Performance Pay in Insurance Markets: Evidence from Medicare

Michele Fioretti, Hongming Wang

SciencesPo
DEPARTMENT OF ECONOMICS

Discussion Paper

PERFORMANCE PAY IN INSURANCE MARKETS: EVIDENCE FROM MEDICARE

Michele Fioretti and Hongming Wang

SCIENCES PO ECONOMICS DISCUSSION PAPER

No. 2021-02

Performance Pay in Insurance Markets: Evidence from Medicare*

Michele Fioretti† **Hongming Wang**‡

May 31, 2021

Abstract

Public procurement bodies increasingly resort to pay-for-performance contracts to promote efficient spending. We show that firm responses to pay-for-performance can widen the inequality in accessing social services. Focusing on the quality bonus payment initiative in Medicare Advantage, we find that higher quality-rated insurers responded to bonus payments by selecting healthier enrollees with premium differences across counties. Selection is profitable because the quality rating fails to adjust for differences in enrollee health. Selection inflated the bonus payments and shifted the supply of high-rated insurance to the healthiest counties, reducing access to lower-priced, higher-rated insurance in the riskiest counties.

JEL classifications: I13, I14, L15
Keywords: pay-for-performance, Medicare Advantage, risk selection, quality ratings, health insurance access

*We would like to thank Ghazala Azmat, Aaron Baum, Zach Brown, Moshe Buchinsky, Thomas Chaney, Alice Chen, Francesco Decarolis, Golvine De Rochembau, Liran Einav, Randall Ellis, Emeric Henry, Aljoscha Janssen, Bora Kim, Marleen Marra, Daria Pelech, Carol Propper, Geert Ridder, Alejandro Robinson-Cortés, Mark Shepard, Andre Veiga, Gianluca Violante and participants of the 2018 ASHECON, the 2019 IIOC, the 2019 APIOC, the 2020 Econometric Society World Congress, the 2020 INFER, the 2021 AEA meetings, and at seminars at the University of Southern California, Sciences Po, Bar-Ilan University, and Keio University for helpful comments and discussions.

†Department of Economics, Sciences Po. email: michele.fioretti@sciencespo.fr
‡Center for Global Economic Systems, Hitotsubashi University. email: hongming.wang@r.hit-u.ac.jp

1 Introduction

Market-based approaches are increasingly popular means to reduce inefficiencies in the provision of public goods. One of them, the pay-for-performance model, is found in a range of settings, from government agencies (Burgess *et al.*, 2017) to education (Biasi, 2018) and tax collection (Khan *et al.*, 2015). In pay-for-performance, firms receive a quality rating of their services, and payments are directly linked to the quality rating. In principle, financial incentives can spur firms to invest in service quality. In reality, however, pay-for-performance can direct resources away from investments if the design of the quality rating is badly aligned with the quality initiative.

The design of the quality rating is especially critical in selection markets like the insurance market. Here, service quality depends directly on the match between the needs, or type, of consumers and the service offered (Veiga and Weyl, 2016). As a result, pay-for-performance can create additional incentives to screen consumers if servicing certain consumer types worsens the quality rating. The selection response can distort the quality rating with potentially adverse effects on consumers. In health insurance markets, for example, selecting on enrollee characteristics like pre-existing conditions or ethnicity (Bauhoff, 2012) can reduce access to care for those who need it the most, ultimately widening health inequality (e.g., Chetty *et al.*, 2016, Currie and Schwandt, 2016). However, we know little about the ways insurers internalize pay-for-performance, or the effect of insurers' responses on quality ratings, payments, and enrollees.

This paper examines how insurers respond to pay-for-performance by exploiting the introduction of the quality bonus payment initiative in the U.S. Medicare Advantage market, where Medicare services are provided by private insurers who receive subsidies from the government.[1] Under pay-for-performance, government subsidies depend on insurance quality through a quality rating that was already available to prospective

[1] Medicare provides near-universal health insurance to Americans over the age of 65. The program costed the U.S. government $750 bn in 2018, or 20.8% of total health expenditure (CMS, 2018). Around one-third of Medicare enrollees receive services from private insurers through the Medicare Advantage program.

enrollees before the reform. Since the reform shifted insurer payments without affecting consumers' knowledge of the quality rating, we exploit the reform to understand the incentive effects on insurers and the resulting impacts on consumers' access to insurance.

We find that insurers with high-quality ratings before the payment reform served less risky enrollees after the reform. These insurance contracts lowered premiums in healthier, low-risk counties and simultaneously raised premiums in riskier ones to select healthier enrollees. Risk selection is profitable because the quality rating relies heavily on health outcome measures, which are not adjusted for enrollees' health conditions. In response, selecting insurers inflated the quality rating by avoiding enrollees with more complicated conditions. Due to selection, the supply of lower-priced, higher-rated insurance shifted to the healthiest counties, reducing access for consumers in the riskiest counties.

We motivate our empirical analysis using a stylized model of insurer pricing. The model predicts that a biased quality rating induces insurers to select healthier enrollees, and the selection incentive increases with bonus payments. Since the payment reform significantly increased the bonus payments to higher-rated insurers, we distinguish insurance contracts by their pre-reform quality ratings and examine the responses of high-rated contracts to the payment reform in a difference-in-differences framework.

Empirically, we document shifts in the distribution of risk scores to the lower percentiles after the payment reform in high-rated insurance, but not in low-rated insurance. Consistent with the model predictions, risk scores of high-rated contracts serving healthier counties before the payment reform decreased even more – in these "high-selection" contracts, risk scores dropped by 4 percentage points. We then ask *how* insurers selected healthier enrollees and *why*.

To address how selection happened, we examine the pricing strategy of insurers across counties. We find that prescription drug coverage premiums increased substantially with county risk scores in high-rated contracts, but not in low-rated contracts. We rule out local socio-economic factors, market concentration, provider cost and quality as drivers of the

premium differences, and show evidence that premiums responded directly to the health of enrollee across counties. Thus, consistent with our model's predictions, high-rated contracts selected healthier enrollees by varying premiums across counties.

To understand why the payment reform incentivized the selection of healthier individuals, we inspect sub-measures of quality exploiting the weights they receive in the final rating linked to payments. For high-selection contracts, around 50% of the quality rating is determined by the health outcome measures. These measures rank contracts based on improvements in chronic conditions over time but fail to adjust for differences in health conditions at the time of enrollment. As such, these measures are sensitive to the risk types of enrollees. We find that healthier enrollees are associated with better outcome ratings, and contracts with greater improvements in the risk pool also experienced greater relative gains in the outcome rating. These results are consistent with insurers selecting healthier enrollees to inflate the quality rating and bonus payments.

We quantify the effect of selection on the quality rating and payments using an instrumental variable strategy. Based on our finding that insurers selected enrollees through premiums, we instrument the risk composition of contracts by the premium differences across counties. We use the IV estimates to calculate rating gains due to the selection of enrollees, and infer actual quality improvements by removing the selection gains from the quality rating. We find that risk selection explained nearly 80% of the health rating gains in high-selection contracts, inflating the overall rating by 0.5 to 1 star (out of 5 stars). As a result, the star rating became less informative for consumers and bonus payments increased by 14% for high-selection contracts.

The selection response has sizeable distributional impacts on enrollees. Since average premiums and enrollee benefits did not differ by quality ratings, premium differences to select healthier enrollees shifted insurance benefits from the sickest to the healthiest enrollees in high-rated insurance. To quantify this shift, the market share of high-rated insurance increased by more than 17% in the healthiest counties compared to the riskiest

ones after the policy. As the supply of high-rated insurance shifted towards the healthiest counties in the North West and the South West, access to low-priced, high-rated insurance worsened in the riskier counties in the South.

Several aspects of the quality rating contributed to the selection responses. First, the current rating measures health improvements relative to a uniform threshold for all risk types. Adjusting the threshold by the expected outcomes of risk types compensates insurers for enrolling riskier individuals, thereby reducing the selection incentive. Moreover, because health outcomes are averaged across enrollees in multiple counties, a stratified risk adjustment based on the risk in the serviced areas can further reduce the selection incentives. The adjustment would also result in more informative star ratings for consumers.

Relation to the Literature. This paper is related to a large literature on pay-for-performance. Our key findings are consistent with the theoretical insight that payment incentives based on biased measures of performance distort efforts (Holmstrom and Milgrom 1991, Baker 1992). Applied to healthcare, the distortions are heightened when multiple procedures are rewarded, in which case standard payment methods such as capitation may be sub-optimal (Sherry 2016, Eggleston 2005). Empirically, pay-for-performance has modest impacts on provider behavior (Rosenthal and Frank 2006, Mullen *et al.* 2010) and can result in patient selection (Shen 2003) and strategic reporting (Gravelle *et al.* 2010) for outcome-based performance measures. We add to this literature by providing the first evidence on insurers' responses to pay-for-performance and the distortions on prices and the quality rating.

Our results also inform the literature on risk adjustment, without which insurers would have financial incentives to enroll the healthy and avoid the sick. The goal of risk adjustment is to explain the predictable portion of an enrollee's health cost variation (Ellis and McGuire, 2007). Despite improvements in prediction models (Van De Ven and Ellis, 2000), adjustments based on statistical prediction of service costs can lead to inefficient

care provision and selection (Glazer and McGuire 2000, Newhouse *et al.* 2015).[2] Moreover, prediction errors distort benefit design and shift selection to enrollees cheaper than their predicted costs (e.g., Brown *et al.* 2014, Carey 2017, Lavetti and Simon 2018, Geruso *et al.* 2019). Adding to this literature, we document selection through premiums across geographies in response to inadequate adjustments of pay-for-performance measures, even though revenues are indeed risk-adjusted. We further show that selection shifted the spatial distribution of insurance, hurting in particular consumers in the riskiest counties.[3]

This paper also contributes to an emerging literature on the value-based initiatives of the ACA. Layton and Ryan (2015) find that the quality of MA insurance did not improve in counties with larger benchmark bonuses. Relatedly, Abaluck *et al.* (2020) find that the mortality benefits of MA contracts are not correlated with the star rating published by CMS.[4] In the hospital setting, penalties applied to low-quality hospitals improved re-admission rates but also induced patient selection (Gupta, 2021). These findings suggest that pay-for-performance can positively impact health, but may be less effective in markets where selection can substantially improve ratings.

The remainder of the paper is organized as follows. Section 2 introduces the Quality Bonus Payment demonstration and a conceptual framework of insurer behavior. Following this framework, we examine the effects of bonus payments on risk scores in Section 3 and the pricing responses across counties in Section 4. We inspect the rating design as the source of the selection incentive in Section 5 and the distributional impacts across counties in Section 6. Section 7 discusses the results and concludes.

[2]Risk adjustment has improved substantially since Newhouse *et al.* (1997) discussed the need to condition adjustments on diagnoses (see also Breyer *et al.*, 2011). Recent results show that health-based risk adjustments could also improve market stability in ACA Exchanges by reducing the adverse selection in consumer sorting (Handel *et al.* 2015, Layton 2017).

[3]Our findings suggest that insurer responses to payment incentives could contribute to the disparities in healthcare spending, prices, and health outcomes in the US (e.g., Skinner 2011, Cooper *et al.* 2018, Finkelstein *et al.* 2016, 2019).

[4]Consistent with the star rating being a noisy signal of insurance quality, the rating has only modest impacts on consumer welfare (Charbi, 2020) or enrollment (Darden and McCarthy, 2015).

2 Quality Ratings and Payments in Medicare Advantage

Medicare provides near-universal health insurance to the elderly population (65+) in the US. Enrollees choose between Traditional Medicare, also known as Fee-For-Service Medicare (FFS), and private Medicare insurance from the Medicare Advantage (MA) market. MA plans provide additional benefits over FFS, for which enrollees are charged the "Part C premium." Most plans also provide prescription drug coverage, which results in the "Part D premium." An insurance contract's service area determines the counties where enrollees can purchase plans offered within the contract. Premiums and benefit design can vary across plans within a contract but cannot vary by enrollees in the same plan. Despite premiums, MA insurance plans critically rely on government subsidies from the Center for Medicare and Medicaid Services (CMS) to operate, which account for over 80% of the cost of covering an enrollee (Curto *et al.*, 2019).[5]

CMS introduced the star rating in 2009 as a summary measure of insurance quality. Through our study period 2009-2014, the rating is computed each year on a scale of 1 to 5 stars with half-star increments. It is displayed to consumers together with premiums and benefits on the plan choice website. With the introduction of pay-for-performance in MA under the Affordable Care Act (ACA) in 2012, the star rating also became the basis of bonus payments to high quality insurance contracts.

The star rating summarizes a large number of measure-level ratings focusing on specific aspects of insurance quality. Measure-level ratings are assigned based on a cluster analysis of performance data.[6] In 2009-2011, the overall star rating is a simple average of measure ratings. Starting 2012, measures of enrollee health outcomes receive the largest weights

[5] Vilsa Curto and coauthors counterfactually estimate that the average enrollee costs a MA plan $805 (see their Table 3), while plans demand subsidies to CMS for $746 on average (see the notes to their Figure 3-4). The remainder (20%) is charged to enrollees.

[6] The cluster analysis generates cut-points of star ratings such that contracts with similar performances receive the same star rating. Cut-points of specific measures are available in the yearly Technical Notes published by the CMS.

(3.0) in the overall rating. Measures of access and customer services receive 1.5 weights, and measures of managed care processes such as preventive care receive 1.0 weights.[7] Because performance data are collected from all enrollees in the contract, subsidiary plans share the same quality ratings as the contract.

Not all quality measures account for differences in enrollee characteristics when computing the star rating. For instance, according to the health outcome measures, a chronic condition is "managed" if results from related medical tests meet a pre-determined threshold, which however is *not* adjusted for the severity of conditions. Part D measures of drug safety and adherence for patients with diabetes or hypertension may suffer from similar biases. These measures of chronic conditions are derived from the Healthcare Effectiveness Data and Information Set (HEDIS), which lacks information on diagnoses to adjust health outcomes by disease conditions.[8] By contrast, survey-based measures on access and customer service are adjusted for the age, education, and the general health status of enrollees.[9] Thus, if riskier patients have chronic conditions that are more challenging to manage, they can worsen the health outcome measures and the overall contract rating.

2.1 Conceptual Framework

Before considering the implications of enrollee risk types for firm strategy under pay-for-performance, we start by describing how CMS disburses subsidies. CMS payments to MA plans are determined by comparing the plan's asking price, or bid, with its benchmark,

[7] Appendix Table D2 lists all quality measures in the 2013 rating, together with the weight, the underlying data source, and the period over which data are collected for each measure.

[8] The HEDIS measures were first introduced in 1991 to monitor patients' health outcomes and compare them across health plans, but it was not designed to measure a plan's value added because it does not collect information on diagnoses (Mainous III and Talbert, 1998). The health outcome measures use the lab test data in HEDIS to monitor the management of chronic conditions. For diabetes, hemoglobin A1c and low-density lipoprotein cholesterol (LDL-C) test results are collected, and the condition is managed if hemoglobin A1c is tested below 9%, and LDL-cholesterol level is below 100 mg/dL. Details of the outcome measures are available in the yearly Technical Note published by CMS.

[9] The access measures are based on the Consumer Assessment of Healthcare Providers and Systems (CAHPS) dataset, where respondents rate the health plan in terms of getting needed care, complaint resolution, and customer service. As explained in AHRQ (2017), adjusting "makes it more likely that reported differences are due to real differences in performance, rather than differences in the characteristics of enrollees or patients."

which is predetermined by CMS. The bid (denoted b) reflects the projected cost of an average enrollee in the plan plus an administrative load. Equation 1 below shows that, if the bid is below the benchmark (denoted B) times a quality adjustment (θ^{star}), the payment equals the plan's bid plus a rebate. By law, the rebate is passed on to enrollees as premium discounts or additional benefits.[10] Since payments are capped at the benchmark, a plan charges enrollees an extra premium if it bids over its benchmark. Formally,

$$payment = \begin{cases} b + rebate & \text{if } b < \theta^{star} \cdot B \\ \theta^{star} \cdot B & \text{if } b \geq \theta^{star} \cdot B. \end{cases} \tag{1}$$

Before 2012, $\theta^{star} = 1$ and the rebate was 75% of the positive difference between B and b for all contracts. Under the ACA, θ^{star} was set to 1.05 for plans with star ratings above 4.0.

A key issue of regulated insurance markets is that different enrollees require different health services despite the same premium charged to all enrollees. To reduce the scope for selection, CMS updates per-capita payments to reflect each enrollee's expected cost, thereby making potential enrollees equally profitable to insurers. Instead, under pay-for-performance, the per capita subsidy is, in turn, a function of quality measures that may depend on some features of the enrolled population if not adequately risk-adjusted. Thus, selecting healthier enrollees may result in better quality measures, implying higher subsidies to the insurer.

The first question we raise in this paper is: How do insurers react to pay-for-performance in the absence of an adequate risk adjustment? To guide our analysis, Appendix A.1 presents a stylized model of MA insurer behavior which shows that even with perfect risk adjustments on benchmarks, a biased star rating that is responsive to the contract's risk pool would push insurers to select healthier enrollees. As a result, insurers will price-discriminate across counties to select healthier enrollees as premiums

[10] Rebate equals $\gamma^{star} \cdot (\theta^{star} \cdot B - b)$, which increases for lower bidding plans according to a fixed percent γ^{star}. Similar to the quality adjustment on benchmark, the rebate percent γ^{star} increased with the star rating after 2012.

are allowed to differ across counties for the same contract.

Focusing on an insurer offering one insurance contract in two counties, Appendix A.1 finds that the insurer will set a lower premium in the healthier county if the star rating rewards a healthier risk pool. The price drop in this county compared to the risky one is proportional to the difference in FFS risk scores across counties ($\Gamma_1^{FFS} - \Gamma_2^{FFS}$) and is mediated by the change in the benchmark bonus due to a change in the contract's average risk score ($\frac{dB}{dq} \cdot \frac{\partial q}{\partial r}$), as in

$$\Delta p_1 - \Delta p_2 \propto -\frac{dB}{dq} \cdot \frac{\partial q}{\partial r} \cdot \left(\Gamma_1^{FFS} - \Gamma_2^{FFS}\right), \tag{2}$$

where we denote the premium difference in county l before and after the quality initiative by Δp_l, the star rating by q, the risk score of an average enrollee in the contract by r, and the risk score of the FFS population in county l by Γ_l^{FFS}. The level difference in FFS risk scores naturally drives the premium difference because drawing additional enrollees from counties with lower Γ_l^{FFS} improves a contract's risk pool and quality rating.[11] Thus, we expect premiums to drop in low-risk counties compared to high-risk ones. This reasoning also applies under oligopoly, with the caveat that the selection incentive weakens with the number of firms as raising premiums also means losing revenues to competing insurers (see Appendix A.2).

The second question we raise relates the selection strategy to changes in the market shares of high-rated insurance across counties. We show in Appendix A.4 that the premium responses could lead to rising inequality in the access to high-rated insurance across counties, which may disadvantage consumers in the riskiest counties of Medicare. As a result, welfare may decrease if enrollees in the riskier counties value MA insurance more than those in the healthy counties. We empirically explore the potential impacts on welfare examining the shifts in the spatial distribution of insurance and discuss the

[11] This argument is net of the effect of the risk score of the marginal enrollees across counties, which we account for using fixed effects of contract-county pairs in the empirical analysis.

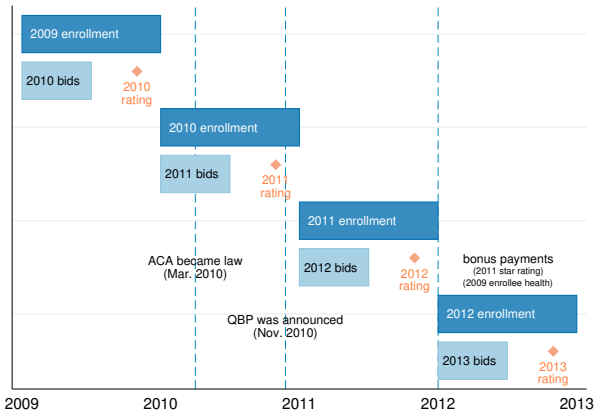
implications for consumers.

2.2 Quality Bonus Payment Demonstration

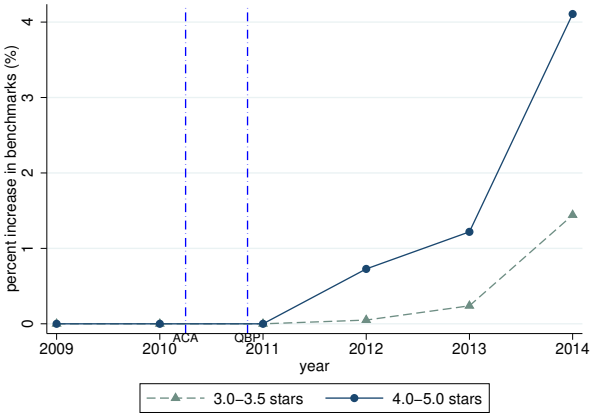
The ACA model was not immediately implemented in MA. Bonus payments were determined by the QBP demonstration between 2012 and 2014, which CMS introduced on November 10th, 2010 as a phase-in of the ACA model. We summarize the QBP bonus rates for benchmarks and rebates in 2009-2014 and the ACA rates (2015) in Appendix Table D1.

Figure 1: Star rating computation and its implications for benchmarks

(a) Timeline of bidding, enrollment, and star rating disclosure, 2009-2012



(b) Growth in rating-adjusted benchmarks after the payment reform



Notes: Panel (a) plots the timeline of Medicare Advantage (MA) enrollment, plan bidding, and star rating disclosure for enrollment years 2009-2012. Bonus rates for 2012 are calculated from the 2011 star rating (released in the Fall of 2010), which in turn is derived from the health outcomes of enrollees in 2009. Panel (b) plots the percent increase in rating-adjusted benchmarks after the payment reform, for contracts below and above the ACA cut-off (4.0 stars) in the baseline period (2009-2010). We distinguish contracts by the maximum quality rating in 2009-2010, and use the baseline rating to determine the bonus rates applicable to the contract in 2012-2014.

Bonus payments under QBP rewarded a contract's *past performances*. As illustrated in panel (a) of Figure 1, a three-year lag exists between enrollment in year t and the payout of bonus payments in year $t+3$. This is because payments for year t are adjusted by the star rating in $t-1$, where most quality measures are based on enrollee data collected two years

prior in $t-3$.[12] The three-year delay effectively links enrollees serviced in 2012 (and their outcomes) with payments in 2015, when the ACA model restricts benchmark bonuses only to contracts rated 4.0 stars and above. This implies that contracts may begin selecting healthier enrollees immediately after bonus payments became law. With the passage of the ACA in early 2010, we examine insurer selection responses on premiums and risk scores treating 2011 as the first post-reform year.[13]

To understand the magnitude of benchmark bonuses, panel (b) of Figure 1 predicts benchmarks both for contracts rated 4.0 stars and above in 2009-2010 and for lower-rated contracts.[14] By 2014, the year when bonus payments aligned with the ACA model for higher-rated contracts, benchmarks increased by 4.1% for higher-rated contracts, or by $33 per enrollee-month above the 2009-2010 levels.[15]

2.3 Data

We draw data from the administrative registry of all MA insurance plans offered in 2009-2014 (the "Landscape File"). The data contain information on plan characteristics such as premiums and drug deductibles across the service areas (counties) covered by each plan. We drop Regional Preferred Provider Organization (PPO) plans and plans with missing star ratings for payment purposes since these plans are subject to a different set of payment rules. We further restrict the sample to a homogeneous set of plans covering both medical and prescription drug expenditures, or the MA-PD plans. Details of the sample

[12] In particular, the HEDIS outcome and process measures are based on enrollee health records from two years prior. Access measure are more up-to-date, with year t ratings derived from CAHPS records from the first half of year $t-1$. Appendix Table D2 list the period of data collection for each quality measure in the 2013 rating.

[13] Appendix Table D3 illustrates the ACA policy variation in bonus rates linking the star rating in year t with the payment model in $t+3$. Relevant for the selection incentive during QBP, benchmark bonuses increased discretely from 0% to 5% above 4.0 stars.

[14] We predict quality-adjusted benchmarks for 2012-2014 using the maximum Part C rating in 2009-2010 as the basis. We adjust the raw county benchmarks with the baseline rating, and use the average benchmark across counties as the predicted benchmark for contracts. In the prediction, we restrict counties to those already covered by the insurance contract prior to the payment reform.

[15] The benchmark increase did not exactly match the 5% bonus rate because raw county benchmarks were generally lower since 2012. We survey more recent policy changes in MA after 2014 in Appendix C.

construction are available in Appendix B.

We merge this data with the Payment File containing plan payments and plan risk scores to examine plan bidding and risk selection. Since the quality rating is calculated at the level of insurance contracts, we focus on contract-level differences by averaging over subsidiary plans using enrollment weights. The first two columns of Table 1 summarize the estimation sample. Panel A looks at contract-year observations, while Panel B expands the contract-year observations by the counties in the contract's service area. On average, a MA-PD contract offers 3.4 plans covering over 25 counties in its service area. Most contracts place bids below the benchmark, generating a rebate of $81.04 per enrollee-month. A large number of contracts charge zero premiums and zero drug deductibles.

