94*	1.15	1.72									
	(0.55)	(0.81)	(1.61)									
$riskscore_{t-2}$				-0.18	0.28	-2.76***						
				(0.49)	(1.00)	(0.80)						
$riskscore_{t-1}$							0.49	1.54**	-1.23*			
							(0.37)	(0.59)	(0.60)			
$riskscore_t$										0.44	0.88	-1.06
										(0.29)	(0.54)	(0.77)
baseline star	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5	3.0-3.5	≥4.0	≥4.5
R^2	0.73	0.66	0.14	0.67	0.64	0.71	0.65	0.62	0.73	0.66	0.57	0.66
N	337	136	38	479	202	64	618	259	82	792	338	118

Notes: Table shows OLS-estimated correlation between final star rating and enrollee risk score across multiple periods. Contemporaneous correlation occurs between year t star rating and year $t-2$ risk score. Correlation with lag risk score in $t-3$ and lead risk score up to year t is also examined. For each time pair, table shows separate correlation for baseline low quality (3.0-3.5 stars), high quality (4.0 stars and above) and very high quality (4.5 stars and above) contracts. All regressions include contract and year fixed effects. Standard errors clustered at the contract level in the parenthesis.
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

B Figures

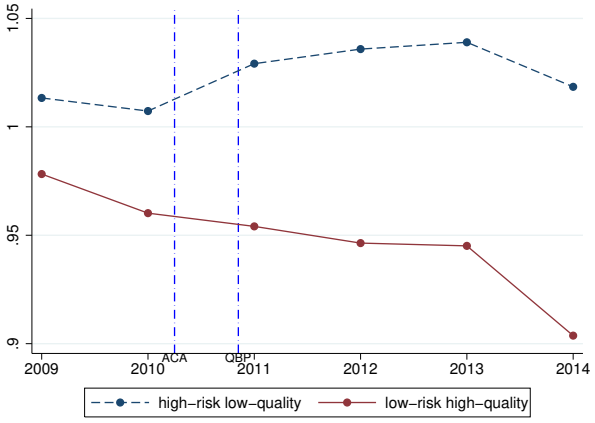
Figure 1: Effect on risk score, event study



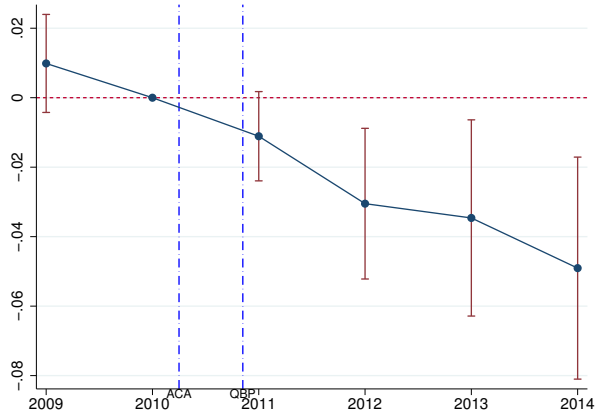
Notes. The left panel shows the raw trend of risk score for baseline high and low quality contracts. The right panel shows the event study estimates of the difference-in-difference model, controlling for contract and year fixed effects, with 95% confidence intervals based on robust standard errors clustered at the level of contract. Risk score is aggregated at the contract level weighted by plan enrollment.

Figure 2: Effect on risk score, by service area risk, event study

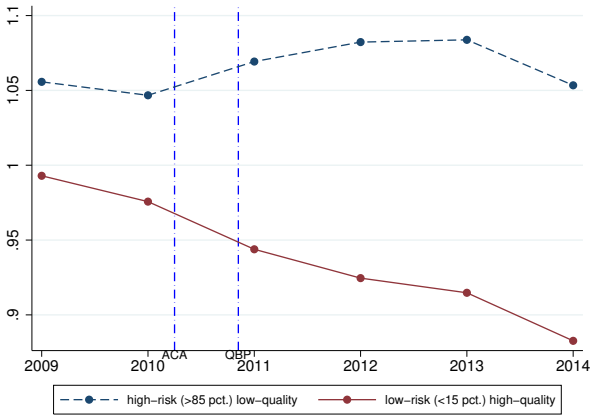
(a) raw trend



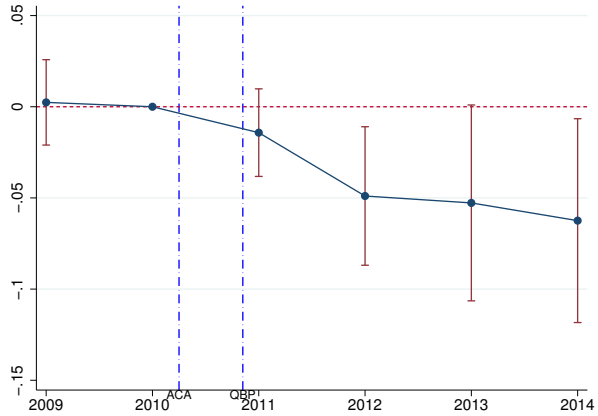
(b) event study



(c) raw trend, 15% tails

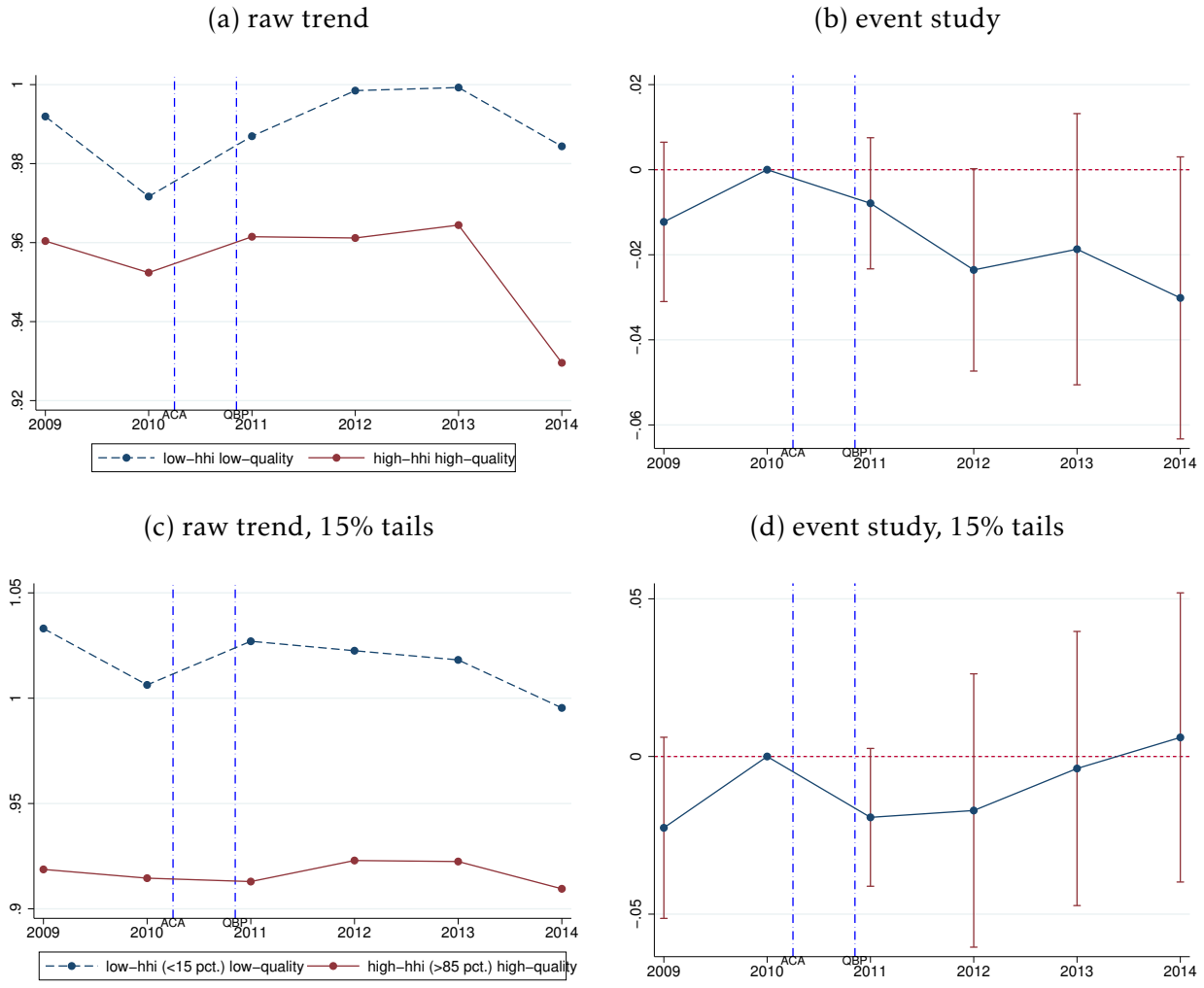


(d) event study, 15% tails



Notes. Panel (a) shows raw trend of contract-level risk score, for high quality contracts with below median service area risk and low quality contracts with above median service area risks. Panel (b) shows the event study estimates in 95% confidence intervals based on robust standard error clustered at the level of contracts. Corresponding raw trend and event study estimates for the 15% tails are in Panel (c) and (d), respectively.

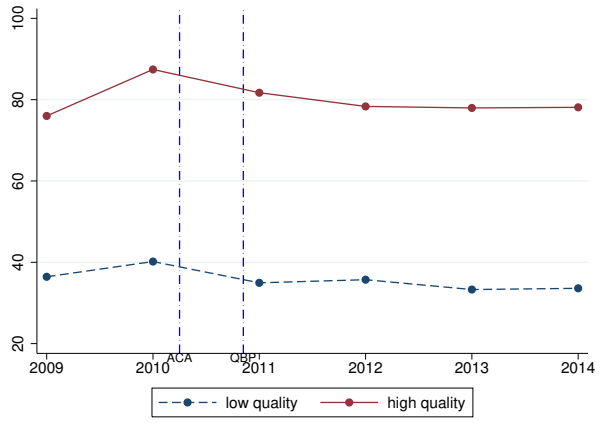
Figure 3: Effect on risk score, by market competition, event study