3 Evidence of Risk Selection

High- and Low-Rated Contracts. To provide evidence on the selection responses, we document shifts in the distribution of risk scores across high- and low-rated insurance contracts. We group insurance contracts by the maximum Part C rating in 2009-2010, our baseline period.[16] High-rated contracts have at least one 4.0-star rating or above in the baseline, whereas low-rated contracts are rated no more than 3.5 stars in the baseline.[17] Over time, risk scores shifted to the lower percentiles in high-rated contracts but not in low-rated contracts (Appendix Figure E1). In particular, risk scores shifted in high-rated insurance in 2011, the first year after quality bonus payments were signed into law under the ACA in March 2010 (Appendix Figure E2).

Responses by Star Ratings. We further examine heterogeneous responses across baseline ratings in Figure 2. We classify contracts by the maximal Part C rating in 2009-2010 and

[16] Part C and Part D ratings are calculated separately for MA-PD contracts in 2009-2010. Because the Part C rating includes two-thirds of all measures in Part C and D, the overall rating (constructed as the average of all measure ratings) is primarily driven by the Part C rating. We find similar selection responses across the overall rating in Appendix Figure E3.

[17] We exclude contracts with a 2.5-star rating or below from our analysis. These contracts are subject to suspension by the CMS if the Part C rating does not improve above 3.0 stars in three years. Since the threat of suspension differs from our focus on bonus payments, we exclude these contracts from the analysis.

Table 1: Summary statistics

	(I)	(II)	(III)	(IV)	(V)	(VI)
	Full Sample		Low-Rated		High-Rated	
	mean	s.e.	mean	s.e.	mean	s.e.
	Panel A: Contract-Year Observations					
Risk Score	0.97	0.007	0.97	0.009	0.96	0.12
Number of Counties	25.09	5.40	25.19	7.74	18.18	2.21
Number of Plans	3.40	0.23	3.53	0.31	3.12	0.28
Service Area Risk	0.99	0.007	1.00	0.009	0.96	0.009
Enrollment (k)	334.75	34.95	328.35	39.19	349.06	71.56
Benchmark	899.95	5.82	909.93	6.70	877.67	10.78
Bid	786.05	6.37	787.09	7.76	783.73	11.15
Benchmark-Bid	113.90	5.71	122.84	7.11	93.94	8.89
Rebate	81.04	3.85	86.45	4.83	68.94	5.89
Part C Premium	30.78	2.55	21.06	2.64	52.47	4.69
Zero Part C Premium (%)	48.74	2.81	59.27	3.29	25.23	3.90
Part D Premium	19.96	1.22	15.42	1.40	30.10	1.77
Zero Part D Premium (%)	44.23	2.87	54.98	3.42	20.23	3.68
Drug Deductible	33.33	4.51	33.51	5.84	32.92	6.53
Zero Drug Deduc (%)	84.21	1.89	84.70	2.36	83.11	3.07
N	1,122		775		347	
	Panel B: Contract-County-Year Observations					
Enrollment (k)	18.25	2.35	17.00	2.48	21.57	4.64
Number of Plans	1.76	0.073	1.59	0.088	2.22	0.093
Part C Premium	33.03	2.75	26.05	2.86	51.53	5.66
Zero Part C Premium (%)	37.36	3.25	43.06	4.03	22.25	4.83
Part D Premium	21.27	1.47	18.29	1.79	29.16	2.18
Zero Part D Premium (%)	35.04	3.27	41.49	4.06	17.97	4.29
Drug Deductible	29.44	6.32	30.99	8.31	25.33	6.30
Zero Drug Deduc (%)	84.26	2.95	83.40	3.87	86.55	2.91
Market Share (%)	33.51	1.72	31.77	1.97	38.12	3.20
N	20,472		14,861		5,611	

Notes: The table summarizes the estimation sample. We aggregate plan characteristics to the contract-year level in Panel A, and to the contract-county-year level in Panel B, both weighting by enrollment. Enrollment is total enrollee-month counts in a year, and price variables are in 2012 dollars per enrollee per month. Bids and benchmarks are risk-adjusted to reflect the cost of a standard-risk enrollee. Column 3-6 show summary statistics by contract rating. High-rated contracts (column 5-6) have at least one 4.0-star rating or above in the baseline (2009-2010). Low-rated contracts (column 3-4) are never rated 4.0 stars or above in the baseline. We exclude contracts rated below 3.0 stars in both 2009 and 2010: these contracts are subject to suspension if the star rating does not improve in 2011. Column 1-2 summarizes the full estimation sample combining high- and low-rated contracts. Details of the sample construction are in Appendix B.

plot the density shifts for each rating from 3.0 stars to 4.5 stars.[18] Risk scores decreased the most in marginal high-rated contracts with a maximum 4.0-star rating in the baseline (panel c), where the density shifted significantly from the middle to the lower percentiles.[19] We find weaker selection responses among higher-rated contracts and no significant shifts in risk scores among low-rated contracts.

Quantile Difference-in-Differences. We then formally estimate the shifts in the distribution of risk scores using a quantile-based difference-in-differences design. We model the κ-th quantile of risk score $y_{qt}(\kappa)$ for quality rating q in year t as

$$y_{qt}(\kappa) = \beta(\kappa) \cdot high_q \cdot post_t + \alpha_q(\kappa) + \tau_t(\kappa) + \epsilon_{qt}(\kappa), \tag{3}$$

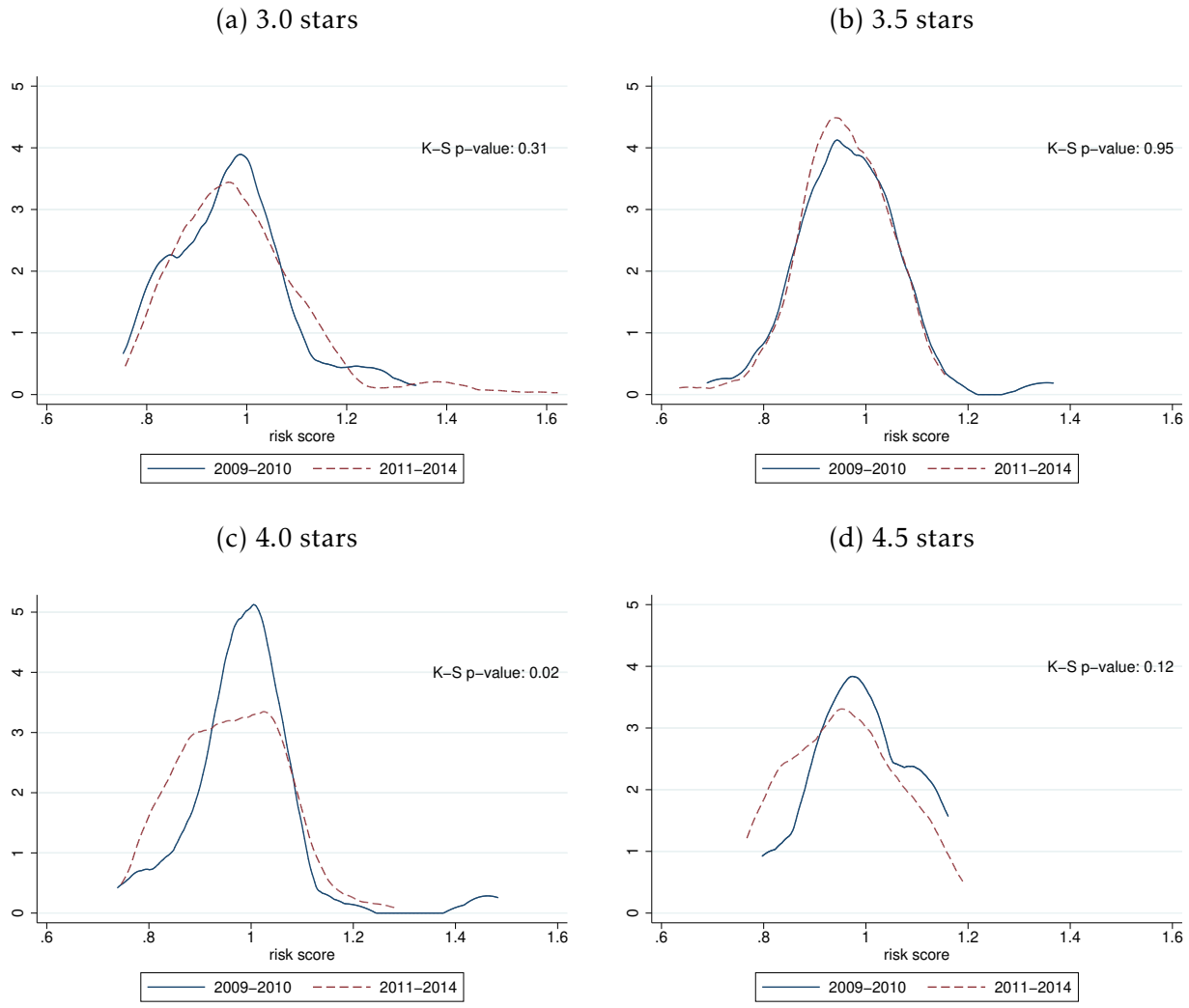
where $high$ indicates high-rated insurance and $post$ indicates the post-reform years (2011 and after). $\beta(\kappa)$ estimates the shift in the κ-th quantile of risk scores of high-rated contracts after the payment reform. We control for rating, $\alpha_q(\kappa)$, and time fixed effects, $\tau_t(\kappa)$. Appendix Figure E5 estimates equation 3 using the group quantile estimator (Chetverikov *et al.*, 2016) in panel (a) and the changes-in-changes (Athey and Imbens, 2006) in panel (b). Both estimates show large and significant reductions in the 20% to 40% of risk scores. In these deciles, risk scores dropped by 4-8 percentage points in high-rated contracts, or by 4%-9% below their baseline levels (Appendix Table D4). The effects on risk scores in the upper deciles are smaller and statistically insignificant.

High-Selection Contracts. The quantile analysis indicates highly heterogeneous responses in risk scores, with most of the reduction concentrated in the lower percentiles of high-rated insurance. At the contract level, this implies that risk scores decreased disproportionately for some, but not all, high-rated contracts. To examine the average

[18] We find similar shifts in risk scores across the maximum overall rating in Appendix Figure E3. We do not separately plot risk scores for 5.0-star contracts because very few obtained such rating at baseline.

[19] We find that contracts closer to 4.0 stars (within a half star radius) show larger drops in risk scores (Appendix Figure E4). Within those contracts, the drop is further concentrated in those with a maximum 4.0-star rating in the baseline (Appendix Figure E3).

Figure 2: Effect on risk scores by the baseline rating, kernel density



Notes: The figure plots the kernel density of risk scores by the baseline rating of contracts. For each rating from 3.0 stars to 4.5 stars, the figure compares the density of risk scores before and after the payment reform, and tests for the null of equal distribution applying the Kolmogorov–Smirnov (K-S) test with the p-value shown next to the density. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

and heterogeneous effects of the payment reform on high-rated contracts, we estimate the following specification

$$y_{ct} = \beta \cdot treat_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}, \tag{4}$$

where y_{ct} is the risk score of contract c in year t. We include contract (α_c) and year (τ_t) fixed effects. $treat$ indicates different sub-groups of high-rated contracts. β estimates the

effect of bonus payments on the risk scores of the high-rated contracts indicated by $treat$.

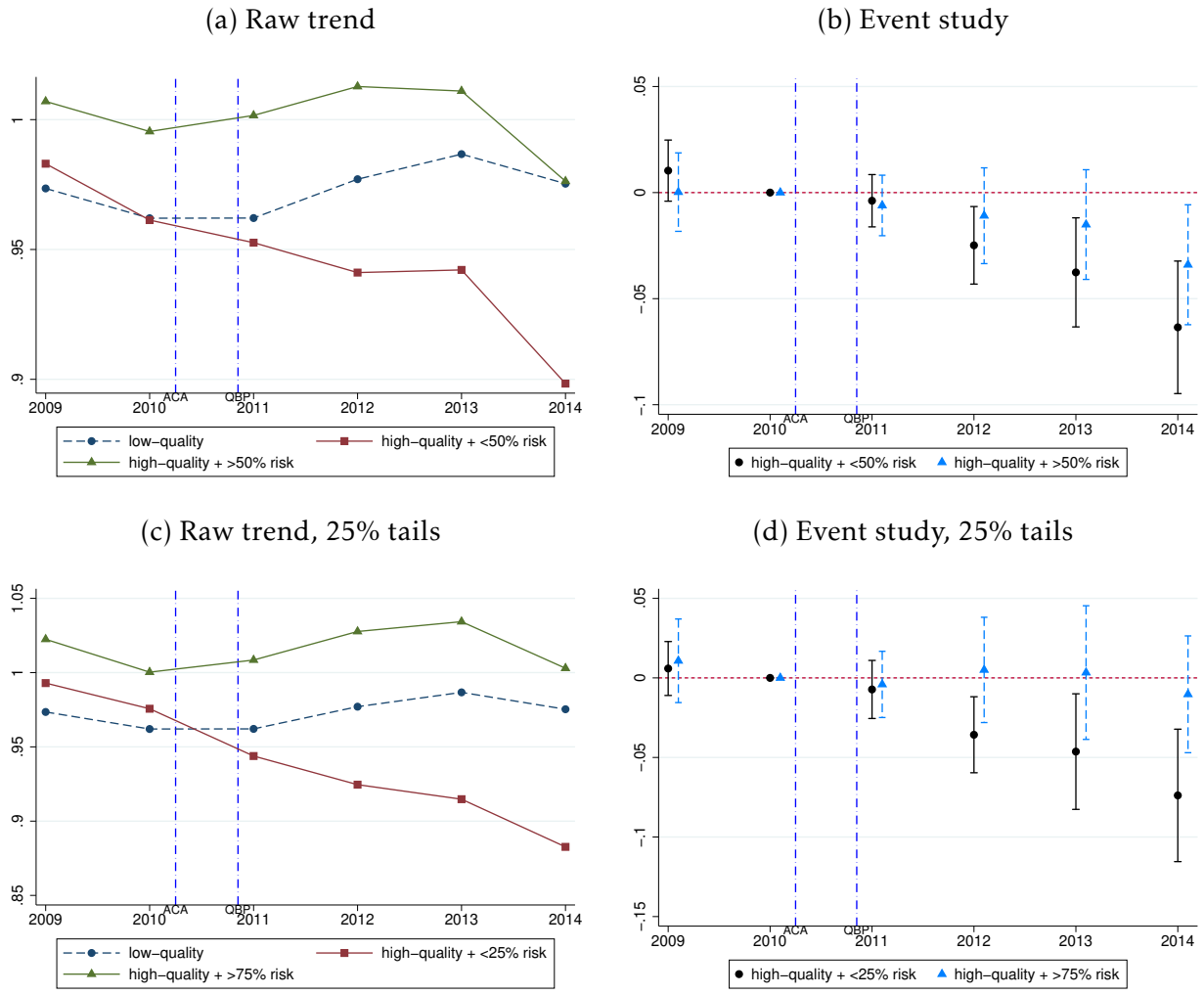
Guided by the model in Section 2.1, we explore heterogeneous effects for contracts with different fee-for-service risk scores in the service area. Since contracts can more effectively select healthier enrollees in counties with lower FFS risk scores (equation 2), risk scores may decrease more for contracts exposed to healthier FFS enrollees in the service area. We hence calculate the service area risk as the average FFS risk score in the service area and consider heterogeneous effects by the median service area risk in high-rated contracts.

Panel (a) of Figure 3 plots the raw trends of risk scores for two groups of high-rated contracts and for low-rated contracts. Risk scores trended similarly for high-rated contracts above the median service area risk and for low-rated contracts, but dropped for high-rated contracts serving healthier locations. Panel (c) shows similar patterns across the lower and upper 25% of service area risks. By contrast, risk scores decreased in high-rated contracts serving the healthiest locations.

Appendix Table D5 estimates the heterogeneous effects on high-rated contracts using equation 4. On average, risk scores decreased by 2.6 percentage points in high-rated contracts (column 1). This effect is driven by high-rated contracts in the lower percentiles of service area risks (column 2-5). Risk scores dropped by 3.7 percentage points below the median service area risk in column 2, and by 4.3 percentage points below the 25th percentile in column 4. Conversely, risk scores did not differ meaningfully between low-rated contracts and high-rated contracts serving riskier locations (column 3 and 5).

To summarize, the overall decrease in risk scores is concentrated in what we term "high-selection" contracts – high-rated contracts with below-median service area risks in the baseline. This heterogeneous effect is consistent with the theoretical prediction that insurers can more effectively select healthier enrollees in counties with lower fee-for-

Figure 3: Effect on risk scores, by service area risks, event study



Notes: The figure shows the changes in the risk scores of high-rated contracts with different service area risks. Panel (a) shows the raw trends of risk scores for high-rated contracts above and below the median service area risk (0.975) and for low-rated contracts. Panel (b) shows the event study estimates for the high-rated contracts in panel (a). Panel (c) shows the raw trends of risk scores for high-rated contracts in the lower and upper 25% of service area risks (below 0.902 or above 1.009) and for low-rated contracts. Panel (d) shows the event study estimates for the high-rated contracts in panel (c). We plot 95% confidence intervals based on robust standard errors clustered at the contract levels in panel (b) and (d).

service risk scores. We next examine pricing responses as the mechanism of selection.

4 How did Insurers Risk Select?

Drawing from the discussion in Section 2.1, we investigate whether high-rated insurance increased (decreased) premiums in riskier (healthier) counties after the payment

reform. Specifically, we implement the following tripe-difference design

$$y_{clt} = \beta_0 \cdot risk_{cl} \cdot high_c \cdot post_t + \beta_1 \cdot risk_{cl} \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt}. \quad (5)$$

The variables *high* and *post* indicate the high-rated contracts and the post reform period as in Section 3. The outcome variables are prices varying at the level of contract c, year t, and county l. In each county, we generate contract-level prices from plan prices weighted by enrollments. The variable $risk_{cl}$ measures the risk score differences across counties in a contract's service area. In particular, we calculate county l's deviation to the median county risk score in the service area of contract c and use the deviation-to-median to measure $risk_{cl}$ in the analysis.[20] By construction, $risk_{cl}$ varies across counties within contracts and varies across contracts within counties.[21]

We include contract-county fixed effects α_{cl} to absorb pre-existing differences in prices and enrollments across contracts and counties.[22] We control for year fixed effects in τ_t. Assuming that premiums in high- and low-risk counties would have followed parallel trends absent the payment reform, β_1 gives the effect of bonus payments on premiums in low-rated contracts. Further assuming that premium differences by county risk scores would have trended similarly between high- and low-rated contracts absent the reform, β_0 gives the differential effect of the payment reform on premiums in high-rated contracts. β_2 gives the effect on premiums in the median risk county served by high-rated contracts.[23]

We also control for time-varying, location-specific payment incentives that may affect prices in these locations. Specifically, X_{lt} includes yearly raw benchmarks, bonus rates, and bonus-adjusted benchmarks.[24] In addition to varying prices, insurers may also enter

[20] Specifically, we derive the median county risk score across all counties covered by a contract and measure risk score differences within contracts relative to the median county. Appendix F explores alternative measures of risk differences within contracts.

[21] Based on the variation in $risk_{cl}$, we cluster standard errors by contracts and counties.

[22] The fixed effects absorb local consumer characteristics which did not vary with the payment reform.

[23] When evaluated at the median county risk, $risk_{cl} = 0$ and interaction terms containing $risk_{cl}$ vanish in equation 5. β_2 gives the price change in the median county for high-rated contracts after the reform.

[24] We use the maximum bonus applied to 5-star contracts to measure a county's benchmark generosity.

high-bonus counties or exit high-risk counties to increase bonus payments. However, we find little evidence of selection over service area characteristics.[25]

4.1 Varying Premiums to Risk Select Enrollees

Part D Premiums. Because the health outcome measures in the quality rating focus on chronic conditions such as diabetes and hypertension, we first examine if premiums of prescription drug coverage (Part D) varied across counties in response to the payment reform. We show estimates of equation 5 in Table 2. For every 10 percentage point increase in the risk score, Part D premiums increased more in high-rated contracts by $1.53 (column 3), or by 8.3% above the average premium. The response is driven by high-rated contracts (column 2), and we do not detect similar differences in low-rated contracts (column 1). To the extent that larger risk differences may exacerbate the premium responses, we also examine premiums across the risk tails of counties in column 4-6. Overall, we find very similar responses in the risk tails.

We illustrate the premium differences showing raw trends and event study estimates in Appendix Figure E6. On the raw trends, we split the service area of each contract into high- and low-risk counties – grouping either by the median or across the 15% tails – and plot premium trends across county risks for an average high- and low-rated contract. In high-rated contracts, Part D premiums deviated from pre-reform parallel trends and increased significantly in the riskiest counties since 2011. We do not find similar increases in low-rated contracts in the event study.

Part C Premiums. We then examine responses in Part C premiums in Appendix Table D7. We do not find significant premium differences across county risk scores in either low- or high-rated contracts (column 1-2). In column 4-6, we also do not find premium differences across the 15% risk tails of counties. Part C premiums trended similarly across both rating-groups over the sample period, and the event study estimates generally show insignificant

[25] Specifically, high-rated contracts did not cover additional counties or change the composition of covered counties based on risk scores or benchmarks. Appendix Table D6 shows the estimates.

Table 2: Effect of the payment reform on Part D premiums, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			15.28**			17.43**
			(6.99)			(8.51)
Risk · Post	-4.29	17.66***	-2.97	-4.01	16.64**	-3.36
	(5.00)	(5.82)	(4.91)	(5.57)	(7.35)	(5.35)
High · Post			1.23			2.38
			(2.35)			(2.00)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	18.29	29.16	21.27	18.05	27.99	20.74
R^2	0.76	0.67	0.75	0.75	0.70	0.75
N	14,861	5,611	20,472	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in Part D premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium differences in low- and high-rated contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-rated contracts. Column 4-6 repeat the analysis but restrict the within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

differences by county risk scores (Appendix Figure E7). Overall, high-rated contracts increased their total Part C and D combined premiums by $4.05 for a ten percentage point increase in the risk score, or by 7.6% above the mean (Appendix Table D9 and Figure E9).[26]

Drug Deductibles. We do not find significant differences in the drug deductibles of high-rated contracts (Appendix Table D10). Across risk tails, both high- and low-rated contracts increased drug deductibles by approximately $3 per ten percentage point risk

[26]We also examine the offering of zero-premium plans across counties as one particular margin of selection. Consistent with the differences in premiums, high-rated contracts offered more plans with zero Part D premiums in low-risk counties and decreased the offering of such plans in high-risk counties (Appendix Table D8). We do not find similar responses for zero Part C premiums, or by low-rated contracts. Appendix Figure E8 plots the raw trends.

score. However, raw trends and event study estimates indicate a significant pre-reform difference in 2009 for high-rated contracts (Appendix Figure E10). Due to the noise in the data, we do not pursue differences in drug deductibles as a potential selection mechanism of high-rated insurance.

4.2 Mechanism

While the premium differences are consistent with the selection of healthier individuals in low-risk counties, similar differences could also emerge from premium responses to other county characteristics correlated with risk scores. For instance, if high-rated contracts targeted high-income markets where risk scores tend to be lower, then the premium differences may be driven by selection over (non-risk) demand factors rather than risk types. Here we consider a range of demand and supply factors that can plausibly generate the premium differences through the correlation with risk.[27]

Socio-Economic Factors. Appendix Table D11 estimates the premium differences by county differences in per capita income and transfer income. We do not find a significant premium differences with either measure of income. Specifically, premiums did not increase in high-transfer counties or decrease in high-income counties, contrary to the risk composition gain in high-rated contracts. Appendix Table D12 finds similar null effects by county demographics such as racial composition and college education.

Special Enrollment Period. Premiums may also differ in response to the Special Enrollment Period (SEP), a policy change in 2012 that allowed enrollees to switch to a 5.0-star MA contract anytime during the year. SEP may increase the risk exposure of 5.0-star contracts and hence trigger additional selection responses (Decarolis and Guglielmo, 2017). However, since very few contracts ever achieved 5.0-star ratings, excluding 5.0-star contracts and counties with 5.0-star contracts has little effect on the premium differences

[27] Details of the county characteristics examined here are provided in Appendix B.2.

across counties (Appendix Table D13).

Market Concentration. We examine the role of market concentration in Appendix Table D14. High-rated insurance increased premiums more in more concentrated markets (column 2),[28] but because these markets also have healthier enrollees,[29] the differences would drive up premiums in healthier markets. However, controlling for the differences in market concentration, premiums increased significantly with county risk scores in high-rated insurance (column 5). These results are consistent with the prediction in Appendix A.2 that competition forces could weaken the selection through premiums.

Provider Quality. We next consider differences in provider costs and quality as alternative drivers of the premium differences. If high-risk counties are associated with lower quality and higher costs, then payments to improve outcomes in these counties can crowd out rebates to enrollees, generating the premium differences over risk scores. To investigate the quality channel, we use hospital readmission rates and preventable hospital stays as measures of inpatient and outpatient quality. However, we do not detect consistent differences over these measures across counties (Appendix Table D15).

Provider Cost. We investigate the cost channel exploiting adjustments on fee-for-service (FFS) costs in Appendix Table D16. Premiums did not differ by FFS costs in low-rated contracts. In high-selection contracts where risk scores decreased more (column 3), premiums tend to increase with FFS costs. Similar patterns hold when we adjust FFS costs for the differences in price levels in columns 5-8.[30] Further adjusting for the risk of enrollees in price-standardized costs, premiums no longer differ across FFS costs in columns 9-12, where the coefficient for high-selection contracts is insignificant.[31] Thus, premiums varied

[28] We measure concentration using the Herfindahl-Hirschman Index (HHI), calculated for county l as $HHI_l = \sum_c (s_{cl})^2$, where s_{cl} is the market share of contract c in the county. Concentration rates across all contracts affect the premiums of high-rated contracts, but concentration within county-quality pairs has no significant impacts on premiums (column 7-9).

[29] Across counties, a ten percentage point increase in the risk score is associate with a 6% decrease in concentration as measured by HHI.

[30] The adjustment uses national input prices to calculate labor and facility costs and replace local reimbursement rates with fixed national ones.

[31] Appendix Table D17 finds similar patterns in the risk tails. We find substantially smaller and insignifi-

with costs through the risk composition across space, rather than differences in the price levels or practice styles.

Coding Intensity. Finally, since counties with more intensive coding of diagnoses have higher risk scores for similar health conditions, premiums could instead respond to the coding intensity of fee-for-service risk scores. To remove risk score differences driven by coding intensity rather than health, Appendix Table D18 adjusts risk scores with the diagnosis intensity factors developed by Finkelstein *et al.* (2017).[32] Upon adjustment, we find a stronger variation of Part D premiums over risk scores relative to the main results in Table 2. Thus, premiums responded directly to the health of enrollees rather than location-specific non-health factors coded in the risk score.

Although it is impossible to consider all correlates of risk, we can rule out common demand and supply factors as drivers of the premium differences over county risks. Moreover, exploiting adjustments on costs and risk scores, we show that premium responded directly to the health of enrollees in the county, but not to local price levels, practice style, or other non-health factors coded in the risk score.

4.3 Insurance Generosity

Other price and non-price designs of the insurance contract may also vary in favor of healthier individuals. To understand the extent of insurance generosity that can be explained by premiums, we estimate equation 5 using rebates as the dependent variable in column 4 of Appendix Table D19. The estimate suggests that rebates increased by \$5.63 less in high-rated contracts for every ten percentage point increase in the county risk score. Of the \$5.63 loss of rebate, \$4.07 was added onto premiums in high-rated contracts (Appendix Table D9). Put together, premium differences account for 72% of the

cant premium differences over costs after adjusting for the risk composition across counties.

[32] These adjustors are generated from movers in the elderly FFS population who have similar underlying health conditions but different risk scores due to location-specific coding intensity. By construction, the adjustors remove cross-space differences in risk scores for a given level of underlying health conditions.

differences in the overall generosity by quality.[33]

In contrast to the significant premium differences across county risk scores, average rebates and premiums did not increase for high-selection contracts. Specifically, we estimate a contract-level difference-in-differences (equation 4) where the outcome variable is the average premium and rebate for enrollees in a high-rated contract (Appendix Table D21). We estimate a similar null effect on rebates for high-rated contracts more generally (Appendix Table D22). We thus conclude that insurers selected healthier enrollees by shifting insurance benefits – in particular premium discounts – from riskier to healthier counties, without changing the average benefit levels of high-rated insurance.

5 Why Does the Payment Reform Induce Risk Selection?

5.1 Selection in the Health Outcome Measures

The quality rating is a weighted average of different measure-level ratings, whose weights increased differentially across measures in 2012 (see Section 2). Although all measures received unit weights before 2012, CMS increased the weight of health outcome measures to 3.0, the largest of all weights in the quality rating. The weight change significantly increased the contribution of outcome measures to the final rating linked to payments, especially for high-rated contracts (Appendix Table D23). Here, we explore biases in the health outcome measures as a potential driver of selection through two different empirical strategies. We then examine variation in premiums across counties as a source of selection to specifically improve health outcome measures.

Cross-Contract Evidence. The first strategy exploits the payment reform and the cross-contract differences over baseline risk scores in a difference-in-differences analysis analo-

[33] Similar calculation for high-selection contracts suggests that premium differences (Appendix Table D20) account for about 65% of the rebate differences between low-rated and high-selection contracts.

gous to equation 4. Specifically, we estimate

$$y_{ct} = \beta \cdot risk_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}, \tag{6}$$

where $risk_c$ is the baseline enrollee risk score in contract c. The specification compares the health outcome rating y_{ct} across contracts that started out with different risk scores in the baseline. The results in Appendix Table D24 show that a 10 percentage point increase in the baseline risk score is associated with a loss of 0.12 stars (over a range of 1-5 stars) in subsequent outcome ratings (column 1).[34] This correlation is driven by the HEDIS measures of chronic conditions (column 3), which improved significantly for contracts with healthier enrollees in the baseline.[35]

This correlation may reflect the fact that the HEDIS measures are not adjusted for the prevalence or severity of health conditions. In turn, this affects the ranking of contracts if contracts differ significantly by the case-mix of health conditions.[36] In the presence of such bias, outcome ratings should improve more for selecting contracts when the HEDIS outcomes of their enrollees enter the quality rating. This observation motivates our second empirical strategy.

Evidence Over Time. The second empirical strategy examines the relationship between outcome ratings in year t and risk scores in year $t-2$ with the following specification

$$y_{ct} = \beta \cdot risk_{ct-2} + \alpha_c + \tau_t + \epsilon_{ct}. \tag{7}$$

[34] In this analysis we consider only outcome measures that consistently appear in the quality rating from 2009 to 2014. Later introduced measures, such as hospital re-admission measures, drug adherence measures, and quality improvement measures, are not included in the difference-in-differences analysis. In Section 5.3 we consider the effect of risk scores on all quality measures using an instrumental variable approach.

[35] Appendix Figure E11 plots the raw trends and event study estimates.

[36] The health literature has raised similar concerns over the lack of risk adjustments on the HEDIS quality measures. In the case of blood sugar control, for instance, Safford *et al.* (2009) show that adjusting for diabetes severity and co-morbidities meaningfully altered the quality ranking and outlier status of facilities in the Veteran Health Administration. Specific to the Medicare Advantage star ratings, Nichols *et al.* (2018) shows that patients with multiple co-morbidities are associated with worse medication adherence and blood sugar control.

We lag risk scores by two years because outcome ratings are based on the medical records of enrollees from two years prior (see Section 2). This implies that if riskier individuals have worse health outcomes, then the negative effect on the outcome rating would appear after a two-year delay. Consistent with selection on health outcomes, we find that lowering risk scores by ten percentage points improves outcome ratings by 0.30 stars for high-selection contracts two years later (column 6 of Appendix Table D25). We do not find similar correlation patterns for low- or high-rated contracts across other lag or lead periods.

Premiums and Outcome Measures. The selection incentive implies that premiums may respond to the chronic conditions targeted by the health outcome measures. We inspect such pricing responses here. Adopting the triple-difference design in equation 5, we compare premiums across counties with different diabetes prevalence rates in Appendix Table D26.[37] High-selection contracts increased Part D premiums by \$9.47-\$12.44 per ten percentage point increase in the prevalence rate (column 6-7), or by 47%-63% above the mean. Appendix Figure E12 shows the raw trends and the event study. We also find similar patterns but smaller magnitudes for hypertension (Appendix Table D27).

To summarize, high-selection contracts significantly varied premiums in favor of healthier counties with lower prevalence rates of chronic conditions. Both the risk pool and the health outcome rating improved for these contracts after the payment reform. Building on these results, we develop an instrumental variable strategy to quantify the extent of selection in the health outcome measures.

5.2 Quantifying Risk Selection in the Health Outcome Measures

This section quantifies the effect of risk scores on the HEDIS outcomes by developing an instrumental variable (IV) strategy that relies on our finding that insurers varied premiums across counties to attract healthier individuals and improve the risk pool.

[37] We multiply the raw prevalence rates by the coding-adjusted risk score to construct health-adjusted prevalence rates for our analysis. Prevalence rates are adjusted downward if enrollees in the county have fewer and milder conditions. We detail the prevalence rates in Appendix B.

Adjusting for Risk Score. We assume that the health outcome measures are determined by a contract-specific component and a component due to the risk scores of enrollees. Specifically, we estimate the following equation

$$y_{ct} = \alpha_c + \gamma_c \cdot post_t + \beta \cdot risk_{ct-2} + \tau_t + \epsilon_{ct}, \tag{8}$$

where y_{ct} is the health outcome (as measured by HEDIS) of contract c in year t. Since HEDIS outcomes are measured from enrollees two years prior, $risk_{ct-2}$ denotes the concurrent risk score of these enrollees at the contract level. We focus on HEDIS outcomes in 2011-2014 (corresponding to risk scores in 2009-2012) and define $post = 1$ for 2013-2014.

The intercept α_c is a contract fixed effect. We interpret α_c as the contract's ability to improve the chronic conditions of a unit-risk enrollee. Other than quality, outcomes may also improve due to selected risk types in $risk_{ct-2}$. Selection invalidates the ordinary-least-square (OLS) estimate of β. We employ an IV strategy to estimate the effect of risk scores on outcomes, and use it to "risk-adjust" the health outcome y_{ct}. Controlling for risk types, we infer the health improvement of a standard risk type from $\gamma_c \cdot post$, which we interpret as the change of insurance quality over time.[38] We estimate β specifically for high-selection contracts, where risk scores decreased more after the payment reform.

Instrument. We exploit the premium differences over county risk scores to construct instruments for $risk_{ct-2}$. Specifically, we construct the instrument $riskiv_{ct-2}$ as

$$riskiv_{ct-2} = Corr(p_{ct-2}, R_c) = \frac{1}{|N_c|} \sum_{l \in N_c} \frac{(p^l_{ct-2} - \bar{p}_{ct-2}) \cdot (R^l_c - \bar{R}_c)}{\sigma_{p_{ct-2}} \cdot \sigma_{R_c}}, \tag{9}$$

where p_{ct-2} stacks county l premiums, $\left(p^l_{ct-2}\right)_{l \in N_c}$, in the service area N_c of contract c. The denominator $|N_c|$ refers to the number of counties in N_c. Similarly, R_c stacks the fee-for-service risk scores of counties covered by contract c in 2009-2010, $\left(R^l_c\right)_{l \in N_c}$. We capture

[38] Since controlling for $\alpha_c \cdot \tau_t$ would absorb all the variation in our key variable of interest, $risk_{ct-2}$, we estimate the change in quality before and after the payment reform by $\gamma_c \cdot post$.

the premium differences across county risk scores using the covariance $\frac{1}{|N_c|}\sum_{l\in N_c}(p^l_{ct-2} - \bar{p}_{ct-2})(R^l_c - \bar{R}_c)$, where $\bar{p}_{ct-2}$ and $\bar{R}_c$ are the cross-county averages. We normalize the covariance by the standard deviation of premiums $\sigma_{p_{ct-2}}$ and risk scores σ_{R_c}, and use the correlation coefficient $Corr(p_{ct-2}, R_c)$ as the instrument $riskiv_{ct-2}$.[39]

The instrument summarizes the responsiveness of premiums to county risk scores. Contracts with larger $riskiv_{ct-2}$ price-discriminate more on the basis of risks when setting premiums across counties. These contracts potentially have healthier enrollees and hence lower risk scores due to the premium differences. We therefore predict contract risk scores using premium differences across counties as instruments in the first stage. We isolate premium differences by the health of enrollees using coding-adjusted risk scores for R_c (Finkelstein *et al.*, 2017) in equation 9. We construct additional instruments exploiting premium differences over diabetes and hypertension prevalence rates based on our results in Section 5.1.[40]

For the instruments to be valid, premium differences should impact the risk score of contracts but otherwise have no direct impact on the contract's health outcome measures. This requires that premium differences affected the composition of enrollee risk types, but did not affect unobserved determinants of health outcomes through the error term ϵ_{ct}. We examine the plausibility of the exclusion restriction based on the results in Section 4.2. Specifically, we show that premiums differed significantly with the health of enrollees but are not correlated with the supply or quality of providers, demand characteristics, or the competitiveness of the insurance market across counties. These results lend support to the exclusion restriction.

Selection in Health Outcome Measures. We report estimates of equation 8 in Table 3.

[39] The normalization adjusts for level differences in $\sigma_{p_{ct-2}}$ and σ_{R_c} by contracts, and gives a standardized measure of premium differences comparable across contracts.

[40] We capture premium differences over diabetes prevalence rates with $diabiv_{ct-2} = Corr(p_{ct-2}, D_c)$, where D_c is the vector of baseline diabetes prevalence rates across counties in contract c. Similarly, we use $hyptiv_{ct-2} = Corr(p_{ct-2}, H_c)$ to capture premium differences across hypertension, where H_c is the vector of baseline hypertension prevalence rates in contract c.

The OLS estimates do not indicate significant effects of risk scores on the HEDIS outcomes. Based on these estimates, health outcomes improved by 1.8 percentage points in high-rated contracts (column 2) and by 1.68 percentage points in high-selection contracts (column 4). However, risk selection can bias the OLS estimates and the implied improvements in health outcomes.

Table 3: Effect of selection on the HEDIS outcome

	(I)	(II)	(III)	(V)	(VI)
			Panel A: OLS		
Risk Score	-0.29	-19.20	-6.33	-38.84	-73.83*
	(10.10)	(17.02)	(20.77)	(25.34)	(36.63)
$\gamma_c \cdot$ Post	1.96	1.81	1.27	1.68	1.42
			Panel B: TSLS		
Risk Score		-93.28*		-94.09***	-160.57**
		(53.70)		(35.71)	(65.29)
First-stage F-stat	2.00	9.12	3.54	10.09	26.35
Over-id p-value	–	0.29	–	0.39	0.13
$\gamma_c \cdot$ Post		1.07		0.24	-1.03
ΔRisk $\cdot \widehat{\beta_{TSLS}}$		1.31		1.79	3.85
Contracts	low	high	high	high	high
Service area risk			>50%	≤50%	≤25%
y mean	65.66	71.04	71.85	70.37	64.51
N	1,946	669	413	228	116

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

Notes: The table shows the effect of risk scores on the HEDIS outcomes. HEDIS outcomes of a contract are measured by the percentage of enrollees who have controlled their chronic conditions (i.e., by testing below the medical thresholds). Panel A shows OLS estimates regressing HEDIS outcomes on contract risk scores. Panel B shows two-stage-least-squares (TSLS) estimates instrumenting contract risk scores by the premium differences across counties.

To correct for this endogeneity, we show two-stage-least-squares (TSLS) estimates in Panel B where we instrument risk scores, $risk_{ct-2}$, with the premium differences over

county risk scores, diabetes prevalence rates, and hypertension prevalence rates. The premium differences significantly predict risk scores in high-rated contracts (column 2) and particularly in high-selection contracts (column 4-5).[41] For these contracts, we find significant and negative effects of risk scores on the outcome measures, with a ten percentage point increase in risk score lowering health outcome measures by 9 percentage points in high-rated contracts.[42]

Applying the TSLS estimates, we decompose the gains in the health outcome measures into a selection component and a component reflecting the health gains of a standard-risk enrollee. We calculate the selection component using $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}}$, where ΔRisk is the risk score change relative to low-rated contracts after the payment reform. In high-selection contracts, ΔRisk is 1.9 percentage points, and the selection increased health outcome measures by $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}} = 1.79$ percentage points.[43] We infer the health gains of a standard-risk enrollee from estimates of $\gamma_c \cdot post$ in equation 8. Adjusted for risk, health outcomes improved by $\gamma_c \cdot post = 0.24$ percentage points on average in high-selection contracts. Compared to the 1.68 percentage point increase in health outcome measures (panel a), selection of healthier enrollees accounted for 86% of the health measure gains in high-selection contracts.[44]

5.3 Quantifying Risk Selection in the Bonus Payments

Selection Gains in the Star Rating. We apply the IV strategy to quantify the selection gains in the overall star rating and the bonus payments. In Appendix Table D29, we group measures by their weights in the overall rating and quantify the selection gains for the

[41] We show the first-stage results in Appendix Table D28. We explore alternative combinations of instruments for the health outcome rating in Appendix Table D31.

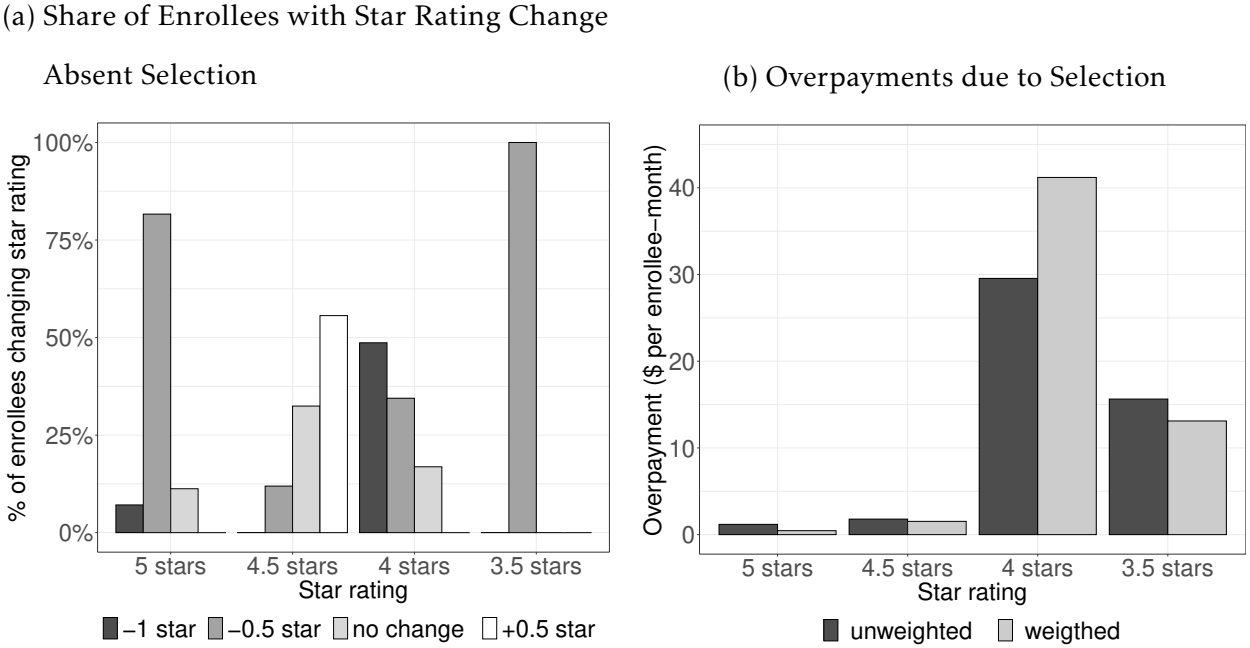
[42] To give a sense of the magnitude, a 9 percentage point increase in health outcomes roughly closes 56% of the health outcome gap between the 15th and 85th percentiles of risk scores in high-rated contracts.

[43] Specifically, $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}} = -0.019 \cdot (-94.09) = 1.79$. ΔRisk is the event study coefficient for year 2011-2012 in the contract-level analysis of risk scores (panel b of Figure 3).

[44] Risk-adjusted health improvements explain $\frac{0.24}{1.68} = 14.3\%$ of the health measure gains, and selection explains $1 - \frac{0.24}{1.68} = 85.7\%$.

outcome rating (measures with 3.0 weights), access rating (1.5 weights), and the process rating (1.0 weights). Risk selection impacts significantly the outcome rating (column 1-2), explaining nearly 80% of the rating gains to high-selection contracts.[45] Access and process ratings are not affected by risk scores. Together, these estimates suggest that risk selection increased the overall rating by 0.23 stars among high-selection contracts.[46]

Figure 4: Effects of selection on the quality rating and overpayments



Notes: The figure shows the effect of adjusting risk selection on the overall star ratings of high-selection contracts in panel (a) and on the payments to these contracts in panel (b). For different star ratings in 2014 (horizontal axis), panel (a) shows the percentage of enrollees receiving lower (by 1 star or 0.5 star) or higher (by 0.5 star or unchanged) star ratings upon adjustment for selected risk scores. Based on the changes in panel (a), panel (b) shows changes in 2015 payments by the 2014 star rating. We assume that contracts receiving a downgrade (upgrade) in the star rating adjust bids downward (upward) relative to the new benchmarks such that rebates to enrollees remain unchanged.

We then remove the selection gains in the outcome rating to construct risk-adjusted

[45]The selection effect is comparable to but different from the 86% calculated in Section 5.2 because, 1) we look at ratings on a scale of 2 to 5 stars in this section rather than the raw statistic in each measure, and 2) we group together all measures receiving 3.0 weights and focus on the average rating across measures as the dependent variable. In Section 5.2 we focused only on the three HEDIS measures.

[46]Due to the weighting across measures, the selection gains on the overall rating is half the selection gains on the outcome rating ($50\% \cdot 0.45 = 0.23$ stars). Across health outcome measures, HEDIS outcomes are most sensitive to risk scores (Appendix Table D30), followed by drug related outcomes. These measures explain all of the selection gains in the outcome rating.

star ratings for high-selection contracts in 2014.[47] Figure 4 compares the risk-adjusted rating with the original rating in panel (a), grouping contracts by the original rating on the horizontal axis. The vertical axis reports the percent of enrollees with a contract's rating change after the adjustment. Around 80% of the enrollees in 4.0-star contracts would receive a lower rating for their contracts after risk adjustment. In the 3.5-4.0 star range, 98% of enrollees are in marginal high-rated contracts with continuous ratings below 4.0 stars in 2014. Even small increases in the risk score could downgrade their contracts to 3.5 stars and below. By contrast, risk adjustment has smaller impacts in the 4.5-5.0 range, where all contracts maintain at least a 4.0-star rating after risk adjustment.

Selection and Payments. We next quantify the impact of risk adjustment on the bonus payments to insurers.[48] We determine payments under the risk-adjusted rating assuming that insurers adjust bids to match the higher benchmarks, for which we find empirical support in Appendix Table D32.[49] We then infer counterfactual bids by inverting equation 1 holding rebates at the pre-adjustment level.[50]

We compare payments before and after risk adjustment in panel (b) of Figure 4. We interpret the difference as overpayments that rewarded the selection gains in the star ratings rather than actual improvements in quality. Overpayments are largest in the 3.5-4.0 star range, with average bonus payment gains of $25 from selection. Absent selection, most of these contracts would be rated 3.5 stars or below (panel a) and hence ineligible for bonus payments under the ACA model. We find smaller overpayments ($0.59 per contract) in the 4.5-4.0 star range where the risk-adjusted rating remains 4.0 stars and above for all

[47] Recall that the overall rating is based on a weighted average of measure ratings. Here, we subtract the selection gains in the outcome rating and round the new average rating to the nearest half star to construct the risk-adjusted rating.

[48] Since the 2014 star rating determines payments in 2015, we re-calculate 2015 payments using the risk-adjusted rating and payment rules in Table D1.

[49] In particular, high-selection contracts submitted higher bids after the introduction of benchmark bonuses, narrowed the distance between bids and benchmarks, but did not significantly increase the rebates to enrollees. Appendix Figure E14 plots the raw trends of the bidding adjustment.

[50] For insurers bidding above the benchmark, the payment is the benchmark at the new star rating.

contracts.[51]

To assess magnitude, we compare the overpayments with the benchmark bonus to high-selection contracts. Relative to low-rated contracts, high-selection contracts received $87 more in bonus-adjusted benchmarks in 2015.[52] Selection increased bonus payments to high-selection contracts by $12, or 14% of the benchmark bonus in 2015. For marginally high-rated contracts (3.5-4.0 stars), selection increased payments by 29% of the benchmark bonus, or by $59.8 million in 2015 alone.[53]

5.4 Implications for Risk Adjustment

Our results suggest that the star rating is a poor indicator of the health benefits of insurance due to insurer selection. CMS's standard way to neutralize selection is to risk-adjust per capita payments so that each enrollee is *predicted* to be equally profitable (e.g., Van De Ven and Ellis, 2000). However, for health outcome measures in the star rating, an insurer's marginal payment does not exclusively depend on the health status of the marginal enrollee as in MA, but also on that of all its enrollees. To provide a concrete example, the measure *Diabetes Care – Controlling Blood Sugar* is measured by the fraction of diabetic enrollees with hemoglobin A1c level below 9%. Enrolling healthier individuals can improve the overall test results and increase bonus payments for all enrollees in the contract. Because the star rating examines a wide range of health outcomes, we advocate risk-adjusting the rating at the level of individual measures while accounting for enrollee heterogeneity in risk types.

We propose a simple risk-adjustment design in Appendix A.3 to counter the selection incentive. The adjustment predicts health outcomes for different risk types and assesses

[51] We find nearly identical results for overpayments when we calculate the risk score change for each contract using the synthetic control method (Appendix Figure E13).

[52] The estimate is the event study coefficient for 2015 in an extended analysis of contract benchmarks using the difference-in-differences model in equation 4.

[53] Specifically, $25/$87 = 29%. Scaled by the enrollment-months in high-selection contracts, overpayments amounted to $68.5 million annually in 2015, with $59.8 m concentrated in marginal high-rated contracts.

health improvements relative to the predicted outcome as the threshold. The patient-specific threshold neutralizes the gains from selecting healthier individuals, re-directing efforts to improving the health of enrollees. In practice, the prediction could be implemented joining the HEDIS outcome data with diagnoses in the claims data. However, prediction errors could in turn become a source of selection that distorts payments and contract design (Brown *et al.* 2014, Lavetti and Simon 2018, Geruso *et al.* 2019, Carey 2017).[54] The imperfections call for more sophisticated models to fully counter the selection incentive.