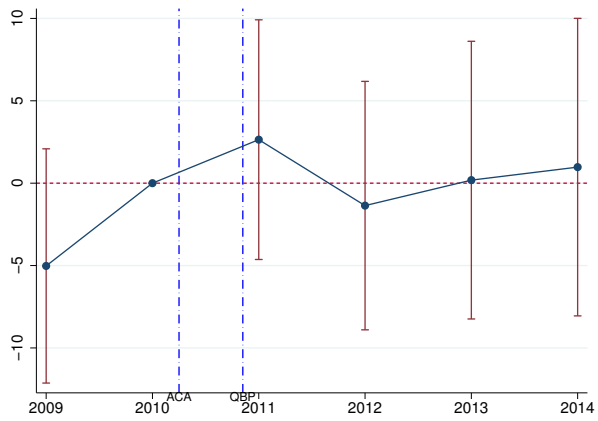
Notes. Panel (a) shows raw trend of contract-level risk score, for high quality contracts with below median HHI and low quality contracts with above median HHI. Panel (b) shows the event study estimates in 95% confidence intervals based on robust standard error clustered at the level of contracts. Corresponding raw trend and event study estimates for the 15% tails are in Panel (c) and (d), respectively.

Figure 4: Effect on premium and drug deductible, event study

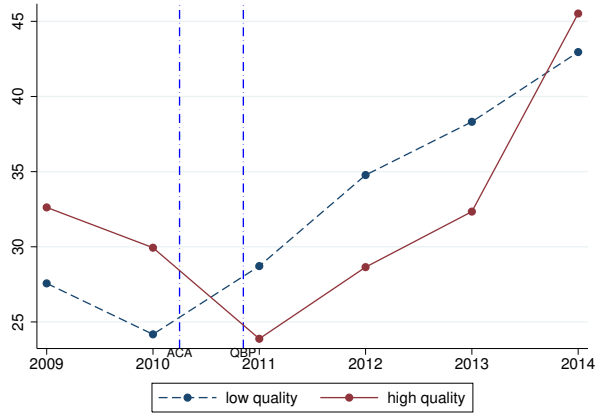
(a) raw trend, premium



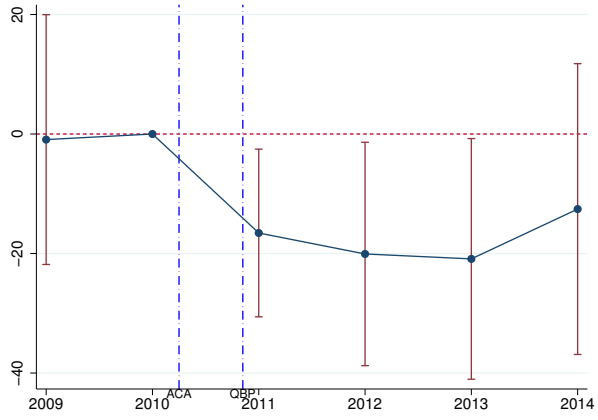
(b) event study, premium



(c) raw trend, drug deductible



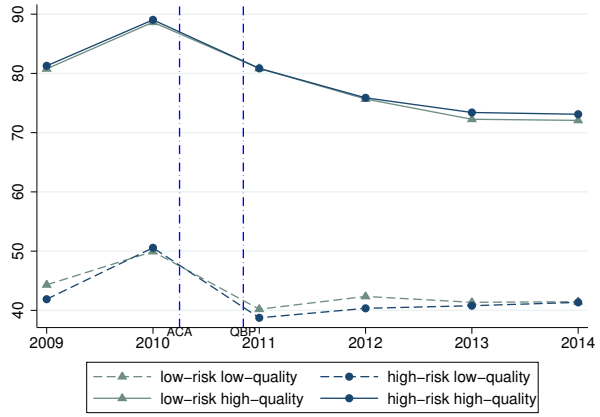
(d) event study, drug deductible



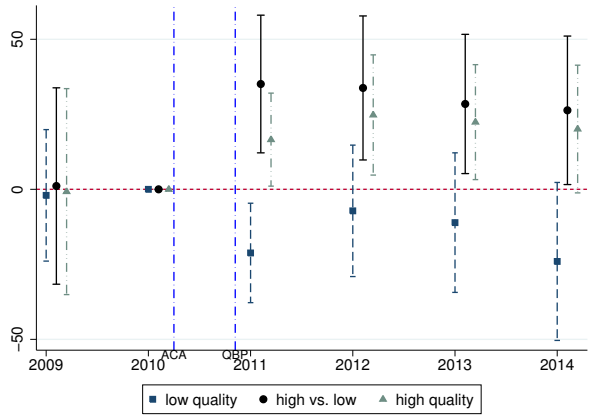
Notes. Panel (a) shows the raw trend of average premium at the contract level. Panel (b) shows the event study estimates for premium with 95% confidence intervals based on robust standard error clustered at the level of contracts. Panel (c) and (d) show the raw trend and event study estimates for average drug deductible at the contract level.

Figure 5: Effect on premium, within-contract cross county variation, event study

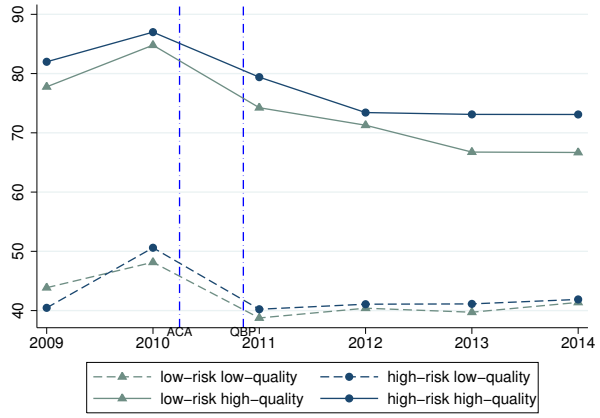
(a) raw trend



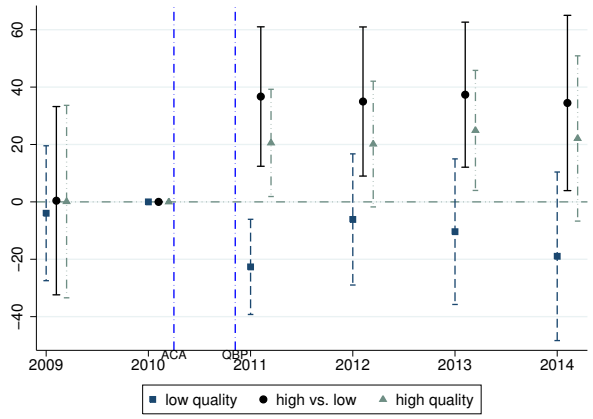
(b) event study



(c) raw trend, 15% tails

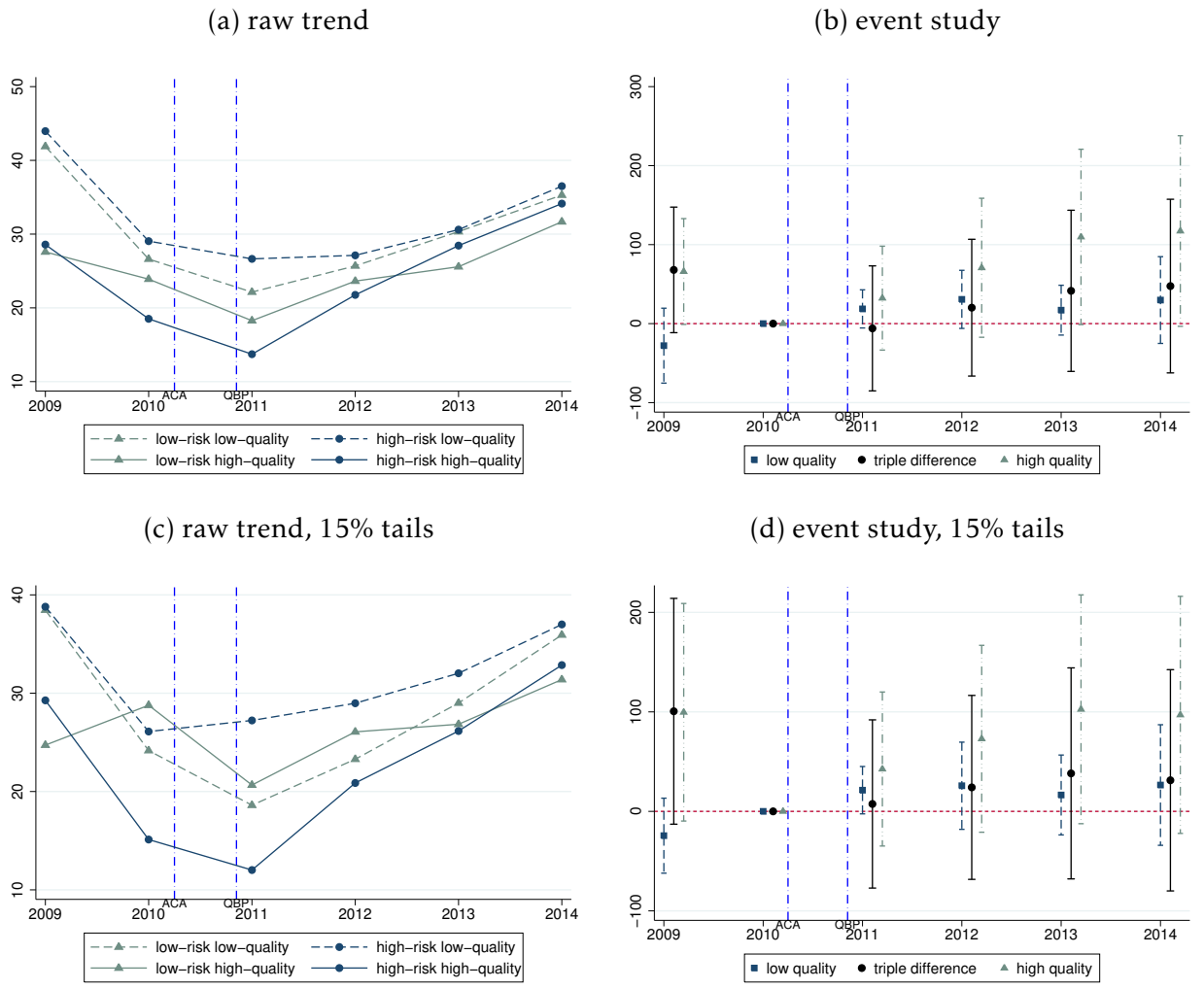


(d) event study, 15% tails



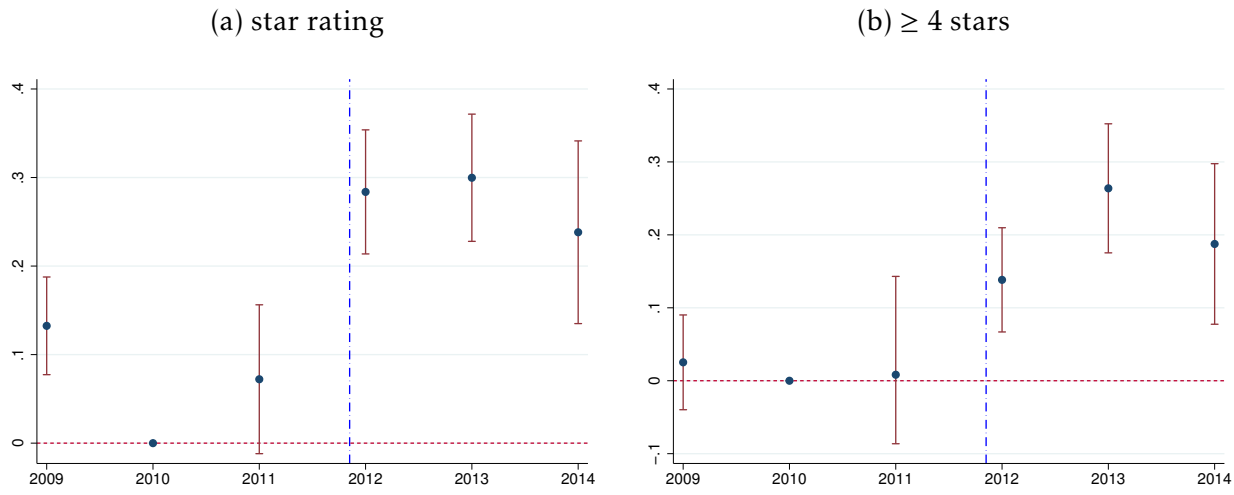
Notes. Panel (a) shows the raw trend of premium for high and low quality contracts, and their respective high and low risk regions. A high risk region given contract has baseline FFS risk score above the median in the market set. Panel (b) shows the event study estimates for low quality contracts (left line), high quality contracts (right line), and the differential effect on high quality contracts (middle line). Plotted 95% confidence intervals are based on robust standard errors clustered two-way at the level of county and contract. Panel (c) and (d) show the corresponding raw trend and event study estimates, limiting counties to those in the lower or upper 15% of the risk distribution given contract.

Figure 6: Effect on drug deductible, within-contract cross county variation, event study