Because of the geographic price discrimination we detect in this paper, we propose that adjustments be fine-tuned for contracts with similar risk exposure in the service area. This could be important if contracts serving riskier counties are exposed to consumers systematically different from those in healthier counties. In this case, stratifying risk adjustmens across quantiles of risk scores could reduce selection for contracts serving similar types of consumers. Following recent proposals by the Medicare Payment Advisory Commission (2019), we suggest stratifying adjustments also for additional risk factors.[55] While implementing the proposed adjustments creates additional costs and bureaucratic burden, several HEDIS measures are also used in other value-based programs (e.g., the California Cal MediConnect, California Department of Health Care Services, 2018), which may reduce the cost of setting up a new data infrastructure while increasing its value.

6 Distributional Impacts Across Counties

We next explore the impacts of selection on the spatial distribution of high-rated insurance across counties. Specifically, we compare the market share of high-rated insurance across counties with different enrollee risk scores in 2009-2010. Because high-rated insurers decreased (increased) premiums in healthier (riskier) counties after the payment

[54] We discuss how imperfect risk adjustments on both the quality rating and benchmarks could impact selection in Appendix A.3.3.

[55] Recent studies show that stratified risk adjustment can reduce the bias in quality measurement for both hospitals (McCarthy *et al.*, 2019) and MA plans (Durfey *et al.*, 2018).

reform, the enrollment in high-rated insurance may diverge across counties with the healthiest and the riskiest enrollees. We examine this hypothesis estimating the distribution of high-rated contracts across county risk scores through

$$y_{clt} = \beta_0 \cdot risk_l \cdot high_c \cdot post_t + \beta_1 \cdot risk_l \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta_3 \cdot high_c \cdot risk_l \quad (10)$$
$$+ \beta \cdot X_{lt} + \alpha_c + \gamma_l + \tau_t + \epsilon_{clt},$$

where y_{clt} is the market share of contract c in county l and year t. The key independent variable $risk_l$ is the baseline fee-for-service risk score in county l. Therefore, we examine the changes in market share y_{clt} as risk scores increase from the healthiest to the riskiest counties in $risk_l$.[56] We control for the baseline differences across county risk scores in β_3, the growth of market shares across counties in β_1, and the growth of high-rated insurance in β_2. β_0 estimates the differential growth of high-rated insurance across county risk scores.

We plot the raw trends of market shares and the event study estimates from equation 10 in Appendix Figure E15. Market shares raw trends diverged markedly across the 15% healthiest and riskiest counties. At the contract level (panel a), market shares of high-rated contracts increased in the healthiest counties (gray lines) and decreased in the riskiest counties (blue lines). Panel (c) finds a similar divergence looking at the overall market shares of high- and low-rated contracts across risk tails (solid lines). Across all counties, Appendix Table D33 estimates that a 10 percentage point increase in risk scores lowered the market share of high-rated contracts by 11.8 percentage points (column 2), or by 9 percentage points more compared to low-rated contracts (column 3).[57] We examine robustness and estimate the distributional effects on premiums in Appendix H.

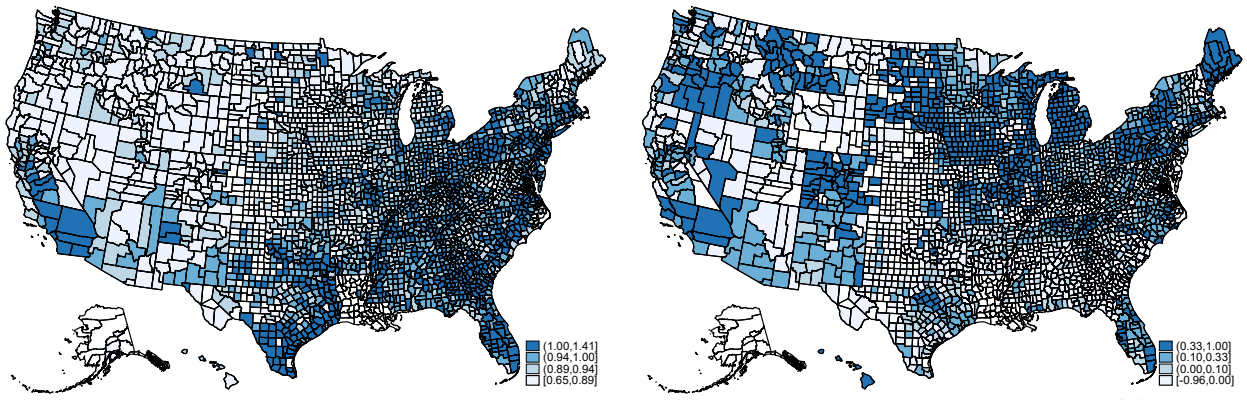
[56] This is different from equation 5 where the $risk_{cl}$ variable also depends on contract c. Here, we use the cross-county differences in $risk_l$ to examine the distribution of high-rated insurance across space. X_{lt} controls for the same set of county variables in equation 5.

[57] To validate the magnitude of the selection response, we estimate the premium elasticity for high-rated contracts using an IV strategy similar to that in Section 5.2. Appendix G finds premium elasticities similar to those in related papers (e.g., Lucarelli *et al.*, 2012, Decarolis *et al.*, 2020, Starc and Town, 2020).

Figure 5 illustrates the growth of high-rated insurance across US counties in 2009-2014. High-rated market shares increased the most in the healthiest counties in the North West and the South West and generally decreased in the riskier counties in the South. The median increase was 16% in the healthiest counties compared to 5% in the riskiest ones (Appendix Figure H4).

Figure 5: Distribution of county risk scores and changes in the high-rated market share

(a) County risk scores (b) Changes in the high-rated market share



Notes: The figure plots the cross-county distribution of fee-for-service risk scores in panel (a) and the changes in the market share of high-rated insurance in panel (b). We calculate the market share of high-rated insurance based on the contract's contemporaneous rating (4.0 stars and above), and plot the increase in 2013-2014 market shares compared to the baseline in 2009-2010 across counties. Lighter blues in panel (a) indicate lower county risk scores whereas darker blues in panel (b) indicate larger increases in the high-rated market share. Gradients of colors indicate the inter-quartile range of the outcome variable in each panel.

The next section discusses the implications of inequality in perspective of CMS's efforts to improve quality of service and consumer information among its value based programs.

7 Discussion and Conclusion

This paper studies how suppliers respond to regulations linking government subsidies to the quality of social services, and its implications for consumers. Our results on the Medicare Advantage market suggest that policy efforts to improve the quality of social services can have unintended consequences on enrollees' access to the same services.

The previous section highlights the shifts in the spatial distribution of high-rated contracts that resulted from insurer selecting favorable consumer types. Because the payment reform did not affect consumers' information of the star rating or their preferences for quality, their choice of insurance products would have followed parallel trends after the payment reform in healthy and risky counties. The differential growth of high-rated insurance in the healthiest counties, therefore, is consistent with insurer selection incentivized by the payment model.[58] Selection resulted in lower premiums and better access to high-rated insurance among the healthiest counties in Medicare but worsened the access to high-rated insurance in the riskiest counties. While establishing the policy impacts on aggregate welfare would require strong assumptions on consumer preferences, our results suggest that the selection behavior of individual contracts impacts the distribution of insurance coverage across space, hurting the more vulnerable beneficiaries in the adversely affected locations.

More broadly, firm responses could reduce welfare for all Medicare beneficiaries in counties facing higher premiums. For consumers currently with low demand for MA insurance, the premium increase would reduce their access to high-rated managed care when they transition into a state where effective management of chronic conditions is more desirable. Thus, in addition to weakening the risk protection of current enrollees, selection lowers the option value of insurance for all beneficiaries in riskier counties as their chronic conditions evolve over time (Handel *et al.*, 2015).

Selection also worsened the information quality of the star rating, with broader implications for consumer choice. As we show in panel (a) of Figure 4, risk-adjusted star ratings would be lower for over 50% of the enrollees in contracts with a 4.0-star rating. The bias in the star rating may complicate plan choice in a market already fraught with sub-optimalities and inconsistencies (Abaluck and Gruber, 2011). Instead, better informa-

[58] This interpretation is also supported by our finding that the availability of high-rated plans did not decrease across county-risk (Appendix Figure H5), which implies that consumers responded to price discrimination rather than plan availability.

tion quality and more personalized delivery of information can result in better choices and consumer wellbeing (Kling *et al.*, 2012, Gruber *et al.*, 2020).

Providing consumer information was the first aim of the star rating. To improve the star rating, the Medicare Payment Advisory Commission (MedPAC) recently proposed sweeping changes to the way the quality measures are collected (Medicare Payment Advisory Commission, 2019). First, reducing the number of submeasures and basing them on contract-specific predetermined prospective standards may decrease the uncertainty related to quality investments while making a contract's star rating more comparable across counties. Second, as we also propose in Section 5.4, MedPAC views stratified risk adjustments as a necessary step to counter poor informativeness. By showing that risk selection exacerbates the poor information content of the star rating, our paper provides an additional rationale for CMS to consider MedPAC policy proposals.

Going forward, we believe that our results are of importance also for other pay-for-performance insurance programs both directly and indirectly. First, CMS's other value based programs are based on the same data sources as those we study in this paper (e.g., the HEDIS), and thus our results could directly apply there (e.g., Cal MediConnect). Second, it is natural for legislators to base performance payments on already existing measurement requirements. Despite being readily available, these performance measures may not be fully consistent with the aim of the policy and could result in unintended consequences.

References

Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, **105** (490), 493–505.

Abaluck, J., Bravo, M. M. C., Hull, P. and Starc, A. (2020). *Mortality effects and choice across private health insurance plans*. Tech. rep., National Bureau of Economic Research.

— and Gruber, J. (2011). Choice inconsistencies among the elderly: evidence from plan choice in the medicare part d program. *American Economic Review*, **101** (4), 1180–1210.

AHRQ (2017). Analyzing CAHPS survey data. https://www.ahrq.gov/cahps/surveys-guidance/helpful-resources/analysis/index.html, Agency for Healthcare Research and Quality, Accessed: 2020-05-10.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, **74** (2), 431–497.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, **100** (3), 598–614.

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.

Biasi, B. (2018). The labor market for teachers under different pay schemes. *Mimeo, Yale University*.

Breyer, F., Bundorf, M. K. and Pauly, M. V. (2011). Health care spending risk, health insurance, and payment to health plans. In *Handbook of Health Economics*, vol. 2, Elsevier, pp. 691–762.

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.

Burgess, S., Propper, C., Ratto, M., Tominey, E. *et al.* (2017). Incentives in the public sector: Evidence from a government agency. *Economic Journal*, **127** (605), 117–141.

California Department of Health Care Services (2018). Cal mediconnect performance dashboard metrics summary. Accessed Apr. 21, 2021.

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

Charbi, A. (2020). The fault in our stars! quality reporting, bonus payments and welfare in medicare advantage. *Mimeo, University of Texas at Austin*.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A. and Cutler, D. (2016). The association between income and life expectancy in the United States, 2001-2014. *Jama*, **315** (16), 1750–1766.

Chetverikov, D., Larsen, B. and Palmer, C. (2016). Iv quantile regression for group-level treatments, with an application to the distributional effects of trade. *Econometrica*, **84** (2), 809–833.

CMS (2018). NHE fact sheet. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html, accessed: 2020-08-28.

Cooper, Z., Craig, S. V., Gaynor, M. and Van Reenen, J. (2018). The price ain't right? hospital prices and health spending on the privately insured. *The Quarterly Journal of Economics*, **134** (1), 51–107.

Currie, J. and Schwandt, H. (2016). Mortality inequality: the good news from a county-level approach. *Journal of Economic Perspectives*, **30** (2), 29–52.

Curto, V., Einav, L., Levin, J. and Bhattacharya, J. (2019). Can health insurance competition work? Evidence from Medicare Advantage, Mimeo, Harvard University, Stanford University, Stanford University, Stanford University.

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*, **56**, 383–396.

—, Polyakova, M. and Ryan, S. P. (2020). Subsidy design in privately provided social insurance: Lessons from medicare part d. *Journal of Political Economy*, **128** (5), 1712–1752.

Durfey, S. N., Kind, A. J., Gutman, R., Monteiro, K., Buckingham, W. R., DuGoff, E. H. and Trivedi, A. N. (2018). Impact of risk adjustment for socioeconomic status on medicare advantage plan quality rankings. *Health Affairs*, **37** (7), 1065–1072.

Eggleston, K. (2005). Multitasking and mixed systems for provider payment. *Journal of health economics*, **24** (1), 211–223.

Ellis, R. P. and McGuire, T. G. (2007). Predictability and predictiveness in health care spending. *Journal of health economics*, **26** (1), 25–48.

Finkelstein, A., Gentzkow, M., Hull, P. and Williams, H. (2017). Adjusting risk adjustment–accounting for variation in diagnostic intensity. *The New England Journal of Medicine*, **376** (7), 608.

—, — and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, **131** (4), 1681–1726.

—, — and Williams, H. L. (2019). *Place-based drivers of mortality: Evidence from migration*. Tech. rep., National Bureau of Economic Research, National Bureau of Economic Research, No. w25975.

Geruso, M., Layton, T. and Prinz, D. (2019). Screening in contract design: evidence from the ACA health insurance Exchanges. *American Economic Journal: Economic Policy*, **11** (2), 64–107.

Glazer, J. and McGuire, T. G. (2000). Optimal risk adjustment in markets with adverse selection: An application to managed care. *American Economic Review*, **90** (4), 1055–1071.

Gravelle, H., Sutton, M., Ma, A. *et al.* (2010). Doctor behaviour under a pay for performance contract: Treating, cheating and case finding? *Economic Journal*, **120** (542), 129–156.

Gruber, J., Handel, B., Kina, S. and Kolstad, J. (2020). Managing intelligence: Skilled experts and ai in markets for complex products. *NBER Working Paper*, *w27038*.

Gupta, A. (2021). Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, **111** (4), 1241–83.

Handel, B., Hendel, I. and Whinston, M. D. (2015). Equilibria in health exchanges: Adverse selection versus reclassification risk. *Econometrica*, **83** (4), 1261–1313.

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

Khan, A. Q., Khwaja, A. I. and Olken, B. A. (2015). Tax farming redux: Experimental evidence on performance pay for tax collectors. *The Quarterly Journal of Economics*, **131** (1), 219–271.

Kling, J. R., Mullainathan, S., Shafir, E., Vermeulen, L. C. and Wrobel, M. V. (2012). Comparison friction: Experimental evidence from medicare drug plans. *The quarterly journal of economics*, **127** (1), 199–235.

Lavetti, K. and Simon, K. (2018). Strategic formulary design in Medicare Part D plans. *American Economic Journal: Economic Policy*, **10** (3), 154–92.

Layton, T. J. (2017). Imperfect risk adjustment, risk preferences, and sorting in competitive health insurance markets. *Journal of health economics*, **56**, 259–280.

— 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.

Lucarelli, C., Prince, J. and Simon, K. (2012). The welfare impact of reducing choice in medicare part d: A comparison of two regulation strategies. *International Economic Review*, **53** (4), 1155–1177.

Mainous III, A. G. and Talbert, J. (1998). Assessing quality of care via hedis 3.0: is there a better way? *Archives of family medicine*, **7** (5), 410.

McCarthy, C. P., Vaduganathan, M., Patel, K. V., Lalani, H. S., Ayers, C., Bhatt, D. L., Januzzi, J. L., de Lemos, J. A., Yancy, C., Fonarow, G. C. *et al.* (2019). Association of the new peer group–stratified method with the reclassification of penalty status in the hospital readmission reduction program. *JAMA network open*, **2** (4), e192987–e192987.

Medicare Payment Advisory Commission (2019). Chapter 8: Redesigning the medicare advantage quality bonus program. In: Report to the congress: Medicare and the health care delivery system. Accessed Apr. 21, 2021.

Mullen, K. J., Frank, R. G. and Rosenthal, M. B. (2010). Can you get what you pay for? pay-for-performance and the quality of healthcare providers. *The Rand Journal of Economics*, **41** (1), 64–91.

Newhouse, J. P., Buntin, M. B. and Chapman, J. D. (1997). Risk adjustment and medicare: taking a closer look. *Health Affairs*, **16** (5), 26–43.

—, Price, M., Hsu, J., McWilliams, J. M. and McGuire, T. G. (2015). How much favorable

selection is left in Medicare Advantage? *American Journal of Health Economics*, **1** (1), 1–26.

Nichols, G. A., Raebel, M. A., Dyer, W. and Schmittdiel, J. A. (2018). The effect of age and comorbidities on the association between the Medicare STAR oral antihyperglycemic adherence metric and glycemic control. *Journal of Managed Care & Specialty Pharmacy*, **24** (9), 856–861.

Rosenthal, M. B. and Frank, R. G. (2006). What is the empirical basis for paying for quality in health care? *Medical Care Research and Review*, **63** (2), 135–157.

Safford, M. M., Brimacombe, M., Zhang, Q., Rajan, M., Xie, M., Thompson, W., Kolassa, J., Maney, M. and Pogach, L. (2009). Patient complexity in quality comparisons for glycemic control: An observational study. *Implementation Science*, **4** (1), 2.

Shen, Y. (2003). Selection incentives in a performance-based contracting system. *Health Services Research*, **38** (2), 535–552.

Sherry, T. B. (2016). A note on the comparative statics of pay-for-performance in health care. *Health economics*, **25** (5), 637–644.

Skinner, J. (2011). Causes and consequences of regional variations in health care. In *Handbook of Health Economics*, vol. 2, Elsevier, pp. 45–93.

Starc, A. and Town, R. J. (2020). Externalities and benefit design in health insurance. *The Review of Economic Studies*, **87** (6), 2827–2858.

Van De Ven, W. P. and Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. In *Handbook of health economics*, vol. 1, Elsevier, pp. 755–845.

Veiga, A. and Weyl, E. G. (2016). Product design in selection markets. *The Quarterly Journal of Economics*, **131** (2), 1007–1056.

Performance Pay in Insurance Markets: Evidence from Medicare

Michele Fioretti[1] **Hongming Wang**[2]

Online Appendix

[1]Department of Economics, Sciences Po. email: michele.fioretti@sciencespo.fr
[2]Center for Global Economic Systems, Hitotsubashi University. email: hongming.wang@r.hit-u.ac.jp

Online Appendix

Table of Contents

A Theoretical Section

A.1 A Monopolistic Model of Risk Selection

A monopolist insurer sells Medicare Advantage (MA) insurance in two counties. The insurer's revenues depend on the premiums it charges in the two counties (p_1 and p_2) and on the county benchmark, B. Under pay-for-performance, B increases with the insurer's quality rating, q. The demand for the contract in county $l = \{1, 2\}$ depends on the premium p_l charged in the county according to $s_l = s_l(p_l)$.[3] The insurer average risk score is a function of the risk score of the enrollees in the two counties. Therefore, it depends on p_1 and p_2 as in $r(p_1, p_2)$. The insurer can increase the quality rating either through investments which incur a marginal cost c, or through risk selection which lowers the risk score. The insurer's pricing problem is to maximizes total profits $\sum_{l=1}^{2}(p_l + B - c) \cdot s_l$ by choosing p_1 and p_2.[4]

To illustrate the selection incentive due to a biased rating, we examine the case where the insurer selects healthier enrollees with premiums p_l, but does not increase costly investments c.[5] We assume *perfect risk adjustments on benchmarks and prices* so that the insurer is fully compensated for the health costs of enrollees. In this world, risk selection would have no bearing on the insurer's profit absent the linkage with quality and bonus payments. From the first order conditions of the insurer problem, the optimal premium in county $l = \{1, 2\}$ solves

$$p_l = c - B + \left(1 + \frac{dB}{dq} \cdot \frac{\partial q}{\partial r} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_l + s_{-l}}{s_l}\right) \cdot |\varepsilon_l|^{-1}, \tag{A1}$$

[3] We assume that demand is responsive to changes in premium p_l, but not responsive to changes in the quality score q. We make this simplifying assumption because premium is the main lever of selection across markets, whereas the quality rating does not vary across markets. Empirically, Darden and McCarthy (2015) provides supporting evidence that the demand response to the star rating is fairly weak.

[4] To calculate insurer revenue, we follow Curto *et al.* (2019) and express premium p_l as the "excess bid," or the difference between payments to the insurer and the benchmark B.

[5] In practice, insurers adjust both investments and premiums to improve the quality rating. However, as we show in the main text, premium responses are driven by the bias in the quality rating. Endogenizing investment in quality c does not affect the qualitative predictions on premiums.

where ε_l is the semi-elasticity of demand to premium in county l.

Before the payment reform, $\frac{dB}{dq}$ is zero and the optimal premium equals marginal cost plus a mark-up, which is inverse to demand semi-elasticity. After the reform, equation A1 shows that premiums also respond to the selection incentive due to a biased quality rating through the term

$$\Delta p_l \equiv p_l^{post} - p_l^{pre} = \frac{dB}{dq} \cdot \frac{\partial q}{\partial r} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_l + s_{-l}}{s_l} \cdot |\varepsilon_l|^{-1}. \tag{A2}$$

The selection term is switched on when $\frac{\partial q}{\partial r} \neq 0$. In what follows, we examine the case where risk score r biases the quality rating downward, i.e., $\frac{\partial q}{\partial r} < 0$.

The selection incentive affects premiums more in counties with larger $\frac{\partial r}{\partial p_l}$. Specifically, given $\frac{\partial q}{\partial r} < 0$, the premium will be lower in county one relative to county two if, other things equal, county one is more conducive to risk selection (i.e., $\frac{\partial r}{\partial p_l} > \frac{\partial r}{\partial p_{-l}}$). Relative to the pre-reform levels, premiums will drop ($\Delta p_l < 0$) in counties with $\frac{\partial r}{\partial p_l} > 0$, where lower premiums decrease the risk score. These price responses will decrease the risk score after the payment reform according to

$$\Delta r \equiv r^{post} - r^{pre} = \Delta p_l \frac{\partial r}{\partial p_l} + \Delta p_{-l} \frac{\partial r}{\partial p_{-l}} < 0.$$

The difference in the price change across counties can be shown to depend on the difference in the fee-for-service risk score Γ_l^{FFS} across counties.[6] To relate the price change Δp_l to the fee-for-service risk score Γ_l^{FFS}, we focus on the term $\frac{\partial r}{\partial p_l}$. The term gives the responsiveness of contract-level risk score r to a small price change in county l. The contract risk score in turn depends on the weighted average of enrollee risk scores from

[6] Medicare enrollees who did not purchase a Medicare Advantage plan are automatically enrolled in the fee-for-service program. The average risk of these enrollees is the fee-for-service risk score.

the two counties. Specifically,

$$
\begin{aligned}
r &= \frac{s_l \cdot \Gamma_l^{MA} + s_{-l} \cdot \Gamma_{-l}^{MA}}{s_l + s_{-l}} \\
&= \frac{\bar{\Gamma}_l - (1 - s_l) \cdot \Gamma_l^{FFS} + \bar{\Gamma}_{-l} - (1 - s_{-l}) \cdot \Gamma_{-l}^{FFS}}{s_l + s_{-l}},
\end{aligned} \tag{A3}
$$

where Γ_l^{MA} is the average risk score of enrollees in the MA contract in county l. Γ_l is the average risk of all consumers in county l. Equation A3 therefore expresses contract risk score r in terms of enrollment share s_l and the average fee-for-service risk Γ_l^{FFS} in each county, exploiting the fact that $\bar{\Gamma}_l = s_l \cdot \Gamma_l^{MA} + (1 - s_l) \cdot \Gamma_l^{FFS}$. Taking derivative of equation A3 w.r.t. premium in county l yields

$$
\frac{\partial r}{\partial p_l} = \frac{s_l'}{s} \cdot \left(\frac{1 + s_{-l}}{s} \cdot \Gamma_l^{FFS} + \frac{1 - s_{-l}}{s} \cdot \Gamma_{-l}^{FFS} - \frac{\bar{\Gamma}_l + \bar{\Gamma}_{-l}}{s} \right) - \frac{1 - s_l}{s} \cdot \frac{\partial \Gamma_l^{FFS}}{\partial p_l}, \tag{A4}
$$

where $s = s_l + s_{-l}$. The first bracket in equation A4 captures the cross-county composition effect on the contract risk score. A small increase in p_l lowers enrollment by $s_l' = \frac{\partial s_l}{\partial p_l}$, increasing the *relative* enrollment from the other county, $-l$. Contract risk score r decreases more at lower enrollee risk scores in the other county, $-l$. In county l, the price change affects both the market share s_l and the relative enrollment, allowing enrollee risk scores to have larger impacts on r. In both counties, enrollee risk scores are negatively related to Γ_l^{FFS}, and the relationship is exact up to a marginal term $\frac{\partial \Gamma_l^{FFS}}{\partial p_l}$.