Notes. Panel (a) shows the raw trend of drug deductible for high and low quality contracts, and their respective high and low risk regions. A high risk region given contract has baseline FFS risk score above the median in the market set. Panel (b) shows the event study estimates for low quality contracts (left line), high quality contracts (right line), and the differential effect on high quality contracts (middle line). Plotted 95% confidence intervals are based on robust standard errors clustered two-way at the level of county and contract. Panel (c) and (d) show the corresponding raw trend and event study estimates, limiting counties to those in the lower or upper 15% of the risk distribution given contract.

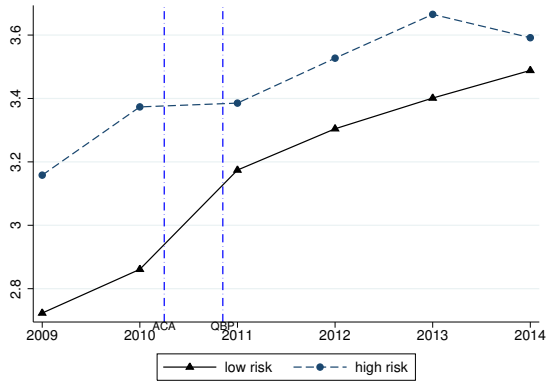
Figure 7: Weight increase of outcome measures in quality rating



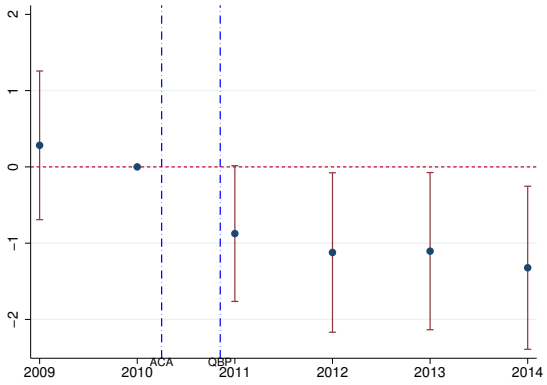
Notes. The figures show the event study trends of outcome measure weights in the quality star rating. The left panel shows that outcome ratings receive higher weights after 2012. The right panel shows a similar weight increase in outcome ratings for achieving at least a 4.0 star final rating. 95% confidence intervals are plotted based on robust standard errors clustered at the contract level.