Equation A4 states that for a similar enrollment response, a premium response in county l can more effectively decrease r if the fee-for-service risk score is lower in county l. To induce the enrollment response, premiums need to adjust more in counties with smaller demand elasticity (equation A2). Substituting equation A4 into equation A2 nets out the semi-elasticity term, $\varepsilon_l = s_l'/s_l$. The resulting price change Δp_l relative to the

pre-reform level is given by

$$\Delta p_l = -\frac{d\,B}{d\,q}\frac{\partial q}{\partial r}\cdot\left(\frac{1+s_{-l}}{s}\cdot\Gamma_l^{FFS}+\frac{1-s_{-l}}{s}\cdot\Gamma_{-l}^{FFS}-\frac{\bar{\Gamma}_l+\bar{\Gamma}_{-l}}{s}-\frac{1-s_l}{s_l'}\cdot\frac{\partial\Gamma_l^{FFS}}{\partial p_l}\right),$$

where the terms in the parentheses on the right hand side are evaluated at pre-reform levels of prices, markets shares, and fee-for-service risk scores. Focusing on the differences by Γ_l^{FFS}, the relative price change between counties is

$$\Delta p_l - \Delta p_{-l} \propto -\frac{d\,B}{d\,q}\frac{\partial q}{\partial r}\cdot\left(\Gamma_l^{FFS}-\Gamma_{-l}^{FFS}\right). \tag{A5}$$

Equation A5 states that other things equal, premiums should increase more in counties with larger fee-for-service risk scores. On the other hand, the full equation for the relative price change is given by

$$\Delta p_l - \Delta p_{-l} = -\frac{d\,B}{d\,q}\frac{\partial q}{\partial r}\cdot\left(\Gamma_l^{FFS}-\Gamma_{-l}^{FFS}-\frac{1-s_l}{s_l'}\cdot\frac{\partial\Gamma_l^{FFS}}{\partial p_l}+\frac{1-s_{-l}}{s_{-l}'}\cdot\frac{\partial\Gamma_{-l}^{FFS}}{\partial p_{-l}}\right).$$

Compared to equation A5, the full equation also includes the difference in $\frac{1-s_l}{s_l'}\cdot\frac{\partial\Gamma_l^{FFS}}{\partial p_l}$ across counties. The additional terms are determined by the consumer characteristics in each county. Exploiting the fact that the payment reform is a supply-side regulation that did not affect consumers' knowledge of the quality rating or preferences, when evaluating $\frac{1-s_l}{s_l'}\cdot\frac{\partial\Gamma_l^{FFS}}{\partial p_l}$ at pre-reform prices and market shares, we absorb these terms using contract-county fixed effects. Controlling for consumer characteristics, equation A5 predicts that premiums increase more relative to the pre-reform levels in riskier counties. We hence examine heterogeneous responses across baseline fee-for-service risk scores in the empirical analysis.

A.2 Extension: Competition Across Insurers

Extending the monopoly model to allow for multiple competing insurers yields similar results on premiums. The main difference we find is that, all else equal, premium responses are weaker with greater competition between insurers.

For simplicity, we begin with the case of two insurers a and b setting premiums in county l and $-l$. Given premiums p_l^a, p_l^b, consumer demand for the contract of insurer $i = a, b$ is given by $s_l^i(p_l^a, p_l^b)$ in county l.[7] To maximize profit, insurer a sets premium p_l^a according to the following condition

$$p_l^a = c^a - B + \left(1 + \frac{dB}{dq^a} \cdot \frac{\partial q^a}{\partial r^a} \cdot \frac{\partial r^a}{\partial p_l^a} \cdot \frac{s_l^a + s_{-l}^a}{s_l^a}\right) \cdot |\varepsilon_l|^{-1}, \tag{A6}$$

where r^a is the risk score of enrollees in the contract offered by insurer a.

When risk score affects the contract's quality rating q^a ($\frac{\partial q^a}{\partial r^a} \neq 0$), insurers have incentives to select favorable risk types to boost the quality rating and hence payments. This selection incentive is captured in equation A6 in the term

$$\Delta p_l^a = \frac{dB}{dq^a} \cdot \frac{\partial q^a}{\partial r^a} \cdot \frac{\partial r^a}{\partial p_l^a} \cdot \frac{s_l^a + s_{-l}^a}{s_l^a} \cdot |\varepsilon_l|^{-1}.$$

Because the insurer attracts enrollees from both counties, the risk score of enrollees depends on the contract's exposure to risk types across counties. Specifically, enrollee risk score in contract a can be written as

$$\begin{aligned} r^a &= \frac{s_l^a \cdot \Gamma_l^a + s_{-l}^a \cdot \Gamma_{-l}^a}{s_l^a + s_{-l}^a} \\ &= \frac{\bar{\Gamma}_l - s_l^b \cdot \Gamma_l^b - (1 - s_l^a - s_l^b) \cdot \Gamma_l^{FFS} + \bar{\Gamma}_{-l} - s_{-l}^b \cdot \Gamma_{-l}^b - (1 - s_{-l}^a - s_{-l}^b) \cdot \Gamma_{-l}^{FFS}}{s_l^a + s_{-l}^a}, \end{aligned}$$

where $(\Gamma_l^i)_{i=a,b}$ is the average risk score of enrollees in contract i, and Γ_l^{FFS} is the risk score

[7] To simplify the analysis, we assume that the own-price elasticity of insurer i is greater in magnitude than the cross price elasticity of insurer $-i$ to a change in i's premium, or that $|s_l^{a'}| > |s_l^{b'}|, \forall l$.

of fee-for-service (FFS) enrollees in county l. The second equation applies the law of total probability to the population risk score $\bar{\Gamma}_l = s_l^a \cdot \Gamma_l^a + s_l^b \cdot \Gamma_l^b + (1 - s_l^a - s_l^b) \cdot \Gamma_l^{FFS}$, which sums over enrollees in contract a, contract b, and the FFS program.

A marginal increase in p_l^a changes the risk score in contract a as follows

$$\frac{\partial r^a}{\partial p_l^a} \propto \frac{s_l^{a'}}{s^a} \left[\left(\frac{s_l^{b'}}{s_l^{a'}} + \frac{1 + s_{-l}^a - s_l^b}{s^a} \right) \Gamma_l^{FFS} + \frac{1 - s_{-l}^a - s_{-l}^b}{s^a} \Gamma_{-l}^{FFS} \right], \tag{A7}$$

where $s_l^{i'} = \frac{\partial s_l^i}{\partial p_l^a}$ gives the enrollment response to an increase in p_l^i in county l for $i = \{a, b\}$. $s^a = s_l^a + s_{-l}^a$ is the total enrollment across counties for insurer a. Compared to equation A4, equation A7 states that the risk composition effect of a price change depends on the competitive forces between insurers, captured in the term $\frac{s_l^{b'}}{s_l^{a'}}$. The term has larger magnitude in markets where a price increase in contract a results in larger enrollment gains in the competing contract b. The selection incentive Δp_l^a can then be written as

$$\Delta p_l^a \equiv p_l^{a,post} - p_l^{a,pre} = \propto -\frac{dB}{dq^a} \cdot \frac{\partial q^a}{\partial r^a} \cdot \left[\left(\frac{s_l^{b'}}{s_l^{a'}} + \frac{1 + s_{-l}^a - s_l^b}{s^a} \right) \cdot \Gamma_l^{FFS} + \frac{1 - s_{-l}^a - s_{-l}^b}{s^a} \cdot \Gamma_{-l}^{FFS} \right],$$

and the relative selection incentive across counties is

$$\Delta p_l^a - \Delta p_{-l}^a \propto -\frac{dB}{dq^a} \cdot \frac{\partial q^a}{\partial r^a} \cdot \left(1 - \frac{s_l^{b'}}{|s_l^{a'}|} \right) \cdot \left(\Gamma_l^{FFS} - \Gamma_{-l}^{FFS} \right). \tag{A8}$$

Equation A8 states that insurer a faces greater incentives to increase premiums in county l if FFS enrollees are riskier in the county ($\Gamma_l^{FFS} > \Gamma_{-l}^{FFS}$). However, because increasing premium p_l^a also increases enrollments in competing contracts ($s_l^{b'} \geq 0$), the competition tends to weaken the premium responses across counties. Larger enrollment gains in the competitor's insurance imply weaker selection incentives compared to the monopoly case in equation A5, where the term $\frac{s_l^{b'}}{s_l^{a'}}$ is absent.

More generally, the selection incentive applies to cases with an arbitrary number of insurers. Let $\sum_{i \neq a} s_l^{i'}$ be the sum of enrollment gains in competing contracts after insurer a

increases premium p_l^a in county l.[8] The relative selection incentive across counties is given by

$$\Delta p_l^a - \Delta p_{-l}^a \propto -\frac{dB}{dq^a} \cdot \frac{\partial q^a}{\partial r^a} \cdot \left(1 - \frac{\sum_{i \neq a} s_l^{i'}}{|s_l^{a'}|}\right) \cdot \left(\Gamma_l^{FFS} - \Gamma_{-l}^{FFS}\right). \tag{A9}$$

Equation A9 predicts that insurer a would increase premiums more in counties with riskier FFS enrollees, and the magnitude of the response depends on the extent of selection bias in the quality rating $\frac{\partial q^a}{\partial r^a}$ and on the financial return to selection $\frac{dB}{dq^a}$. It further depends on market competition which may constrain the selection incentive through the term $\frac{\sum_{i \neq a} s_l^{i'}}{|s_l^{a'}|}$. We empirically examine the premium responses across county risk scores in Section 4.

A.3 Modeling the Selection Incentive in the Outcome Ratings

To illustrate how the outcome measures could generate the selection incentive, we explicitly model the dependence of the outcome ratings on enrollee risk scores by

$$q = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[1\left\{h(r_i, c) + \epsilon_i \leq h_0\right\}\right] = \frac{1}{N} \sum_{i=1}^{N} \Phi(r_i, c; h_0), \tag{A10}$$

where the health outcome $h(r_i, c)$ of enrollee i depends on her risk score r_i and the insurer's investment c.[9] The regulator sets an external threshold h_0 and determines the outcome rating q based on the share of enrollees whose health outcome $h(r_i, c)$ falls below the threshold h_0. Allowing for classical measurement error ϵ_i, the probability of enrollee i testing below the threshold h_0 is given by $\Phi(r_i, c; h_0)$, where Φ is the cumulative distribution function of ϵ_i. For a contract with N enrollees, outcome rating q is given by $\frac{1}{N}\sum_{i=1}^{N} \Phi(r_i, c; h_0)$, or the expected share of enrollees testing below the threshold.

Equation A10 implies that the insurer can manipulate the outcome rating with risk selection rather than investments. If healthier individuals are more likely to test below the threshold, then the insurer can improve the outcome rating with healthier enrollees

[8] We still assume for simplicity that the indirect gain to other contracts is smaller than the direct enrollment loss to insurer a. For instance, some enrollees may revert to traditional Medicare.

[9] The model can be extended to allow the distribution of ϵ_i to vary by enrollee characteristics.

without increasing investments c. Formally, enrolling an individual with risk score r_i on the margin updates the outcome rating q according to $\frac{\Phi(r_i, c; h_0)-q}{N+1}$, which in turn affects payment per enrollee according to

$$\frac{\Phi(r_i, c; h_0) - q}{N+1} \cdot \frac{dB}{dq}. \tag{A11}$$

Equation A11 states that an enrollee of risk score r_i can increase plan profits if the enrollee has better health outcomes than the average, or $\Phi(r_i, c; h_0) > q$. Thus, marginally healthier enrollees are more profitable to insurers, leading to lower risk scores and larger selection bias in the outcome rating.[10]

The next subsection extend this framework to consider (i) risk adjusting outcome measures, (ii) stratified risk adjustment, and (iii) the interaction between standard MA risk adjustment and the outcome-based risk adjustment at point (i).

A.3.1 Risk Adjustment

We suggest risk-adjusting the health outcome threshold h_0 as a potential solution to the selection incentive in $\Phi(r_i, c; h_0)$. The adjusted health outcome measures would compare outcome of enrollee i with the predicted outcome $\tilde{h}_0(r_i)$ based on her risk type r_i. If riskier enrollees have worse outcomes, $\tilde{h}_0(r_i)$ would adjust for this predictable difference by risk types and hence avoid penalizing insurers for enrolling risky individuals. To illustrate, let $\tilde{h}_0(r_i) = h_0^{FFS} + \beta^{FFS}(r_i - 1)$ be the predicted health outcome of risk type r_i. We assume that the prediction model is estimated using the sample of fee-for-service enrollees, and h_0^{FFS} is the outcome of the average enrollee whose risk score is normalized to 1. Thus, $\beta^{FFS}(r_i - 1)$ captures the predictable differences in outcomes when risk score r_i differs from the average enrollee.

To risk-adjust the outcome measures, we adjust the original threshold h_0 in equation

[10] For simplicity, we do not pursue the weighting of health outcome measures in the overall star rating q^* in this discussion. In addition, we assume for simplicity that selection improves ratings through healthier enrollees in the contract rather than the clustering of health outcomes across contracts.

A10 by the predictable differences across risk types. This effectively means setting $\tilde{h}_0(r_i) = h_0^{FFS} + \beta^{FFS}(r_i - 1)$ as a new threshold. The probability of risk type r_i testing below the threshold $\tilde{h}_0(r_i)$ is the follows

$$\tilde{\Phi}(r_i, c; h_0) = \Pr\Big\{\epsilon_i \leq h_0^{FFS} + \underbrace{\beta^{FFS}(r_i - 1)}_{\text{risk adjustment}} - h(r_i, c)\Big\}, \tag{A12}$$

where $\beta^{FFS}(r_i - 1)$ is the risk adjustment on thresholds. We assume that incurring costly investment c improves health outcomes by $i(c)$, which is a concave function increasing in c. The improved health outcome is $h(r_i, c) = h_0^{FFS} + \beta^{FFS}(r_i - 1) - i(c)$. It follows that adding the risk adjustment term in equation A12 offsets the selection gain of healthier enrollees in $h(r_i, c)$. The probability of testing below the adjusted threshold is simply $\tilde{\Phi}(c) = \Pr\{\epsilon_i \leq i(c)\}$, which does not differ across risk types but depends on investment c and the resulting health benefit $i(c)$. Therefore, the risk-adjusted rating rewards contracts for more effective management of chronic conditions rather than the selection of conditions.

A.3.2 Stratified Risk Adjustment

In equation A10, we assumed that the health outcomes of all enrollees in a contract are drawn from the same distribution underlying the random variable ϵ_i. In this case, the risk adjustment described in equation A12 can effectively reduce incentives to risk select individuals based on their risk score.

However, as we find in our empirical analyses, enrollee health statuses vary substantially across U.S. counties. Because the health outcome measures average across enrollees in all counties covered by the contract, the random variable ϵ_i may also depend on the location where individual i comes from, with some locations more equipped at delivering good outcome ratings due to better infrastructure or trust in the medical professionals. These broader, location-specific factors may be difficult to control for directly, but could impact health outcomes through the error term ϵ_i and hence bias the outcome rating.

The risk-adjustment term in equation A12 may not be sufficient to remove the resulting selection incentives.

Stratified risk-adjusting (or peer-group risk-adjusting) is a potential way forward and was already applied to star ratings in the Hospital Readmission Reduction Program (HRRP), another value-based initiative under the ACA. With stratification, health outcomes are compared within peer groups of hospitals treating similar types of patients (McCarthy *et al.*, 2019). In MA, stratified risk-adjustment implies computing separate coefficients β^{FFS} in equation A12 across the distribution of risk scores. While we abstract from the technical implementation, we note that our analyses suggests that β^{FFS} be computed based on quantiles of FFS risk scores in the service area. Thus, contracts serving relatively healthy and risky populations will see different updating coefficients for predicted outcomes, further limiting the profitability of risk selection on the predictable part of enrollee health.

A.3.3 Risk-Adjusting Benchmarks and Health Outcomes

The conceptual framework assumes for simplicity that the benchmark payments are perfectly risk adjusted based on enrollee risk scores. This has the double advantage of simplifying the model as well as highlighting that risk selection may arise from biased quality ratings even if the payments themselves are perfectly adjusted for risks. To understand why it has not been a standard practice to risk-adjust all health outcome measures, we note that health outcomes included in the star rating and MA risk adjustments have very different origins. For each enrollee, MA risk adjustment applies a risk factor that summarizes the total cost of treating the health conditions an enrollee suffers from. The health outcome measures instead focus only on enrollees with specific chronic conditions; while some of these enrollees have worse conditions or further complications, these differences are not accounted for in the health outcome measures. The lack of risk adjustment, especially for the HEDIS health outcomes, is partly due to the way these data are sourced. The HEDIS outcomes rely on bureaucratic and administrative data, which are not directly

interpretable in terms of value added or health improvements.[11] Risk-adjusting the health outcomes would require comprehensive records of health conditions and severity, which are available in the claims data rather than HEDIS.

Since the period between 2009 and 2014 shows no major changes in MA risk adjustment, in this paper we study risk adjustments on the health outcome measures independently from existing adjustments on benchmarks. In this section we extend the analysis in previous sections to explore the selection incentives when both the outcome ratings and benchmarks are adjusted for risk but neither adjustment is perfect. Specifically, we assume that enrollees differ in their health outcomes and costs due to differences in the risk type r_i and a non-risk characteristic ψ_i. For instance, ψ_i may indicate enrollees' socio-economic status (SES) such as poverty or dual eligibility for Medicaid. These characteristics may impact health outcomes and insurers' cost of covering the individual, but are not included in the adjustment model to predict either base payments (benchmarks) or the star ratings. The omission of ψ_i creates "prediction errors" in the risk adjustment model, allowing insurers to exploit the errors and select individuals cheaper than their predicted costs (e.g., Brown *et al.* 2014, Carey 2017, Lavetti and Simon 2018, Geruso *et al.* 2019).

Formally, let the $B(r_i, \psi_i) = B_0^{FFS} + \beta_1(r_i - 1) + \beta_2\psi_i$ be the expected cost of enrollee i with risk score r_i and SES ψ_i. Similarly, let $h(r_i, \psi_i) = h_0^{FFS} + \theta_1(r_i - 1) + \theta_2\psi_i$ be the expected health outcome of the enrollee. The intercepts, B_0^{FFS} and h_0^{FFS}, are the outcomes of the average FFS enrollee whose risk score is normalized to 1.[12] The risk adjustment model excludes SES ψ_i in the prediction, and hence estimates the following adjustment formula using only risk score r_i

$$\begin{aligned}\tilde{B}(r_i) &= B_0^{FFS} + (\beta_1 + \beta_2\tilde{\delta})(r_i - 1),\\ \tilde{h}(r_i) &= h_0^{FFS} + (\theta_1 + \theta_2\tilde{\delta})(r_i - 1),\end{aligned} \tag{A13}$$

[11] Mainous III and Talbert (1998) provide a historical account on the emergence and applications of the HEDIS health outcomes as indicators of quality.

[12] We normalize ψ_i so that the average FFS enrollee has $\psi_i = 0$.

where $\tilde{\delta}$ is the coefficient regressing ψ_i on r_i. When $\tilde{\delta} \neq 0$, the prediction coefficient $\beta_1 + \beta_2\tilde{\delta}$ differs from β_1 due to the correlation between the omitted factor ψ_i and risk score r_i. Unless ψ_i and r_i are perfectly correlated, predictions based on risk scores alone do not recover the true expected costs and health outcomes of enrollee i. Specifically, we define prediction errors as the difference between the model predictions and the true expected values as follows

$$\begin{aligned} D^B(\psi_i \,|\, r_i) &= \beta_2\tilde{\delta}(r_i - 1) - \beta_2\,\psi_i\,, \\ D^h(\psi_i \,|\, r_i) &= \theta_2\tilde{\delta}(r_i - 1) - \theta_2\,\psi_i\,. \end{aligned} \tag{A14}$$

Equation A14 states that, conditional on risk type r_i, enrollee i is cheaper than her predicted cost if $D^B(\psi_i \,|\, r_i) > 0$, or if $\psi_i < \tilde{\delta}(r_i - 1)$ for $\beta_2 > 0$. Similarly, conditional on risk type r_i, enrollee i is healthier than her predicted outcome if $D^h(\psi_i \,|\, r_i) > 0$, or if $\psi_i < \tilde{\delta}(r_i - 1)$ for $\theta_2 > 0$. Thus, equation A14 indicates the bias in the prediction models when ψ_i differs given risk type r_i.

The overall selection incentives implied by the prediction errors (equation A14) depend on the financial returns to selection determined by the payment model. Specifically, consider a marginal enrollee of risk type r_{N+1} joining an insurance contract with N enrollees and outcome rating q. The probability of the marginal enrollee testing below the risk-adjusted threshold is given by $\tilde{\Phi}\left(D^h(\psi_{N+1} \,|\, r_{N+1})\right)$.[13] The enrollment increases the outcome rating by $\Delta q = \frac{\tilde{\Phi}(\psi_{N+1} \,|\, r_{N+1}) - q}{N+1}$, and affects the plan profit according to

$$\begin{aligned} \Delta\pi &= \Delta q \cdot \sum_{i=1}^{N+1} \tilde{B}(r_i) + q \cdot \tilde{B}(r_{N+1}) + D^B(\psi_{N+1} \,|\, r_{N+1}) \\ &= \underbrace{\left[\tilde{\Phi}\left(D^h(\psi_{N+1} \,|\, r_{N+1})\right) - q\right] \cdot \tilde{B}(r)}_{\Delta \text{ benchmark bonus}} + q \cdot \tilde{B}(r_{N+1}) + \underbrace{D^B(\psi_{N+1} \,|\, r_{N+1})}_{\Delta \text{ benchmark}}, \end{aligned} \tag{A15}$$

[13] Compared to the risk-adjusted threshold $\tilde{h}$, health outcome h is below the threshold if $h(r_{N+1}, \psi_{N+1}) + \epsilon_{N+1} \leq \tilde{h}(r_{N+1})$, or if $\epsilon_{N+1} \leq D^h(\psi_{N+1} \,|\, r_{N+1})$.

where $r = \frac{1}{N+1}\sum_{i=1}^{N+1} r_i$ is the average risk score, and $\tilde{B}(r)$ is the average benchmark adjusted for risk scores.

Equation A15 indicates that the relative health outcome of the marginal enrollee, $\tilde{\Phi}\left(D^h(\psi_{N+1}\,|\,r_{N+1})\right) - q$, impacts the the benchmark bonus for *all* enrollees in the contract. This is because in the payment model, the quality rating is a multiplier that adjusts benchmarks for marginal *and* infra-marginal enrollees. In addition, risk adjustment over-predicts the cost of the marginal enrollee by $D^B(\psi_{N+1}\,|\,r_{N+1})$, and the insurer is paid the benchmark bonus $q \cdot \tilde{B}(r_{N+1})$ for the enrollment.

Based on equation A14 and A15, we examine whether the insurer has incentives to select particular types of ψ_{N+1} given risk score r_{N+1}. When risk adjustments under-predict both costs and health outcomes for higher values of ψ_{N+1} (i.e., when $\beta_2 > 0$ and $\theta_2 > 0$), the insurer would always prefer enrollees with lower values of ψ_{N+1} given risk score r_{N+1}. For instance, if enrollees from counties with higher poverty rates and dual eligibility status have worse outcomes and higher costs than enrollees with similar risk scores, then adjusting health outcomes solely based on risk would exacerbate the selection against such enrollees. On the contrary, low-spending enrollees healthier than their predicted outcomes would be increasingly favored by insurers as a result of risk adjustment.

Alternatively, when the prediction errors in equation A14 imply opposing selection incentives (i.e., when $\beta_2 \cdot \theta_2 < 0$), the overall impact on selection depends on the relative weight of benchmarks and bonus payments for plan profit. Specifically, enrollment increases the benchmark payment by $\frac{d\,\tilde{\Phi}(\psi_{N+1}\,|\,r_{N+1})}{d\,\psi_{N+1}} = -\beta_2$ as the marginal enrollee has higher ψ_{N+1}, and increases total bonus payments by $-\theta_2 \cdot \tilde{\phi}\,(\cdot) \cdot \tilde{B}(r)$ with higher ψ_{N+1}.[14] Thus, the overall impact of selecting higher types of ψ_{N+1} is given by

$$\frac{d\,\Delta\pi}{d\,\psi_{N+1}} = -\beta_2 - \theta_2 \cdot \tilde{\phi}(\cdot) \cdot \tilde{B}(r), \tag{A16}$$

[14] $\tilde{\phi}(\cdot)$ is the derivative of $\tilde{\Phi}\left(D^h(\psi_{N+1}\,|\,r_{N+1})\right)$ w.r.t. ψ_{N+1}.

which is a weighted average of β_2 and θ_2, the true impacts of ψ_{N+1} on costs and health outcomes. Given benchmark $\tilde{B}(r)$ and the distribution of measurement error $\phi(\cdot)$ of health outcomes, selecting individuals healthier than their predicted outcomes could be more profitable especially when the existing risk adjustment on benchmark is more sophisticated (i.e., β_2 closer to zero).

A.4 Welfare

To illustrate the welfare implications of selection for consumers, we allow consumers to differ in their valuation of insurance based on their risk type r and wealth w. Enrollees live in two counties denoted by $l = \{1, 2\}$. Wealth is distributed over $[\underline{w}, \overline{w}]$ according to $g_l(w)$ in county l. Risk type $r \in [0, 1]$ indicates the probability of having a health event. The conditional distribution of r given w follows $f_l(r|w)$. Absent the health event, enrollees incur cost v managing their chronic conditions. In the health event, the enrollee requires more intensive care and the out-of-pocket cost in this case is capped at m under the fee-for-service program. Thus, the expected utility of FFS insurance for consumer (w, r) is $\mathbb{E}_r u(w; v, m) = (1 - r) \cdot u(w - v) + r \cdot u(w - m)$.

MA insurance reduces the cost v of managing chronic conditions. For simplicity, we assume that enrollees pay zero out-of-pocket cost ($v = 0$) in the healthy state, and pay the FFS cost sharing m as FFS in the health event. Thus, MA insurance may be less desirable for patients in need of more intensive care. Given premium p_l, the expected utility of MA insurance is $\mathbb{E}_r u(w; p, m) = (1 - r) \cdot u(w - p) + r \cdot u(w - p - m)$ for consumer (w, r). The marginal enrollee indifferent between the MA and FFS insurance has risk type $\bar{r}(w; p_l) = \frac{u(w-p_l)-u(w-v)}{u(w-p_l)-u(w-v)-u(w-p_l-m)+u(w-m)}$, and healthier risk types below $\bar{r}(w; p_l)$ would

purchase MA insurance.[15] The resulting consumer surplus CS_l in county l is given by

$$\begin{aligned} CS_l = & \int_{\underline{w}}^{\overline{w}} \int_0^{\bar{r}(w;p_l)} \mathbb{E}_r u(w; p_l, m) \cdot f_l(r|w) g_l(w)\, dr dw \\ & + \int_{\underline{w}}^{\overline{w}} \int_{\bar{r}(w;p_l)}^1 \mathbb{E}_r u(w; v, m) \cdot f_l(r|w) g_l(w)\, dr dw, \end{aligned} \tag{A17}$$

which sums the surpluses accruing to MA (first term) and FFS enrollees (second term), respectively.

Assuming optimal insurance choice for marginal enrollees, an increase in premium p_l affects consumer welfare only through the surplus of infra-marginal enrollees. Specifically, $\frac{d\,CS_l}{d\,p_1} = -\mathbb{E}_w^l[\Lambda_l(w; p_l)]$, where $\Lambda_l(w; p_l) \equiv \int_0^1 1\{r \leq \bar{r}\} \cdot \mathbb{E}_r u'(w; p_l, m)\, f_l(r|w) dr$ indicates the change in consumer surplus at each wealth level for a small change in premium from p_l. Adding across counties, the total change in the consumer surplus is given by

$$\begin{aligned} \Delta CS &= -\mathbb{E}_w^1[\Lambda(w; p_l)] \cdot \Delta p_1 - \mathbb{E}_w^2[\Lambda(w; p_2)] \cdot \Delta p_2, \\ &= -\underbrace{\mathbb{E}_w^1[\Lambda(w; p_1)] \cdot (\Delta p_l + \Delta p_2)}_{\text{transfer value}} - \underbrace{\Delta p_2 \cdot \Delta \mathbb{E}_w[\Lambda(w; p)]}_{\text{insurance value}}. \end{aligned} \tag{A18}$$

The first term in equation A18 gives the transfer value of premiums. The second term indicates the loss in insurance values after a premium increase in county 2, where the valuation of MA insurance exceeds that in county 1 by $\Delta \mathbb{E}_w[\Lambda(w; p)] = \mathbb{E}_w^2[\Lambda(w; p_2)] - \mathbb{E}_w^1[\Lambda(w; p_1)]$. In the event that the insurer transferred premiums from risky to healthy counties without affecting the average premium, the term $\Delta p_1 + \Delta p_2$ is close to zero. This implies that if the marginal utility from insurance is constant across consumers – and hence $\Delta \mathbb{E}_w[\Lambda(w; p)] = 0$ in equation A18 – a pure transfer of insurance premiums would have no impact on consumer welfare. When marginal utilities differ, however, the transfer impacts welfare through the difference in insurance value, $\Delta \mathbb{E}_w[\Lambda(w; p)]$.

[15] The assumption that MA attracts healthier enrollees is consistent with the empirical evidence for advantageous selection in MA (e.g., Newhouse *et al.* 2012, Brown *et al.* 2014, Han and Lavetti 2017, Cabral *et al.* 2018) and in related markets (e.g., Fang *et al.*, 2008).

This model accounts for two sources of variations in insurance value. First, consumers differ in their wealth level w. Second, consumers differ in the risk type r given wealth. While full calibrations of consumer types and welfare are outside the scope of this paper, we illustrate the qualitative implications for welfare imposing simplifying assumptions. Specifically, we assume that at each wealth level, the distribution of risk types in the risky county 2 conditionally first-order stochastic-dominates the distribution in the healthy county 1. This implies that consumers are on average riskier in county 2, and facing similar premiums, enrollees in county 2 would value MA insurance more due to greater exposure to risks.[16] In this case, the premium transfers would result in greater losses of insurance values in county 2 than the surplus gains in county 1, on net reducing consumer welfare by $-\Delta p_2 \cdot \Delta \mathbb{E}_w[\Lambda(w; p)]$.

Moreover, differences in the wealth distribution between counties can complicate the insurance values based on risk types. If the wealth distribution in the healthy county 1 first-order stochastically dominates that in the risky county 2, then the disparity in $\Delta \mathbb{E}_w[\Lambda(w; p)]$ would widen due to greater insurance values at lower wealth. The overall impact on consumer welfare therefore depends on the joint distribution of risk types and wealth as well as the relative distributions between counties. In general, transfers of insurance premiums would disadvantage consumers in counties with lower wealth and worse health, or the more vulnerable populations of Medicare.

[16] Formally, assume that $p_1 = p_2 = p$. The insurance value at wealth w and premium p equals $\Lambda_l(w; p) \equiv \int_0^{\bar{r}(w;p)} \mathbb{E}u'(w; p, m) f_l(r|w) dr$, where the function inside the integral weakly increases in r. Because $\frac{F_2(r|w)}{F_2(\bar{r}|w)} \leq \frac{F_1(r|w)}{F_1(\bar{r}|w)}$ due to conditional first-order stochastic dominance, it follows that $\Lambda(w; p)$ is greater in county 2 for each w and premium p. Moreover, the average consumer and the average FFS enrollee are both riskier in county 2.

B Data Appendix

B.1 Estimation Sample

This section documents the construction of the estimation sample from administrative datasets provided by the Centers for Medicare and Medicaid Services (CMS). The basis of the analysis is the roster file of all Medicare Advantage plans, also known as the landscape file, which provides information on the plan's issuer, plan name and ID, and across the plan's service area, premium and prescription drug coverage (if any) at the county level. The roster file does not include plans in the Program of All-Inclusive Care for the Elderly (PACE plans), Special Needs Plans, Part B only plans, Medicaid plans, or employer-sponsored Medicare Advantage plans. Annual files from 2009 to 2014 can be downloaded at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/index.html?redirect=/PrescriptionDrugCovGenIn/.

We exclude from the samples Regional Preferred Provider Organization (PPO) Plans, which follow a different bidding process than the rest of Medicare Advantage plans. We also exclude plans that do not offer integrated prescription drug coverage. We obtain separate Part C (for Medicare Part A and B coverage) and Part D (prescription drug) premium from the Premium Source File, available in a separate folder for year 2009-2012 at the url above. The first three columns in Appendix Table B1 summarize the number of plan-county observations in the raw files, and the remaining sample after dropping regional PPOs and Part C only plans.

Plan risk scores, payments, and rebates are available at https://www.cms.gov/Medicare/Medicare-Advantage/Plan-Payment/Plan-Payment-Data.html?DLSort=0&DLEntries=10&DLPage=1&DLSortDir=ascending. We observe bids and rebates for plans bidding below the benchmark. We do not directly observe the plan-specific benchmark, but infer the benchmark from the rebate formula. Also available is the Part C risk score used to adjust Medicare Advantage benchmarks and payments. The risk score is calculated

from a hierarchical model that accounts for the severity of conditions and the interaction of conditions from multiple diagnoses. Plans with missing payment information and risk scores are dropped from the sample.

Moreover, in the Quality Bonus demonstration, star rating in year t-1 is used to adjust bonus payments in year t. Payments to plans without a quality rating in the previous year are subject to a different set of rules. For continuing contracts with missing rating data due to small enrollments, a fixed star rating is applied to all such contracts to determine benchmark and rebate bonuses.[17] Since the incentive structure is generally different from that of rated contracts in the same year, we drop contract-year observations where the payment-relevant quality rating is missing. This affects a tiny fraction of the estimation sample, since the vast majority of contracts rated 3.0 stars and above at least once in the baseline continue to receive quality ratings over the sample period.[18] Data on measure ratings and overall ratings are available at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html. The crosswalk file linking plans and contracts over time is available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MCRAdvPartDEnrolData/Plan-Crosswalks.html.

We merge in enrollment counts at the plan-year-county level from monthly enrollment counts from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MCRAdvPartDEnrolData/Monthly-Enrollment-by-Contract-Plan-State-County.html. Annual enrollment sums over enrollee-months over a 12-month period. However, exact counts are masked for counties with fewer than 10 enrollees. We include the full range of service areas

[17] In 2012, a uniform 3.0 star rating is applied to benchmark bonuses in such cases. The rebate bonus is uniformly set at the level of 4.5 stars. New contracts do not receive a star rating in the first three years. Instead, a weighted average of existing contracts offered by the organization is used to impute a star rating for payment purposes.

[18] Less than 1% of the rated contracts in year t have missing star ratings in $t+1$ in the estimation sample. Less than 4% of the baseline contracts have a missing star rating in 2011-2014. Dropping these contracts from the estimation sample gives very similar results.

Table B1: Construction of the estimation sample

	2009	2010	2011	2012	2013	2014
Landscape File observations	99,147	66,674	36,689	40,637	39,548	31,784
Contract observations	539	495	413	463	461	473
Dropping Regional PPOs	-6,181	-7,883	-7,497	-6,877	-6,171	-6,317
Dropping Part C only plans	-42.867	-22.489	-9,674	-10,550	-9,423	-6,343
Plan-county observations	50,099	36,302	19,518	23,210	23,954	19,156
Contract observations	514	470	391	443	442	455
Missing payment/risk score	-2,449	-2,129	-2,899	-3,819	-3,709	-3,090
Missing quality rating star	-21,987	-15,078	-6,712	-5,314	-3,915	-1,426
Plan-county observations	25,663	19,095	9,907	14,077	16,330	14,640
Plan observations	1,183	1,092	829	1,090	1,246	1,349
Contract observations	244	234	248	313	333	336
Linked contract observations			406			
Continuing from baseline			244			
excluded: less than 3.0 stars in 2009 and 2010			54			
low quality rating: less than 4.0 stars, at least one rating ≥3.0 stars			135			
high quality rating: at least one rating ≥4.0 stars			55			
high selection (<50% service area risk)			27			

Notes: The table shows the step-by-step construction of the estimation sample from yearly Landscape Files. Contracts continuing from baseline are those first appearing in the data in 2009 or 2010. Contracts rated below 3.0 stars in both years of 2009-2010 are excluded from the analysis. Low-rated contracts are rated less than 4.0 stars in both 2009 and 2010, but have at least one rating between 3.0 stars and 3.5 stars in 2009-2010. High-rated contracts have at least one 4.0-star rating or above in 2009 or 2010. High-selection contracts are high-rated contracts in service areas where the average fee-for-service risk score is below 0.975, the median of high-rated contracts.

when constructing the within-contract differences in county characteristics, but exclude county-plans with missing enrollments when aggregating prices to the county-contract level. These missing enrollments affect about one-fourth of the county-contract prices. Results are similar without dropping low-enrollment county-plans.

In the difference-in-differences analysis, we summarize the location variation using service area variables at the contract level, and drop the duplicate observations by location. We end up with a little over 1,000 plans each year, for a total of 6,789 plan-year observations from 2009-2014. These plans are offered by 406 distinct contracts, of which 244 continued from the baseline in 2009-2010. For these baseline contracts, 65 received at least one 4.0-star rating or above in 2009-2010. 149 are rated less than 4.0 stars in both years but have at least one rating at or above 3.0 stars. The remaining contracts are rated below 3.0 stars in both 2009 and 2010. These contracts are subject to cancellation after three consecutive ratings below 3.0 stars. We do not include the last set of lowest-rated contracts in the analysis.

In the triple-difference analysis, we consider a range of county characteristics to un-

derstand the within-contract differences in prices. We summarize the county variables below.

B.2 County Characteristics

County fee-for-service (FFS) risk scores and costs are from the Medicare Geographic Variation Public Use File at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_PUF.html. We use the 2009-2010 average for the baseline. The risk scores are calculated from the same Hierarchical Condition Category (HCC) model that generates Medicare Advantage risk scores. Payments to providers in the FFS Medicare are adjusted for the case-mix of patient conditions coded in the risk score. We use the differences in FFS risks scores as measures of potential gains from selection for Medicare Advantage insurers across the service area.

Three variables measure the cost of medical practices in the FFS program. The first, unadjusted cost is calculated as the total Part A and Part B claim costs of medical practices divided by the number of beneficiaries attributed to the practices. The second measure adjusts the raw average cost by local price factors outside the physician's control. Specifically, a national payment scheme is applied to override state-specific fee schedules, and input prices such as labor and facility costs are standardized at the national level.[19] The price-standardized cost is further adjusted for patient case-mixes in the third, risk-adjusted cost measure, which captures local costs of medical practices holding fixed both prices and risk. The adjustments reveal the relevant component in costs which relates to the differences in prices. The first four rows of Appendix Table B2 summarizes the FFS risk scores and costs by county.

Diabetes prevalence rates by county are available from the Center of Disease Control

[19]More details of the price adjustments are available at http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350.

Table B2: Summary of county characteristics

	(I)	(II)	(III)		(IV)	(V)	(VI)
Health Risks and Costs	mean	s.e.	*N*	Socio-Economic Factors	mean	s.e.	*N*
FFS risk score	0.95	0.002	1,852	Per capita income (k)	35.52	0.20	1,828
Per capita FFS Cost (k)	8.84	0.032	1,852	Per capita transfer income (k)	8.25	0.038	1,828
– price adjusted (k)	8.82	0.031	1,852	Non-White (%)	11.38	0.31	1,852
– price-risk adjusted (k)	9.54	0.023	1,852	Some college (%)	37.23	0.24	1,852
Diabetes (%)	8.85	0.049	1,852	HHI	0.57	0.005	1,852
Hypertension (%)	37.62	0.12	1,852	Low-rated HHI	0.76	0.007	1,401
Hospital re-admission (%)	17.81	0.062	1,840	High-rated HHI	0.89	0.008	584
Preventable hospitalization (%)	7.13	0.059	1,826				

Notes: The table summarizes the baseline characteristics of counties in the estimation sample. Counties with missing data of the characteristics are not included. Quality rating-specific HHIs are only calculated for counties where enrollment in the measured quality rating is positive in the baseline.

(CDC) at https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html#. The estimates are based on reported diagnoses from adults over age 20 in the Behavioral Risk Factor Surveillance System (BRFSS). We multiply the age-adjusted estimate, which gives the prevalence rate in a standard-age population, by the FFS risk score to account for differences in health conditions: prevalence is adjusted upward in locations where individuals have more diagnoses in the risk score. We apply the diagnosis intensity factors developed in Finkelstein *et al.* (2017) to the FFS risk scores. The resulting prevalence rate accounts for age, risk, and coding differences across counties.

County hypertension prevalence rates are published by the Institute for Health Metrics and Evaluation (IHME) for adults over age 30 in 2001-2009 (http://ghdx.healthdata.org/record/ihme-data/united-states-hypertension-estimates-county-2001-2009). We use the 2009 value for the baseline. The prevalence rate is calculated as the percent of respondents having systolic blood pressure above 140 mm Hg or taking anti-hypertensive medication in the National Health and Nutrition Examination Survey (NHANES) and the BRFSS. The estimates correct for self-report and coding biases, standardized using national age-race distributions. Details of the construction are provided in Olives *et al.* (2013).

Data on hospital re-admission rate and preventable hospital stays are taken from the Area Health Resources File (AHRF, available at https://data.hrsa.gov/topics/health-

workforce/ahrf). We use the 2010 variables for the baseline. The re-admission rate calculates the percent of re-admitted patients within 30 days of discharge from an acute hospital. The measure is associated with the access to and the quality of inpatient care. Preventable hospital stay calculates the percent of hospital discharge of outpatient treatable conditions in the FFS population. Higher rate indicates lower quality of outpatient care.

County demographic data come from the Survey of Epidemiology and End Results (SEER, available at https://www.nber.org/data/seer_u.s._county_population_data.html), which provides population estimates by age groups and race. We focus on the elderly (65+) population and the White vs. non-White categories. Percent with college education is calculated from the American Community Survey (ACS) micro data (Ruggles *et al.*, 2019). Per capita income and transfer income are from the Bureau of Economic Analysis (https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas), where transfer income includes social security, unemployment insurance, disability, medical and income assistance payments from governments, nonprofit organizations, and businesses. Finally, we calculate the Herfindahl-Hirschman Index (HHI) from contract market shares. The denominator of the market share is the sum of member-month enrollments in all rated contracts in a county. We calculate the quality rating-specific HHI for markets at the level of county-rating pairs.

C Recent Policy Changes in Medicare Advantage

In this section, we summarize the main policy changes to the star rating and the payment model since 2014, the year that marks the end of our study-period.

Changes pre-2020. The star rating was not subject to major changes for the period between 2015 and 2020. The weights of outcome (3.0), process (1.0), experience and access (1.5) measures remained constant in this period. The main change was the introduction of two "improvement measures," one for health plan quality (Part C) and another one for drug plan quality (Part D). Each measure indicates how much the health or drug plan's performance has improved or declined from one year to the next.[20] These two new measures receive 5.0 weights, the largest in the star rating. Within these broad groups of measures, CMS adds new measures or excludes old ones from the star rating computation on a yearly basis. Since 2017, new measures included in the ratings receive a weight equal to 1.0 for their first year of inclusion. In subsequent years the weight associated with the measure weighting category is used. Due to space constraints, we do not present here the entire list of measure changes. However, a summary table of measure changes is available in the Attachment J of each yearly technical note. The attachment reports which measures are included in the years from 2009 onward.[21]

As for risk adjustment, CMS introduced the Categorical Adjustment Index (CAI) in 2017, as an interim policy until regulators decide which clinical measures should be risk adjusted and how to do so (Sorbero *et al.*, 2018). In practice, CMS groups contracts based on their percentage of LIS and disabled enrollees, and applies the relevant updating factor

[20] According to the technical note, the improvement measure is a ratio: "the numerator is the net improvement, which is a sum of the number of significantly improved measures minus the number of significantly declined measures, while the denominator is the number of measures eligible for the improvement measure," which may vary year by year. To be included in the computation, a measure must be consistently specified across adjacent years. See the 2015 Technical note for additional details https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/2015StarRatingsTechnicalNotes.pdf.

[21] For instance, the 2019 technical note is available at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/Star-Ratings-Technical-Notes-Oct-10-2019.pdf.

to the overall unadjusted, unrounded star score for each contract within a group.[22] To determine disability and socioeconomic status, CMS uses already available data from (i) monthly enrollment files, (ii) Social Security Administration and Railroad Retirement Board Record System, and (iii) Centers for Medicare & Medicaid Services Integrated Data Repository (IDR). Thus, no new data are created to produce this coarse risk-adjustment. Studying the impact of CAI on the star score, Sorbero *et al.* (2018) show that CAI benefited mostly plans with at least 50% enrollees eligible both for Medicare and Medicaid, or receive Part D low-income subsidies.

2021 and Beyond. The main change in 2021 is that access and complaints measures see their weights increased from 1.5 to 2.0. The other measures maintain the same weights as in previous years.[23] In addition, CMS replaced the 2021 star rating measures calculated based on HEDIS and CAHPS data collections with earlier values from the 2020 star ratings to avoid contamination due to COVID-19.

In 2020, CMS made proposals for substantial changes to the star ratings.[24] These changes will impact the data collection in 2021 and affect measurement of the star ratings for 2023. We report the key changes below.[25]

First, CMS proposed to change the methodology to compute the cut points for non-CAHPS measures to exclude outlier plans.[26] In particular, CMS now relies on the Tukey's fences methodology, which defines outliers as scores that are lower (higher) than the first (third) quartile by more than x times the interquartile range (i.e., the difference

[22]Source: https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/Supplement-for-Categorical-Adjustment-Index-.pdf.

[23]Source: https://www.cms.gov/files/document/2021technotes20201001.pdf-0. This increase is due to the fact that CMS is satisfied about the reliability of these CAHPS measures. Source: https://www.federalregister.gov/d/2020-11342/p-431.

[24]Source: https://www.federalregister.gov/documents/2020/06/02/2020-11342/medicare-program-contract-year-2021-policy-and-technical-changes-to-the-medicare-advantage-program.

[25]Source:https://www.federalregister.gov/documents/2018/11/01/2018-23599/medicare-and-medicaid-programs-policy-and-technical-changes-to-the-medicare-advantage-medicare.

[26]These changes are part of a long process that started in 2018. Source: https://www.federalregister.gov/documents/2018/11/01/2018-23599/medicare-and-medicaid-programs-policy-and-technical-changes-to-the-medicare-advantage-medicare.

between the third and first quartiles). CMS set the multiplier x to three. CMS expects that this methodology will increase the number of plans with 1.0 and 2.0 star rating and result in savings for $935 million in 2025 and $1,449.2 billion in 2030.[27] From the 2022 measurement year (2024 star rating), the outliers will be removed prior to calculating the cut points to further increase the predictability and stability of the Star Ratings system.[28]

Second, the new rating weights patient experience/complaints and access measures more.[29] The new ruling will bring the weights for these measures from 2.0 to 4.0.[30] The measures affected come from the CAHPS survey, and include the Members Choosing to Leave the Plan, Appeals, Call Center, and Complaints measures.

Third, CMS proposed the removal of the Rheumatoid Arthritis Management measure from the Star Ratings program for performance periods beginning on or after January 1, 2021 due to potential issues at the way the measure is constructed.[31]

Fourth, moving to consolidation, CMS proposed changes on the resulting star rating of merged contracts. In particular, consolidating contracts belonging to the same organization but with different star ratings will obtain a consolidated star rating equal to the weighted average of the star ratings of the consolidating contracts using enrollment as weights.[32] In addition, CMS provide details on star rating computations when consolidating plans have missing values in some measures due to integrity concerns. CMS also changed the definition of a "new MA plan" to "a MA contract offered by a parent organization that has not had another MA contract in the previous 4 years."[33] To this end, CMS proposed rules to compute the star rating for new MA plans, and MA plans with low enrollment.

[27] Source: https://www.mintz.com/insights-center/viewpoints/2146/2020-02-cms-proposes-significant-changes-medicare-advantage-part-d.

[28] Source: https://www.federalregister.gov/documents/2020/06/02/2020-11342/medicare-program-contract-year-2021-policy-and-technical-changes-to-the-medicare-advantage-program.

[29] This change is consistent with our analysis as we show in Appendix Table D29 that finds these measures are not affected by risk selection.

[30] Source: https://www.federalregister.gov/d/2021-00538/p-645.

[31] Source: https://www.federalregister.gov/d/2020-11342/p-422.

[32] Source: https://www.cms.gov/files/document/2021-advance-notice-part-ii.pdf, page 14.

[33] Source: https://www.federalregister.gov/d/2021-00538/p-654.

Finally, in each yearly Policy and Technical changes CMS states that it continues to solicit comments and suggestions regarding risk adjusting star rating measures.[34] However, no specific plan to risk adjust HEDIS measure has been taken to date. Still, one might consider the addition of the new HEDIS measure "follow-up care provided after an emergency department" a step towards the risk adjustment direction, as high values in this measure show that a plan is doing well at preventing the development of more severe complications for populations with complicated case-mixes.[35] However, as noted in some comments to the document, CMS was asked to risk adjust this measure as plans enrolling individuals with a low socio-economic background may be at disadvantage and the CAI adjustment is deemed inadequate.[36]

[34]See, for instance, https://www.federalregister.gov/d/2021-00538/p-1092 for 2021.

[35]The measure is computed as the percentage of ED visits for members 18 years and older who have high-risk multiple chronic conditions who had a follow-up service within 7 days of the ED visit between January 1 and December 24 of the measurement year.

[36]Source: https://www.federalregister.gov/d/2021-00538/p-748.

D Additional Tables

Table D1: Bonus adjustments on benchmarks and rebates

	Star Rating					
Year	≤2.5	3.0	3.5	4.0	4.5	5.0
	Benchmark Bonus θ^{star} = 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%
2015 (ACA)	0.0%	0.0%	0.0%	5.0%	5.0%	5.0%
	Rebate Percentage γ^{star}					
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%
2015 (ACA)	50.0%	50.0%	65.0%	65.0%	70.0%	70.0%

Notes: The table shows the changes in benchmark bonuses and rebate percentages for year 2009-2014 and for the first year of ACA in 2015. The Quality Bonus Payment Demonstration (QBP) became effective in 2012 and ended in 2014, after which the ACA payment model took effect in 2015. For benchmark bonuses, contracts above 4.0 stars continue to receive full benchmark bonuses (5%) from 2015 onward. Contracts below 4.0 stars are no longer eligible for bonus payments. The rebate percentages under ACA are the same as those in 2014. Quality adjustments on benchmarks and rebates are calculated from lagged star ratings in year $t-1$. Source: Centers for Medicare and Medicaid Services, Advance Notice of Methodological Changes, Calendar Year 2009-2015.