Figure 8: Risk score and outcome rating, event study

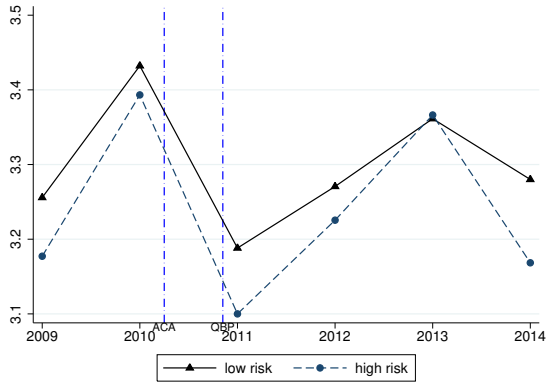
(a) average outcome rating, raw trend



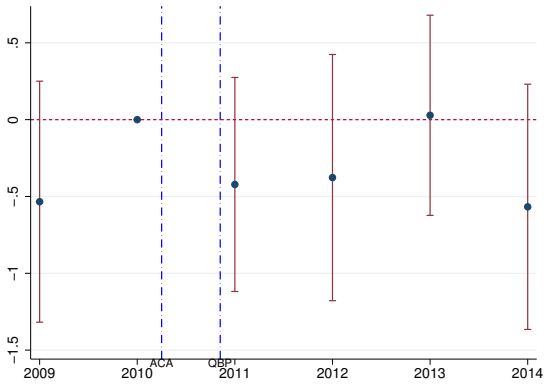
(b) average outcome rating, event study



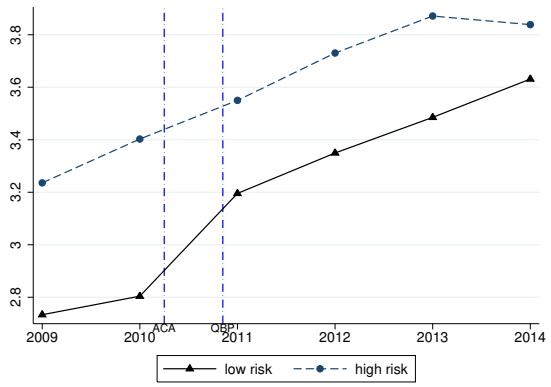
(c) health improved, raw trend



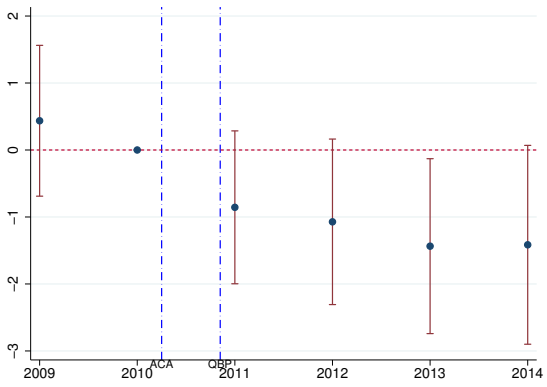
(d) health improved, event study



(e) diabetes and blood pressure, raw trend



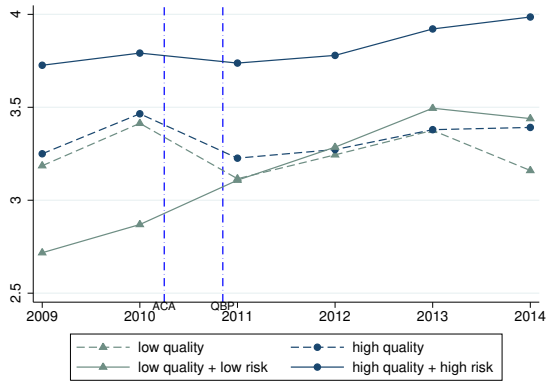
(f) diabetes and blood pressure, event study



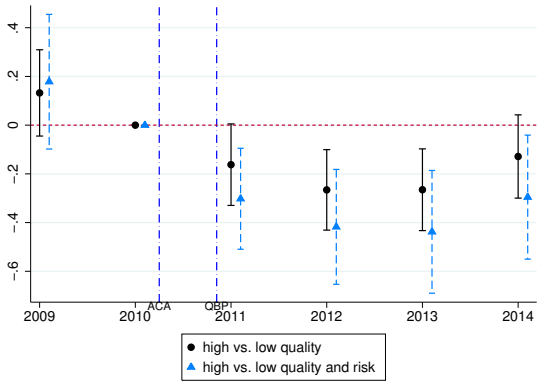
Notes. Figure shows the variation of outcome rating by baseline enrollee risk. In the raw trends (left panels), a high risk contract has baseline (2009-2010) enrollee risk score above the sample median (0.97). Right panels show the event study estimates from difference-in-difference specifications using continuous variation in baseline enrollee risk score. Panel (a) and (b) show the trending of average outcome measure ratings by risk. Panel (c) and (d) show the trending of health improvement measures reported in HOS. Panel (e) and (f) show the trending of diabetes and blood pressure control from HEDIS clinical measures. Event study graphs show 95% confidence intervals based on standard errors clustered at the level of contract.