Table D2: List of measures, weights, and risk adjustments, 2013 star rating

Measure Name	Weight	Data Source	Time Period Measured	Risk Adjustment
A. Outcome Measures (3.0 weights)				
Diabetes Care – Blood Sugar Controlled	3	HEDIS	01/01/2011 - 12/31/2011	N
Diabetes Care – Cholesterol Controlled	3	HEDIS	01/01/2011 - 12/31/2011	N
Controlling Blood Pressure	3	HEDIS	01/01/2011 - 12/31/2011	N
High Risk Medication	3	PDE	01/01/2011 - 12/31/2011	N
Diabetes Treatment	3	PDE	01/01/2011 - 12/31/2011	N
Part D Medication Adherence for Oral Diabetes Medications	3	PDE	01/01/2011 - 12/31/2011	N
Part D Medication Adherence for Hypertension (RAS antagonists)	3	PDE	01/01/2011 - 12/31/2011	N
Part D Medication Adherence for Cholesterol (Statins)	3	PDE	01/01/2011 - 12/31/2011	N
Plan All-Cause Readmissions	3	HEDIS	01/01/2011 - 12/31/2011	Y
Improving or Maintaining Physical Health	3	HOS	04/18/2011 - 07/31/2011	Y
Improving or Maintaining Mental Health	3	HOS	04/18/2011 - 07/31/2011	Y
B. Access Measures (1.5 weights)				
Getting Needed Care	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Getting Appointments and Care Quickly	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Customer Service	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Overall Rating of Health Care Quality	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Overall Rating of Plan	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Complaints about the Health Plan	1.5	CTM	01/01/2012 - 06/30/2012	N
Beneficiary Access and Performance Problems	1.5	CMS	01/01/2011 - 12/31/2011	N
Members Choosing to Leave the Plan	1.5	MBDSS	01/01/2011 - 12/31/2011	N
Plan Makes Timely Decisions about Appeals	1.5	IRE	01/01/2011 - 12/31/2011	N
Reviewing Appeals Decisions	1.5	IRE	01/01/2011 - 12/31/2011	N
Call Center – Foreign Language Interpreter and TTY/TDD Availability	1.5	Call Center	01/30/2012 - 05/18/2012	N
Call Center – Pharmacy Hold Time	1.5	Call Center	02/06/2012 - 05/18/2012	N
Call Center – Foreign Language Interpreter and TTY/TDD Availability	1.5	Call Center	01/30/2012 - 05/18/2012	N
Appeals Auto–Forward	1.5	IRE	01/01/2011 - 12/31/2011	N
Appeals Upheld	1.5	IRE	01/01/2012 - 6/30/2012	N
Getting Information From Drug Plan	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Rating of Drug Plan	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
Getting Needed Prescription Drugs	1.5	CAHPS	02/15/2012 - 05/31/2012	Y
C. Process Measures (1.0 weights)				
Breast Cancer Screening	1	HEDIS	01/01/2011 - 12/31/2011	N
Colorectal Cancer Screening	1	HEDIS	01/01/2011 - 12/31/2011	N
Cardiovascular Care – Cholesterol Screening	1	HEDIS	01/01/2011 - 12/31/2011	N
Diabetes Care – Cholesterol Screening	1	HEDIS	01/01/2011 - 12/31/2011	N
Glaucoma Testing	1	HEDIS	01/01/2011 - 12/31/2011	N
Annual Flu Vaccine	1	CAHPS	02/15/2012 - 05/31/2012	N
Monitoring Physical Activity	1	HOS/HEDIS	04/18/2011 - 07/31/2011	Y
Adult BMI Assessment	1	HEDIS	01/01/2011 - 12/31/2011	N
Care for Older Adults – Medication Review	1	HEDIS	01/01/2011 - 12/31/2011	N
Care for Older Adults – Functional Status Assessment	1	HEDIS	01/01/2011 - 12/31/2011	N
Care for Older Adults – Pain Screening	1	HEDIS	01/01/2011 - 12/31/2011	N
Osteoporosis Management in Women who had a Fracture	1	HEDIS	01/01/2011 - 12/31/2011	N
Diabetes Care – Eye Exam	1	HEDIS	01/01/2011 - 12/31/2011	N
Diabetes Care – Kidney Disease Monitoring	1	HEDIS	01/01/2011 - 12/31/2011	N
Rheumatoid Arthritis Management	1	HEDIS	01/01/2011 - 12/31/2011	N
Improving Bladder Control	1	HOS/HEDIS	04/18/2011 - 07/31/2011	Y
Reducing the Risk of Falling	1	HOS/HEDIS	04/18/2011 - 07/31/2011	Y
Care Coordination	1	CAHPS	02/15/2012 - 05/31/2012	Y
Health Plan Quality Improvement	1	CMS	2012 rating	N
Enrollment Timeliness	1	MARx	01/01/2012 - 06/30/2012	N
Drug Plan Quality Improvement	1	CMS	2012 rating	N
MPF Price Accuracy	1	PDE	01/01/2011 - 09/30/2011	N

Notes: The table lists the 51 quality measures included in the star rating of MAPD contracts in 2013, the weight of each measure, data source, measurement period, and the application of some case-mix (demographic or risk) adjustment to the measure. Of the 51 measures, only 25 measure-level ratings are required to compute an overall star rating as a weighted average of measure ratings. Outcome measures (3.0 weights) account for 50% of the overall rating for high-selection contracts.

Table D3: Year t selection and bonus rates in year $t+3$

	Star Rating					
Year t	≤2.5	3.0	3.5	4.0	4.5	5.0
	$t+3$ Benchmark Bonus θ^{star} = 1 + %					
2012	0.0%	0.0%	0.0%	5.0%	5.0%	5.0%
2013	0.0%	0.0%	0.0%	5.0%	5.0%	5.0%
2014	0.0%	0.0%	0.0%	5.0%	5.0%	5.0%
	$t+3$ Rebate Percentage γ^{star}					
2012	50.0%	50.0%	65.0%	65.0%	70.0%	70.0%
2013	50.0%	50.0%	65.0%	65.0%	70.0%	70.0%
2014	50.0%	50.0%	65.0%	65.0%	70.0%	70.0%

Notes: The table illustrates the three-year lag between enrollment in year t and the payout of bonus payments in year $t+3$. Because bonus payments differ by the star rating in $t+2$, and because health outcome measures are based on enrollee data collected from two years prior in year t, selecting healthier enrollees in year t would affect bonus payments three years later in year $t+3$. During the QBP period in 2012-2014, the return of risk selection depends on the payment model effective in 2015-2017, or the ACA model shown in the table.

Table D4: Distributional effects of the payment reform on risk scores, by deciles

	(I) Difference-in-Differences	(II) Changes-in-Changes	(III) Baseline High
10%	-0.039	-0.049***	0.842
	(0.024)	(0.018)	
20%	-0.085***	-0.073***	0.915
	(0.025)	(0.017)	
30%	-0.057***	-0.056***	0.950
	(0.021)	(0.020)	
40%	-0.036**	-0.046***	0.966
	(0.015)	(0.014)	
50%	-0.032**	-0.039***	0.980
	(0.015)	(0.011)	
60%	-0.014	-0.019	1.002
	(0.016)	(0.018)	
70%	-0.023	-0.018	1.026
	(0.016)	(0.017)	
80%	-0.016	-0.023	1.057
	(0.016)	(0.021)	
90%	-0.038	-0.029	1.096
	(0.031)	(0.020)	

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

Notes: The table shows the effects of the payment reform across deciles of risk scores in high-rated contracts. Column 1 shows estimates from the grouped quantile approach of Chetverikov *et al.* (2016). Column 2 shows the changes-in-changes estimates following Athey and Imbens (2006). In both cases, the standard errors in parenthesis are based on the empirical distribution of estimates from 500 replication samples block-bootstrapped by contracts. To help understand effect sizes, column 3 shows the deciles of risk scores in high-rated contracts in the 2009-2010 baseline.

Table D5: Effect on risk scores, by service area risks

	(I)	(II)	(III)	(IV)	(V)
Treat · Post	-0.026***	-0.037***	-0.016	-0.043***	-0.007
	(0.008)	(0.010)	(0.010)	(0.012)	(0.016)
Treated contracts	high-rated	high-rated		high-rated	
Service area risk		<median	>median	<25%	>75%
y mean	0.97	0.97	0.97	0.97	0.97
R^2	0.86	0.86	0.85	0.86	0.85
N	1,122	920	941	851	858

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

Notes: The table shows the difference-in-differences estimates on the risk score of high-rated contracts. Column 1 estimates the changes in high-rated contracts relative to low-rated ones. Column 2 and 3 divide high-rated contracts by the median service area risk (0.975), and estimate separate effects below (column 2) and above (column 3) the median. Column 4 estimates the effect on high-rated contracts in the lower 25% of service area risk (<0.902), and column 5 estimates the effect in the upper 25% (>1.009). All specifications control for contract fixed effects. Clustered standard errors at the contract level in parenthesis.

Table D6: Effect of the payment reform on service area characteristics

	(I) # Counties	(II) Service Area Risk	(III) Benchmark	(IV) Double-Bonus County	(V) # Plans
High · Post	8.70	0.0028	1.80	-0.020	-0.17
	(8.39)	(0.0024)	(2.94)	(0.021)	(0.23)
y mean	25.09	0.98	795.12	0.72	3.40
R^2	0.73	0.98	0.92	0.90	0.87
N	1,122	1,122	1,122	1,122	1,122

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

Notes: The table shows difference-in-differences estimates on the composition of service areas along measured characteristics. We use 2012 values of county benchmarks and FFS risk scores to construct service area characteristics in column 2-4 at the contract-year level. The constructed variables reflect changes in the composition of service areas by county characteristics, rather than changes in county characteristics over time. Numbers of counties (column 1) and plans (column 5) are counted within contract-years. Estimated effects indicate selection over the composition of service areas along measured characteristics rather than the temporal differences in these characteristics. All regressions include contract fixed effects. Clustered standard errors at the contract level in parenthesis.

Table D7: Effect of the payment reform on Part C premiums, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			23.61**			23.04
			(11.73)			(15.56)
Risk · Post	-10.12	12.00	-11.42	-8.19	10.19	-10.34
	(7.27)	(11.21)	(7.01)	(7.38)	(14.83)	(6.96)
High · Post			-7.86**			-9.26**
			(3.97)			(4.29)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	26.05	51.53	33.03	24.84	49.48	31.24
R^2	0.77	0.85	0.82	0.77	0.84	0.81
N	14,861	5,611	20,472	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in Part C premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium differences in low- and high-rated contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-rated contracts. Column 4-6 repeat the analysis but restrict the within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D8: Effect on zero premiums or drug deductibles, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Zero Part C Premium			Zero Part D Premium			Zero Drug Deductible		
Risk · High · Post			-0.21			-0.50**			0.15
			(0.15)			(0.21)			(0.23)
Risk · Post	0.14	-0.12	0.11	0.24	-0.38**	0.21	-0.17**	-0.093	-0.17*
	(0.12)	(0.11)	(0.11)	(0.15)	(0.18)	(0.14)	(0.087)	(0.22)	(0.091)
High · Post			0.026			0.013			0.062
			(0.033)			(0.033)			(0.054)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	0.46	0.24	0.40	0.45	0.19	0.38	0.85	0.87	0.85
R^2	0.77	0.71	0.77	0.75	0.78	0.77	0.66	0.65	0.66
N	4,393	1,633	6,026	4,393	1,633	6,026	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in the share of zero-premium and zero-drug deductible plans over county risk scores. Specifically, the outcome variable is the percent of plans with zero premiums (or drug deductibles) offered in the contract-county pair, weighted by enrollment. We restrict locations to counties in the lower and upper 15% of county risk scores in the contract's service area. Column 1-2 focus on the percent of plans with zero Part C premiums, showing separate difference-in-differences estimates for low- and high-rated contracts. Column 3 shows the triple-difference estimate giving the differential effect on high-rated contracts. Column 4-6 (7-9) repeat the analysis focusing on the percent of plans with zero Part D premiums (drug deductibles). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D9: Effect of the payment reform on the total premium (Part C and D), within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			38.88***			40.47**
			(12.64)			(16.17)
Risk · Post	-14.41	29.66**	-14.39	-12.20	26.83*	-13.70
	(9.22)	(11.35)	(9.10)	(9.60)	(15.44)	(9.17)
High · Post			-6.64			-6.87
			(4.88)			(5.17)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	44.33	80.69	54.30	43.89	77.47	52.99
R^2	0.85	0.85	0.87	0.85	0.86	0.86
N	14,861	5,611	20,472	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in total premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium differences in low- and high-rated contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-rated contracts. Column 4-6 repeat the analysis but restrict the within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D10: Effect of the payment reform on drug deductibles, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			0.96			-15.73
			(46.13)			(53.70)
Risk · Post	34.54*	60.66	40.03*	30.34*	37.33	34.64**
	(19.24)	(46.27)	(20.66)	(15.52)	(53.12)	(16.62)
High · Post			-13.19			-15.55
			(10.33)			(9.79)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	30.99	25.33	29.44	29.27	25.49	28.25
R^2	0.71	0.59	0.68	0.70	0.65	0.69
N	14,861	5,611	20,472	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in drug deductibles over county risk scores. Column 1-2 show the difference-in-differences estimates for low- and high-rated contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-rated contracts. Column 4-6 repeat the analysis but restrict the within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D11: Effect of the payment reform on Part D premiums over income, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	per capita income (thousands)			per capita transfer income (thousands)		
Treat · High · Post			-0.15			0.10
			(0.11)			(0.75)
Treat · Post	0.060	-0.077	0.063	-0.098	-0.021	-0.11
	(0.061)	(0.088)	(0.061)	(0.33)	(0.70)	(0.33)
High · Post			2.69			2.39
			(2.03)			(2.03)
Counties		15% tails			15% tails	
Contracts	low	high	all	low	high	all
y mean	18.00	27.99	20.72	18.00	27.99	20.72
R^2	0.75	0.70	0.75	0.75	0.70	0.75
N	4,357	1,633	5,990	4,357	1,633	5,990

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in per capita income (column 1-3) and per capita transfer income (column 4-6). County risk score is negatively associated with income, and positively associated with transfer income. We show separate difference-in-differences estimates on low- and high-rated contracts, followed by the triple-difference estimate on high-rated plans. We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis (Table 2). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D12: Effect of the payment reform on Part D premiums over socio-economic status, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	non-White (%)			some college (%)		
Treat · High · Post			0.041			-0.030
			(0.059)			(0.13)
Treat · Post	-0.049	0.026	-0.047	-0.034	-0.053	-0.032
	(0.039)	(0.049)	(0.038)	(0.071)	(0.11)	(0.071)
High · Post			2.34			2.67
			(2.06)			(2.09)
Counties		15% tails			15% tails	
Contracts	low	high	all	low	high	all
y mean	18.05	27.99	20.74	18.05	27.99	20.74
R^2	0.75	0.70	0.75	0.75	0.70	0.75
N	4,393	1,633	6,026	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in socio-economic status (SES), proxied by percent White in column 1-3 and percent having some college education in column 4-6. County risk score is negatively associated with college education and positively associated with percent non-White. We show separate difference-in-differences estimates on low- and high-rated contracts, followed by the triple-difference estimate on high quality rating. We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis (Table 2). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D13: Effect of the payment reform on Part D premiums due to the Special Enrollment Period, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Risk · High · Post					17.43**	17.88**	17.62**
					(8.51)	(8.80)	(8.80)
Risk · Post	-4.01	16.64**	16.73**	16.71**	-3.36	-3.52	-3.27
	(5.57)	(7.35)	(7.57)	(7.59)	(5.35)	(5.34)	(5.31)
High · Post					2.38	2.50	2.47
					(2.00)	(2.03)	(2.03)
Counties				15% tails			
5-star counties	Y	Y	Y	N	Y	Y	N
Contracts	low	high	high	high	(2)-(1)	(3)-(1)	(4)-(1)
5-star contracts		Y	N	N	Y	N	N
y mean	18.05	27.99	28.13	28.04	20.74	20.75	20.82
R^2	0.75	0.70	0.70	0.70	0.75	0.75	0.75
N	4,393	1,633	1,601	1,594	6,026	5,991	5,902

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

Notes: Table shows the within-contract differences in Part D premiums over county differences in county risk scores. Column 1-2 repeats the estimates for low- and high-rated contracts shown in Table 2. Column 3 estimates effects on high-rated contracts excluding contracts with 5.0-star ratings. Due to the Special Enrollment Period which took effect in 2012, 5.0-star contracts are open to new enrollees year round and are hence subject to additional selection risks. Column 4 further excludes all counties covered by 5.0-star contracts. Column 5-7 shows triple-difference effects on high-rated contracts as specified in column 2-4. We restrict counties to those in the lower or upper 15% of county risk scores in the contract's service area. All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D14: Effect of the payment reform on Part D premiums over market concentration, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
County variation in *Treat*:		HHI			FFS risk score			rating-specific HHI	
Treat · High · Post			8.94			21.67***			-2.08
			(8.17)			(7.91)			(5.85)
Treat · Post	7.05	15.24**	6.96	-2.05	22.24***	-1.54	1.06	-1.31	0.93
	(4.62)	(6.84)	(4.60)	(5.85)	(6.40)	(5.63)	(3.12)	(5.07)	(3.13)
High · Post			2.72			2.57			2.68
			(1.94)			(1.89)			(2.17)
HHI · High · Post						11.52			
						(7.89)			
HHI · Post				6.84	17.66***	6.75			
				(4.73)	(6.23)	(4.73)			
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	18.05	27.99	20.74	18.05	27.99	20.74	18.05	27.99	20.74
R^2	0.75	0.71	0.75	0.75	0.71	0.75	0.75	0.70	0.75
N	4,393	1,633	6,026	4,393	1,633	6,026	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in market concentration (column 1-3 and column 7-9) and in risk scores (column 4-6). We measure market concentration by the Herfindahl-Hirschman Index (HHI). HHI is calculated at the level of county l as $HHI_l = \sum_c (s_{cl})^2$, where s_{cl} is the market share of contract c in the county in column 1-3. We calculate HHI for quality-county pairs in column 7-9. Column 4-6 estimate the premium differences over county risk scores while controlling for the effect of HHI on the right hand side, so that we examine jointly the premium differences across risk scores and concentration. We show separate difference-in-differences estimates on low- and high-rated contracts, followed by the triple-difference estimate on high quality rating. We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis (Table 2). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D15: Effect of the payment reform on Part D premiums over provider quality, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	hospital re-admission (%)			preventable hospital stay (%)		
Treat · High · Post			0.42			0.58
			(0.30)			(0.60)
Treat · Post	-0.099	0.39	-0.077	0.16	0.68	0.15
	(0.20)	(0.25)	(0.19)	(0.33)	(0.49)	(0.33)
High · Post			2.31			2.34
			(2.01)			(2.01)
Counties		15% tails			15% tails	
Contracts	low	high	all	low	high	all
y mean	18.10	27.96	20.77	18.07	27.94	20.75
R^2	0.75	0.71	0.75	0.76	0.67	0.75
N	4,372	1,619	5,991	4,356	1,621	5,977

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in provider quality, measured by hospital re-admission for inpatient care in column 1-3, and preventable hospital stay for outpatient care in column 4-6. Risk score is positively associated with both measures, or negatively associated with quality. We show separate difference-in-differences estimates on low- and high-rated contracts, followed by the triple-difference estimate on high quality rating. We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis (Table 2). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D16: Effect of the payment reform on Part D premiums over fee-for-service (FFS) costs, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
County Variation in Treat:	Per Capita FFS Cost Unadjusted (thousands)				Per Capita FFS Cost Price-Standardized (thousands)				Per Capita FFS Cost Price-Standardized Risk-Adjusted (thousands)			
Treat · High · Post				1.39**				1.68**				1.13*
				(0.63)				(0.64)				(0.61)
Treat · Post	-0.10	0.53	1.44**	-0.062	-0.31	0.65	1.44**	-0.28	-0.094	0.79	0.62	-0.14
	(0.28)	(0.45)	(0.55)	(0.27)	(0.38)	(0.51)	(0.56)	(0.38)	(0.43)	(0.75)	(0.56)	(0.42)
High · Post				4.63*				4.61*				4.56*
				(2.41)				(2.42)				(2.42)
Counties			all				all				all	
Contracts	low	high	high +	(3) vs. (1)	low	high	high +	(7) vs. (5)	low	high	high +	(11) vs. (9)
Service area risk			<50%				<50%				<50%	
y mean	18.29	29.16	29.99	20.03	18.29	29.16	29.99	20.03	18.29	29.16	29.99	20.03
R^2	0.76	0.66	0.72	0.77	0.76	0.66	0.72	0.77	0.76	0.66	0.72	0.77
N	14,861	5,611	2,604	17,465	14,861	5,611	2,604	17,465	14,861	5,611	2,604	17,465

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in per capita fee-for-service (FFS) costs. The costs are unadjusted in column 1-4, adjusted for county differences in price levels (both input prices and reimbursement rates) in column 5-8, and further adjusted by FFS risk scores in column 9-12. In each case, we show difference-in-differences estimates on low- and high-rated contracts, as well as on high-selection contracts below the median service area risk (0.975) in the baseline, followed by the triple-difference estimate on high-selection contracts relative to the low-rated controls. We include all counties in the contract's service area. All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D17: Effect of the payment reform on Part D premiums over binary fee-for-service (FFS) costs, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
County Variation in Treat:	Per Capita FFS Cost Unadjusted (thousands)				Per Capita FFS Cost Price-Standardized (thousands)				Per Capita FFS Cost Price-Standardized Risk-Adjusted (thousands)			
Treat · High · Post				1.97***				1.94**				-0.33
				(0.68)				(0.82)				(1.27)
Treat · Post	-0.11	0.90*	1.94***	-0.088	-0.12	1.25**	1.89**	-0.11	0.81	1.12	0.38	0.78
	(0.30)	(0.52)	(0.62)	(0.29)	(0.39)	(0.60)	(0.73)	(0.38)	(0.62)	(1.02)	(1.20)	(0.62)
High · Post				4.90**				4.89**				5.08**
				(2.21)				(2.24)				(2.39)
Counties	15% tails				15% tails				15% tails			
Contracts	low	high	high +	(3) vs. (1)	low	high	high +	(7) vs. (5)	low	high	high +	(11) vs. (9)
Service area risk			<50%				<50%				<50%	
y mean	18.05	27.99	29.08	19.66	18.05	27.99	29.08	19.66	18.05	27.99	29.08	19.66
R^2	0.75	0.70	0.79	0.76	0.75	0.70	0.73	0.76	0.75	0.70	0.73	0.76
N	4,393	1,633	751	5,144	4,393	1,633	751	5,144	4,366	1,633	751	5,144

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

Notes: The table shows the within-contract differences in Part D premiums over county differences in fee-for-service (FFS) costs. The cost variable is unadjusted per capita cost in column 1-4, adjusted for county differences in price levels (both input prices and reimbursement rates) in column 5-8, and further adjusted by FFS risk scores in column 9-12. In each case, we show difference-in-differences estimates on low- and high-rated contracts, as well as on high-selection contracts below the median service area risk (0.975) in the baseline, followed by the triple-difference estimate on high-selection contracts relative to the low-rated controls. We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis (Table 2). All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D18: Effect of the payment reform on premiums and drug deductibles over coding-adjusted risk scores, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			17.05			23.57**			-25.16
			(15.87)			(9.08)			(50.11)
Risk · Post	-7.15	4.90	-9.47	-8.54	18.63**	-7.72	37.95*	32.85	37.73**
	(9.23)	(15.29)	(8.78)	(6.19)	(7.63)	(6.00)	(17.87)	(49.55)	(17.70)
High · Post			-9.15			2.40			-15.59
			(4.29)			(2.00)			(9.72)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	25.84	49.48	32.24	18.05	27.99	20.74	29.27	25.49	28.25
R^2	0.77	0.84	0.81	0.75	0.70	0.75	0.70	0.65	0.69
N	4,393	1,633	6,026	4,393	1,633	6,026	4,393	1,633	6,026

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

Notes: The table re-estimates the within-contract cross-county differences in Part C premiums (Table D7), Part D premiums (Table 2), and drug deductibles (Appendix Table D10), adjusting county risk scores with the diagnosis intensity factors developed in Finkelstein *et al.* (2017). We restrict locations to counties in the lower or upper 15% of county risk scores within the contract's service area, so that we retain the sample of contract-counties used in the main analysis. We show difference-in-differences estimates on low- and high-rated contracts, followed by the triple-difference estimate on high quality rating. All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D19: Effect of the payment reform on rebates, within-contract differences

	(I)	(II)	(III)	(IV)	(V)
Risk · High · Post				-56.32**	-89.71***
				(25.99)	(29.02)
Risk · Post	36.67*	-25.47	-58.57**	36.91*	37.50**
	(18.74)	(18.44)	(23.61)	(18.88)	(18.58)
High · Post				6.28	1.26
				(4.01)	(4.97)
Counties			15% tails		
Contracts	low	high	high +	(2) vs. (1)	(3) vs. (1)
Service area risk			<50%		
y mean	70.38	61.96	50.94	68.10	67.54
R^2	0.80	0.78	0.79	0.80	0.81
N	4,393	1,633	751	6,026	5,144

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

Notes: The table shows the within-contract differences in rebates over county risk scores. We restrict locations to the lower and upper 15% of county risk scores in the contract's service area. Column 1-2 show the difference-in-differences estimates for low- and high-rated contracts, respectively. Column 3 restricts high-rated contracts to those below the median service area risk (0.975) in the baseline, or the high-selection contracts. Column 4 (5) shows the triple-difference estimate on the differential variation in high-rated (high-selection) contracts. All regressions include contract-county fixed effects. Robust two-way clustered standard errors at the contract and county levels in parenthesis.