Figure 9: Quality, risk and outcome rating, event study

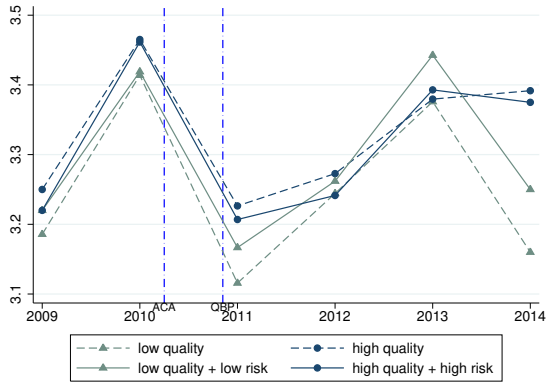
(a) average outcome rating, raw trend



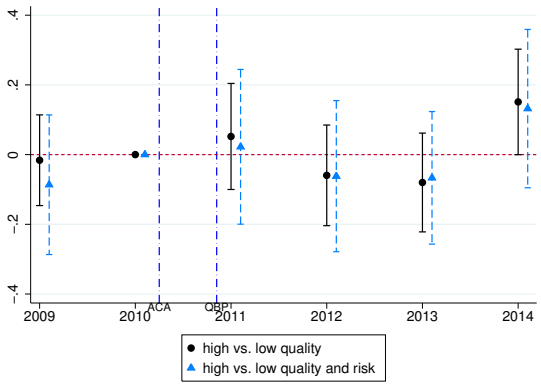
(b) average outcome rating, event study



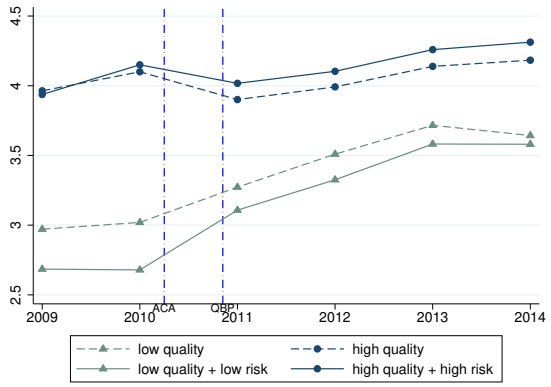
(c) health improved, raw trend



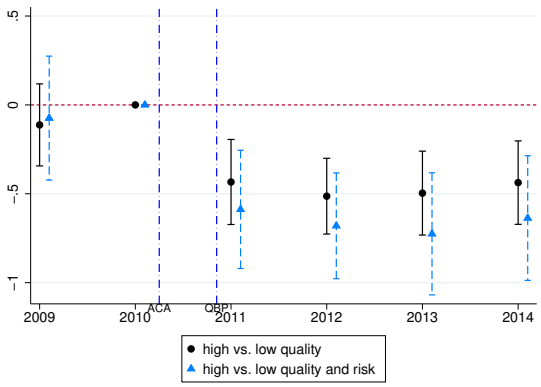
(d) health improved, event study



(e) diabetes and blood pressure, raw trend



(f) diabetes and blood pressure, event study



Notes. Figure shows the variation of outcome rating by baseline quality and enrollee risk. A high risk contract has baseline (2009-2010) enrollee risk score above the sample median (0.97). Left panels show the raw trend of outcome ratings and component ratings for baseline high vs.low contracts in dotted lines, and for baseline high-risk high-quality vs. low-risk low-quality contracts in solid lines. Right panels show the event study estimates from corresponding difference-in-difference specifications. Panel (a) and (b) show the trending of average outcome measure ratings by quality and risk. Panel (c) and (d) show the trending of health improvement measures reported in HOS. Panel (e) and (f) show the trending of diabetes and blood pressure control from HEDIS clinical measures. Event study graphs show 95% confidence intervals based on standard errors clustered at the level of contract.

C Additional Appendix Tables

Table A1: Part C measures in the quality rating, 2013

ID	name	category	weight	source	time frame
Domain 1: Staying Healthy: Screenings, Tests and Vaccines					
C01	Breast Cancer Screening	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C02	Colorectal Cancer Screening	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C03	Cardiovascular Care – Cholesterol Screening	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C04	Diabetes Care – Cholesterol Screening	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C05	Glaucoma Testing	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C06	Annual Flu Vaccine	Process Measure	1	CAHPS	02/15/2012 - 05/31/2012
C07	Improving or Maintaining Physical Health	Outcome Measure	3	HOS	04/18/2011 - 07/31/2011
C08	Improving or Maintaining Mental Health	Outcome Measure	3	HOS	04/18/2011 - 07/31/2011
C09	Monitoring Physical Activity	Process Measure	1	HOS/HEDIS	04/18/2011 - 07/31/2011
C10	Adult BMI Assessment	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
Domain 2: Managing Chronic (Long Term) Conditions					
C11	Care for Older Adults – Medication Review	Process Measure	1	HEDIS	01/01/2011 - 12/31/2011
C12	Care for Older Adults – Functional Status Assessment	Process Measure	1	HEDIS	01/01/2011 - 12/31/2012
C13	Care for Older Adults – Pain Screening	Process Measure	1	HEDIS	01/01/2011 - 12/31/2013
C14	Osteoporosis Management in Women who had a Fracture	Process Measure	1	HEDIS	01/01/2011 - 12/31/2014
C15	Diabetes Care – Eye Exam	Process Measure	1	HEDIS	01/01/2011 - 12/31/2015
C16	Diabetes Care – Kidney Disease Monitoring	Process Measure	1	HEDIS	01/01/2011 - 12/31/2016
C17	Diabetes Care – Blood Sugar Controlled	Intermediate Outcome Measures	3	HEDIS	01/01/2011 - 12/31/2017
C18	Diabetes Care – Cholesterol Controlled	Intermediate Outcome Measures	3	HEDIS	01/01/2011 - 12/31/2018
C19	Controlling Blood Pressure	Intermediate Outcome Measures	3	HEDIS	01/01/2011 - 12/31/2019
C20	Rheumatoid Arthritis Management	Process Measure	1	HEDIS	01/01/2011 - 12/31/2020
C21	Improving Bladder Control	Process Measure	1	HOS/HEDIS	04/18/2011 - 07/31/2011
C22	Reducing the Risk of Falling	Process Measure	1	HOS/HEDIS	04/18/2011 - 07/31/2011
C23	Plan All-Cause Readmissions	Outcome Measure	3	HEDIS	01/01/2011 - 12/31/2020
Domain 3: Member Experience with Health Plan					
C24	Getting Needed Care	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
C25	Getting Appointments and Care Quickly	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
C26	Customer Service	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
C27	Overall Rating of Health Care Quality	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
C28	Overall Rating of Plan	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
C29	Care Coordination	Patients' Experience and Complaints Measure	1	CAHPS	02/15/2012 - 05/31/2012
Domain 4: Member Complaints, Problems Getting Services, and Improvement in the Health Plan's Performance					
C30	Complaints about the Health Plan	Patients' Experience and Complaints Measure	1.5	CTM	01/01/2012 - 06/30/2012
C31	Beneficiary Access and Performance Problems	Measures Capturing Access	1.5	CMS	01/01/2011 - 12/31/2011
C32	Members Choosing to Leave the Plan	Patients' Experience and Complaints Measure	1.5	MBDSS	01/01/2011 - 12/31/2011
C33	Health Plan Quality Improvement	Outcome Measure	1	CMS	2012 rating
Domain 5: Health Plan Customer Service					
C34	Plan Makes Timely Decisions about Appeals	Measures Capturing Access	1.5	IRE	01/01/2011 - 12/31/2011
C35	Reviewing Appeals Decisions	Measures Capturing Access	1.5	IRE	01/01/2011 - 12/31/2011
C36	Call Center – Foreign Language Interpreter and TTY/TDD Availability	Measures Capturing Access	1.5	Call Center	01/30/2012 - 05/18/2012
C37	Enrollment Timeliness	Process Measure	1	MARx	01/01/2012 - 06/30/2012

Notes: Table lists the name of Part C measures in the 2013 quality rating, with detailed information on the data source of the measure, and the relevant measurement period in the source. Weight attached to each measure in the final rating is also listed.

Table A2: Part D measures in the quality rating, 2013

ID	name	category	weight	source	time frame
Domain 1: Drug Plan Customer Service					
D01	Call Center – Pharmacy Hold Time	Measures Capturing Access	1.5	Call Center	02/06/2012 - 05/18/2012
D02	Call Center – Foreign Language Interpreter and TTY/TDD Availability	Measures Capturing Access	1.5	Call Center	01/30/2012 - 05/18/2012
D03	Appeals Auto–Forward	Measures Capturing Access	1.5	IRE	01/01/2011 - 12/31/2011
D04	Appeals Upheld	Measures Capturing Access	1.5	IRE	01/01/2012 - 6/30/2012
D05	Enrollment Timeliness	Process Measure	1	MARx	01/01/2012 - 06/30/2012
Domain 2: Member Complaints, Problems Getting Services, and Improvement in the Drug Plan's Performance (identical to part C domain 4; redundant and not used in the final rating)					
D06	Complaints about the Drug Plan	Patients' Experience and Complaints Measure	1.5	CTM	01/01/2012 - 06/30/2012
D07	Beneficiary Access and Performance Problems	Measures Capturing Access	1.5	CMS	01/01/2011 - 12/31/2011
D08	Members Choosing to Leave the Plan	Patients' Experience and Complaints Measure	1.5	MBDSS	01/01/2011 - 12/31/2011
D09	Drug Plan Quality Improvement	Outcome Measure	1	CMS	2012 rating
Domain 3: Member Experience with the Drug Plan					
D10	Getting Information From Drug Plan	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
D11	Rating of Drug Plan	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
D12	Getting Needed Prescription Drugs	Patients' Experience and Complaints Measure	1.5	CAHPS	02/15/2012 - 05/31/2012
Domain 4: Member Experience with the Drug Plan					
D13	MPF Price Accuracy	Process Measure	1	PDE	01/01/2011 - 09/30/2011
D14	High Risk Medication	Intermediate Outcome Measures	3	PDE	01/01/2011 - 12/31/2011
D15	Diabetes Treatment	Intermediate Outcome Measures	3	PDE	01/01/2011 - 12/31/2011
D16	Part D Medication Adherence for Oral Diabetes Medications	Intermediate Outcome Measures	3	PDE	01/01/2011 - 12/31/2011
D17	Part D Medication Adherence for Hypertension (RAS antagonists)	Intermediate Outcome Measures	3	PDE	01/01/2011 - 12/31/2011
D18	Part D Medication Adherence for Cholesterol (Statins)	Intermediate Outcome Measures	3	PDE	01/01/2011 - 12/31/2011

Notes: Table lists the name of Part D measures in the 2013 quality rating, with detailed information on the data source of the measure, and the relevant measurement period in the source. Weight attached to each measure in the final rating is also listed. Measure D06-D09 in the part D domain "Member Complaints, Problems Getting Services, and Improvement in the Drug Plan's Performance" are identical to measure C30-C33 in the corresponding Part C domain; only C30-C33 are kept for computing the final rating.