Table D20: Effect of the payment reform on the total premium (Part C and D), within-contract differences, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Risk · High · Post					40.67**	58.33**	71.62***
					(16.17)	(23.60)	(24.98)
Risk · Post	-12.20	26.83*	44.69*	63.74**	-13.70	-13.17	-12.60
	(9.60)	(15.44)	(23.04)	(24.70)	(9.17)	(9.29)	(9.46)
High · Post					-6.87	-4.00	-9.13
					(5.17)	(5.49)	(9.11)
Counties				15% tails			
Contracts	low	high (+ service area risk)			(2)-(1)	(3)-(1)	(4)-(1)
Service area risk			<50%	<25%			
y mean	43.89	77.47	94.55	104.24	52.99	51.28	49.22
R^2	0.85	0.86	0.85	0.81	0.86	0.87	0.86
N	4,393	1,633	751	426	6,026	5,144	4,819

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

Notes: The table shows the within-contract differences in total premiums (Part C + D) over county risk scores. We restrict the within-contract locations to the lower or upper 15% of county risk scores in the contract's service area. Column 1 and 2 show the difference-in-differences estimates for low- and high-rated contracts, respectively. Column 3 restricts high-rated contracts to those below the median service area risk (0.975) in the baseline, or the high-selection contracts. Column 4 further restricts high-selection contracts to those below the 25th percentile of service area risk (0.902) in the baseline. Column 5 shows the triple-difference estimate on the differential variation in high-rated contracts relative to the low-rated contracts. Column 6 and 7 show the tripe-difference estimates on the high-selection contracts defined in column 3 and 4, respectively. All regressions include contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D21: Effect of the payment reform on premiums and rebates, high-selection contracts

	(I) Premium (Part C+D)	(II) Zero Premium	(III) Drug Deductible	(IV) Rebate
High · Post	0.099	-0.011	-9.12	-2.59
	(4.43)	(0.031)	(11.29)	(4.55)
y mean	45.41	0.45	33.81	80.68
R^2	0.91	0.88	0.72	0.87
N	920	920	920	937

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

Notes: The table shows difference-in-differences estimates on the premiums and rebates of high-selection contracts. Rebates enhance the insurance benefits by lowering premiums and out- of-pocket costs, and by providing additional coverage such as vision and dental care. We use rebates as a summary measure of overall insurance generosity. Plan-level premiums and rebates are averaged to the contract level using enrollment weights. All regressions include contract fixed effects. Standard errors clustered at the contract level in parenthesis.

Table D22: Effect of the payment reform on premiums and rebates, high-rated contracts

	(I) Part C Premium	(II) Part D Premium	(III) Zero Premium	(IV) Rebate
High · Post	-3.29	0.47	0.032	2.72
	(3.16)	(1.63)	(0.025)	(3.95)
y mean	30.78	19.96	0.41	81.04
R^2	0.87	0.81	0.88	0.87
N	1,122	1,122	1,122	1,122

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

Notes: The table shows difference-in-difference estimates on premiums and rebates. Rebates enhance the insurance benefits by lowering premiums and out-of-pocket costs, and by providing additional coverage such as vision and dental care. We use rebates as a summary measure of overall insurance generosity. Plan-level premiums and rebates are averaged to the contract level using enrollment weights. All regressions include contract fixed effects. Standard errors clustered at the contract level in parenthesis.

Table D23: Weight of measure ratings in the overall star rating

	(1)	(2)	(3)	(4)	(5)	(6)
Measures in Rating	Outcome		Access		Process	
Rating · Post	0.53***	0.71***	-0.16***	0.078	-0.080**	0.089
	(0.045)	(0.087)	(0.030)	(0.10)	(0.039)	(0.11)
Rating	0.36***	0.30***	0.78***	0.66***	1.03***	0.78***
	(0.046)	(0.055)	(0.025)	(0.085)	(0.031)	(0.081)
Contracts	all	high	all	high	all	high
y mean	3.40	4.09	3.40	4.09	3.40	4.09
R^2	0.52	0.50	0.66	0.44	0.68	0.52
N	1,692	338	1,692	338	1,692	338

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

Notes: The table estimates the change in the contribution of outcome, access, and process measures to the overall star rating due to the weight increase in 2012. Column 1-2 estimate the contribution of outcome measures, where the weights increased from 1.0 to 3.0 in 2012. Column 3-4 estimate the contribution of access measures, where the weights increased from 1.0 to 1.5 in 2012. Column 5-6 look at the process measures where the weights remained at 1.0 after 2012. We estimate separate effects for high-rated contracts in even-numbered columns. The contribution of outcome ratings (column 2) increased substantially for high-rated contracts. The contribution of access and process ratings did not change meaningfully (column 4 and 6). Robust standard errors clustered at the contract level in parenthesis.

Table D24: Effect on outcome ratings by baseline risk scores

	(I) Outcome Mean	(II) Health Improved	(III) Diabetes & Blood Pressure
Risk · 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

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

Notes: The table shows the difference-in-differences estimates comparing the rating dynamics across contracts with different baseline risk scores. Column 1 looks at the average rating over outcome measures. Column 2-3 group the outcome measures by the source of measurement. Measures of self-reported health improvement in column 2 come from the Health Outcome Survey (HOS). Measures of managing diabetes and blood pressure conditions in column 3 come from the Healthcare Effectiveness Data and Information Set (HEDIS). All regressions include contract and year fixed effects. Standard errors clustered at the contract level in parenthesis.

Table D25: Within-contract regression coefficients of HEDIS outcome ratings on risk scores, OLS estimates

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
$riskscore_{t-3}$	0.66	-0.29	-1.60*									
	(0.67)	(0.38)	(0.87)									
$riskscore_{t-2}$				0.31	-1.12**	-2.96**						
				(0.42)	(0.47)	(1.05)						
$riskscore_{t-1}$							0.13	0.40	-0.27			
							(0.33)	(0.37)	(0.58)			
$riskscore_{t}$										-0.081	0.058	-0.28
										(0.18)	(0.12)	(0.40)
Contract	low	high	high	low	high	high	low	high	high	low	high	high
Service area risk			≤50%			≤50%			≤50%			≤50%
R^2	0.66	0.85	0.92	0.66	0.85	0.89	0.62	0.81	0.83	0.64	0.81	0.83
N	998	382	160	1,514	597	247	2,196	845	323	3,036	1,214	468

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

Notes: The table shows the within-contract differences of year-t HEDIS outcome ratings in response to risk scores in year t-3 (column 1-3), t-2 (column 4-6), t-1 (column 7-9), and year t (column 10-12). In each case, table shows separate effects for baseline low-rated (3.0-3.5 stars), high-rated (4.0 stars and above) and high-selection (service area risk score below the high-rated median 0.975) contracts. To increase statistical power, we use plan-year observations and regress contract-level HEDIS outcome ratings on plan risk scores while controlling for plan and year fixed effects. Standard errors clustered at the contract level in parenthesis.

Table D26: Effect of the payment reform on Part D premium, within-contract differences over health-adjusted diabetes prevalence rates, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Diabetes · High · Post					110.53	94.74**	124.36***
					(68.60)	(37.04)	(37.54)
Diabetes · Post	-30.92	97.78	81.62**	101.40**	-29.55	-30.09	-30.02
	(22.07)	(61.69)	(33.43)	(38.20)	(22.28)	(22.23)	(22.20)
High · Post					1.76	4.90**	6.05**
					(2.08)	(2.27)	(2.76)
Counties	15% tails						
Contracts	low	high (+ service area risk)			(2) vs. (1)	(3) vs. (1)	(4) vs. (1)
Service area risk			<50%	<25%			
y mean	18.45	28.69	29.76	33.20	21.25	20.15	19.80
R^2	0.74	0.69	0.73	0.68	0.74	0.75	0.75
N	4,400	1652	779	443	6,052	5,179	4,843

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

Notes: The table shows the within-contract differences in Part D premium over county differences in health-adjusted diabetes prevalence rates. The health-adjusted prevalence rate multiples the raw prevalence rate by the coding-adjusted county risk score. We restrict within-contract locations to counties in the lower and upper 15% tails of the baseline prevalence rate. Column 1-2 show the difference-in-differences estimates for low- and high-rated contracts, respectively. Column 3 restricts high-rated contracts to those below the median service area risk (0.975) in the baseline, or the high-selection contracts. Column 4 further restricts high-selection contracts to those in the lower 25% (less than 0.902) of service area risks in the baseline. Column 5 shows the triple-difference estimate on the differential variation in high-rated contracts relative to the low-rated contracts. Column 6-7 show the triple-difference estimates on the high-selection contracts defined in column 3 and 4, respectively. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D27: Effect of the payment reform on Part D premium, within-contract differences over health-adjusted hypertension prevalence rates, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Hypertension · High · Post					27.10	44.49***	37.04***
					(17.29)	(15.68)	(14.06)
Hypertension · Post	-5.34	24.96	40.62**	34.97**	-4.80	-4.97	-4.94
	(7.12)	(15.93)	(14.89)	(12.95)	(7.00)	(7.07)	(7.08)
High · Post					2.70	5.75**	7.08**
					(2.02)	(2.36)	(2.88)
Counties	15% tails						
Contracts	low	high (+ service area risk)			(2) vs. (1)	(3) vs. (1)	(4) vs. (1)
Service area risk			<50%	<25%			
y mean	18.21	28.35	29.38	32.81	20.98	19.86	19.53
R^2	0.75	0.69	0.74	0.68	0.75	0.76	0.76
N	4,457	1,672	771	440	6,129	5,228	4,897

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

Notes: The table shows the within-contract differences in Part D premium over county differences in health-adjusted hypertension prevalence rates. The health-adjusted prevalence rate multiples the raw prevalence rate by the coding-adjusted county risk score. We restrict within-contract locations to counties in the lower and upper 15% tails of the baseline prevalence rate. Column 1-2 show the difference-in-differences estimates for low- and high-rated contracts, respectively. Column 3 restricts higher-rated contracts to those below the median service area risk (0.975) in the baseline, or the high-selection contracts. Column 4 further restricts high-selection contracts to those in the lower 25% (less than 0.902) of service area risks in the baseline. Column 5 shows the triple-difference estimate on the differential variation in high-rated contracts relative to low-rated contracts. Column 6-7 show the triple-difference estimates on the high-selection contracts defined in column 3 and 4, respectively. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table D28: Effect of selection on health outcome measures, first-stage prediction

	(I)	(II)	(III)	(V)	(VI)
$riskiv_{ct-2}$	0.035*	-0.052***	-0.083**	-0.058***	-0.032***
	(0.020)	(0.017)	(0.032)	(0.016)	(0.010)
$diabiv_{ct-2}$	0.075**	0.031**	0.083***	0.007	0.030
	(0.038)	(0.015)	(0.031)	(0.021)	(0.019)
$hyptiv_{ct-2}$	-0.085*	0.011	-0.001	0.012	-0.016
	(0.049)	(0.024)	(0.032)	(0.020)	(0.026)
F-stat	2.11	9.12	3.54	10.09	26.35
Contracts	low	high	high	high	high
Service area risk			>50%	≤50%	≤25%
N	1,280	669	396	228	116

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

Notes: The table shows the first-stage prediction of contract risk scores $risk_{ct-2}$ from three instrumental variables: premium differences over county risk scores in $riskiv_{ct-2}$, premiums differences over diabetes prevalence rates in $diabiv_{ct-2}$, and premium differences over hypertension prevalence rates in $hyptiv_{ct-2}$. The outcome of interest in the second stage is the HEDIS health outcomes of the contract, measured in percentages of enrollees controlling chronic conditions below the medical thresholds. Robust standard errors clustered at the contract level in parenthesis.

Table D29: Effect of selection on the star ratings of outcome, access, and process measures, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)
	Outcome Ratings		Access Ratings		Process Ratings	
			Panel A: OLS			
Risk Score	-2.93*	-1.48	0.69	3.18	-2.26**	-2.06
	(1.67)	(2.97)	(2.80)	(5.05)	(0.93)	(1.78)
$\gamma_c \cdot$ Post	0.22	0.22	-0.19	-0.13	0.18	0.15
			Panel B: TSLS			
Risk Score	-17.91***	-14.47*	-2.26	-0.45	0.054	3.81
	(6.60)	(7.75)	(2.19)	(5.65)	(4.16)	(3.49)
First-stage F-stat	7.04	11.94	7.04	11.94	7.42	11.94
Over-id p-value	0.96	0.22	0.43	0.50	0.33	0.53
$\gamma_c \cdot$ Post	0.12	-0.086	-0.18	-0.19	0.17	0.20
ΔRisk $\cdot \widehat{\beta_{TSLS}}$	0.45	0.52	0.057	0.016	0.00	-0.14
Service area risk	≤50%	≤25%	≤50%	≤25%	≤50%	≤25%
y mean	3.85	3.64	4.18	4.11	3.77	3.60
N	234	122	234	122	234	122

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

Notes: The table shows the effect of risk scores on the star ratings of outcome, access, and process measures in the quality rating. Specifically, outcome measures include all measures receiving 3.0 weights in the overall star rating in a given year. Access (Process) measures include all measures receiving 1.5 (1.0) weights in the overall star rating in a given year. Panel A shows OLS estimates regressing star ratings on contract risk scores. Panel B shows two-stage-least-squares (TSLS) estimates instrumenting contract risk scores by the premium differences across counties. Specifically, we construct instrument $riskiv_{ct-2}$ to summarize premium differences by county risk scores, instrument $diabiv_{ct-2}$ to summarize premium differences by diabetes prevalence rates, and instrument $hyptiv_{ct-2}$ to summarize premium differences by hypertension prevalence rates. The instruments strongly predict risk scores in high-rated contracts (column 2) and particularly in high-selection contracts (column 4-5). For these contracts, we calculate the gains from selection from $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}}$, where ΔRisk is the risk score change (relative to low-rated contracts) after the payment reform in 2011-2012. Removing the selection gains on the star ratings, we infer quality rating improvements for a standard-risk enrollee from $\gamma_c \cdot$ Post. We also include changes in the year fixed effect τ_t after the payment reform in $\gamma_c \cdot$ Post when inferring quality improvements. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. Robust standard errors clustered at the contract level in parenthesis.

Table D30: Effect of selection on the health outcome ratings by types of measures, high-selection contracts

	(I) HEDIS	(II) Drug	(III) Self Report	(IV) HEDIS+Drug
			Panel A: OLS	
Risk Score	-2.47	-3.85	-0.35	-3.01
	(2.59)	(3.49)	(3.71)	(2.64)
$\alpha_i \cdot$ Post	0.29	0.37	-0.049	0.35
			Panel B: TSLS	
Risk Score	-21.52**	-12.12	2.19	-20.85**
	(10.20)	(9.41)	(9.60)	(10.26)
First-stage F-stat	7.04	7.04	7.04	7.04
Over-id p-value	0.63	0.38	0.40	0.99
$\alpha_i \cdot$ Post	0.24	-0.073	0.060	0.14
ΔRisk $\cdot\widehat{\beta_{TSLS}}$	0.54	0.30	-0.055	0.52
y mean	4.02	3.93	3.32	4.02
N	234	234	234	234

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

Notes: The table shows the effect of risk scores on the star ratings of health outcome measures receiving 3.0 weights in the overall rating. We estimate separate effects for HEDIS outcome ratings (column 1), drug outcome ratings from Part D (column 2), self-reported health improvement ratings from HOS (column 3), and the overall effect on HEDIS and drug outcome ratings (column 4). We focus on high-selection contracts serving less risky areas (<50% service area risk) in the table. We construct three instrumental variables to correct for selected risk scores in contracts: instrument $riskiv_{ct-2}$ summarizing premium differences by county risk scores, instrument $diabiv_{ct-2}$ summarizing premium differences by diabetes prevalence rates, and instrument $hyptiv_{ct-2}$ summarizing premium differences by hypertension prevalence rates. We show first-stage estimates for different choices of instruments in Panel A, and show corresponding two-stage-least-square (TSLS) estimates on the effect of contract risk scores in Panel B. Based on the TSLS estimates, we calculate the gains from selection from ΔRisk $\cdot\widehat{\beta_{TSLS}}$, where ΔRisk is the risk score change (relative to low-rated contracts) after the payment reform in 2011-2012. Removing the selection gains on the star ratings, we infer quality rating improvements for a standard-risk enrollee from $\gamma_c \cdot$ Post. We also include changes in the year fixed effect τ_t after the payment reform in $\gamma_c \cdot$ Post when inferring quality improvements. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. Robust standard errors clustered at the contract level in parenthesis.

Table D31: Effect of selection on the health outcome ratings, high-selection contracts

	(I)	(II)	(III)	(IV)
		Panel A: First Stage		
$riskiv_{ct-2}$	-0.032**	-0.039**	-0.045***	-0.045***
	(0.015)	(0.014)	(0.015)	(0.016)
$diabiv_{ct-2}$		0.013		0.002
		(-0.009)		(0.024)
$hyptiv_{ct-2}$			0.020**	0.018
			(0.006)	(0.024)
F-stat	4.24	4.97	10.01	7.04
		Panel B: TSLS		
Risk Score	-17.46**	-17.96***	-17.88***	-17.91***
	(7.90)	(6.84)	(6.64)	(6.60)
Over-id p-value	–	0.79	0.81	0.96
$\gamma_c \cdot$ Post	0.12	0.12	0.12	0.12
ΔRisk $\cdot\widehat{\beta_{TSLS}}$	0.44	0.45	0.45	0.45
y mean	3.85	3.85	3.85	3.85
N	234	234	234	234

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

Notes: The table shows the effect of risk scores on the star ratings of health outcome measures receiving 3.0 weights in the overall rating. We focus on high-selection contracts serving less risky areas (<50% service area risk) in the table. We construct three instrumental variables to correct for selected risk scores in contracts: instrument $riskiv_{ct-2}$ summarizing premium differences by county risk scores, instrument $diabiv_{ct-2}$ summarizing premium differences by diabetes prevalence rates, and instrument $hyptiv_{ct-2}$ summarizing premium differences by hypertension prevalence rates. We show first-stage estimates for different choices of instruments in Panel A, and show corresponding two-stage-least-square (TSLS) estimates on the effect of contract risk scores in Panel B. Based on the TSLS estimates, we calculate the gains from selection from ΔRisk $\cdot\widehat{\beta_{TSLS}}$, where ΔRisk is the risk score change (relative to low-rated contracts) after the payment reform in 2011-2012. Removing the selection gains on the star ratings, we infer quality rating improvements for a standard-risk enrollee from $\gamma_c \cdot$ Post. We also include changes in the year fixed effect τ_t after the payment reform in $\gamma_c \cdot$ Post when inferring quality improvements. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. Robust standard errors clustered at the contract level in parenthesis.

Table D32: Effect of the payment reform on benchmarks, bids, and rebates

	(I) Benchmark	(II) Bid	(III) Benchmark-Bid	(IV) Rebate
High · Post	40.56***	59.17***	-18.61***	-2.22
	(10.19)	(8.74)	(7.32)	(4.58)
y mean	903.00	787.45	115.55	82.26
R^2	0.80	0.84	0.83	0.86
N	920	920	920	920

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

Notes: The table shows difference-in-differences estimates on benchmarks, bids and rebates. We specifically examine the bidding responses of high-selection contracts in High, relative to low-rated contracts. We aggregate plan level benchmarks (inclusive of bonus adjustments), bids, and rebates (inclusive of bonus adjustments) to the contract level using enrollment weights. All regressions include contract fixed effects. Standard errors clustered at the contract level in parenthesis.

Table D33: Effect of the payment reform on market shares, across county risk scores

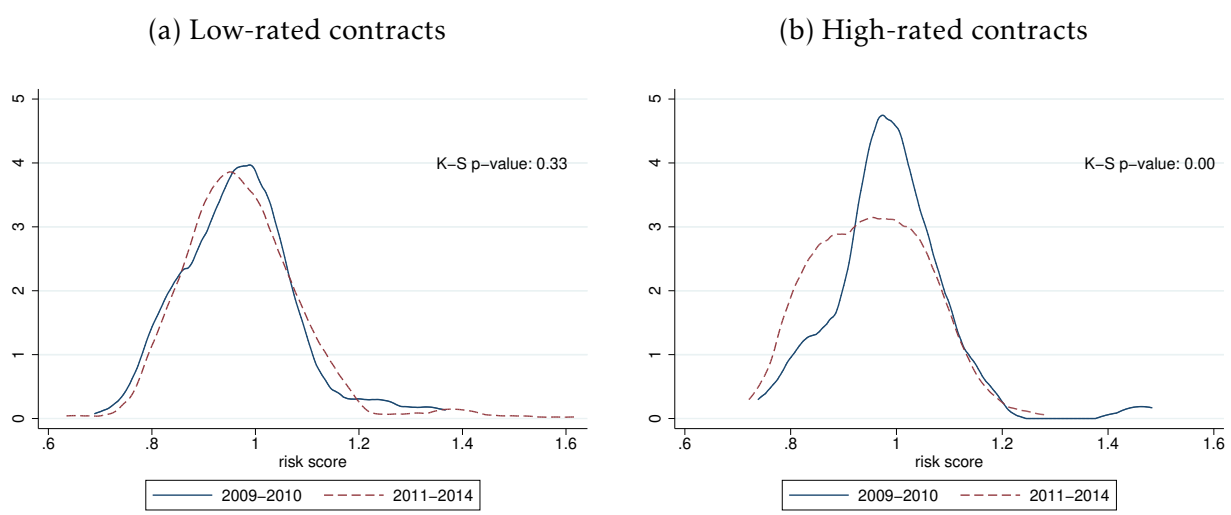
	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-0.90**			-0.19**
			(0.35)			(0.074)
Risk · Post	-0.38***	-1.18***	-0.36***	-0.14***	-0.24***	-0.095*
	(0.12)	(0.34)	(0.11)	(0.050)	(0.045)	(0.050)
High · Post			0.88**			0.15**
			(0.36)			(0.072)
Risk · High			0.65*			-0.83***
			(0.38)			(0.079)
Observations	contract-county-year			rating-county-year (balanced panel)		
Quality rating	low	high	all	low	high	all
y mean	0.31	0.38	0.33	0.28	0.13	0.20
R^2	0.64	0.64	0.58	0.73	0.76	0.33
N	15,327	5,660	21,106	17,236	17,236	34,508

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

Notes: The table shows the effect of the payment reform on the market shares of Medicare Advantage contracts across county risk scores. Column 1-3 estimates the effects on contract market shares using equation 10. Robust two-way clustered standard errors at the contract and county levels in parenthesis. Column 4-6 estimates the effect on the overall market share of high- and low-rated contracts, using a balanced panel of county-years and a specification controlling for county, year, and the rating fixed effects. Robust standard errors clustered at the contract level in parenthesis. Market shares (y mean) are lower in column 4-6 due to the incidence of zero market shares in the balanced panel.

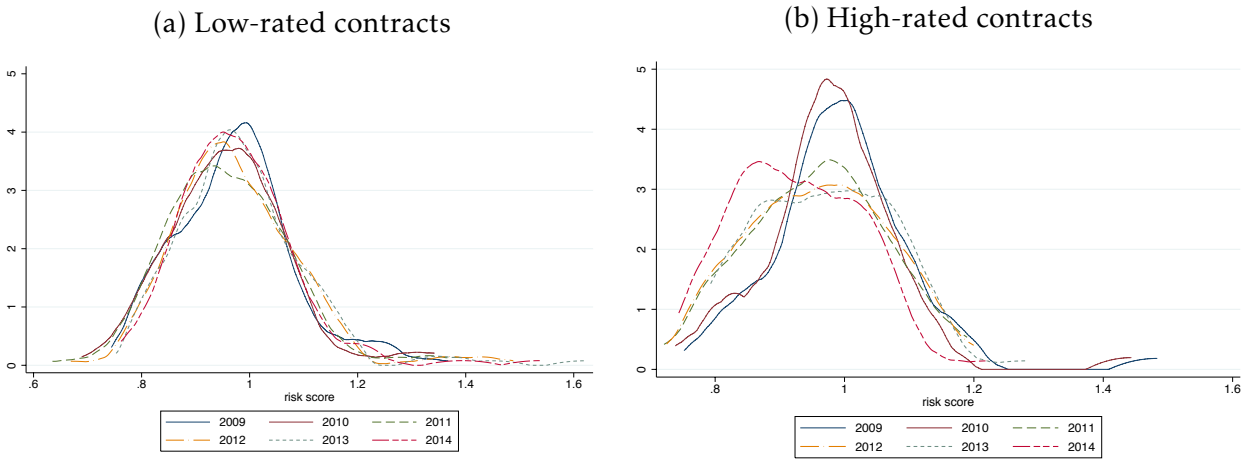
E Additional Figures

Figure E1: Effect on risk scores, kernel density



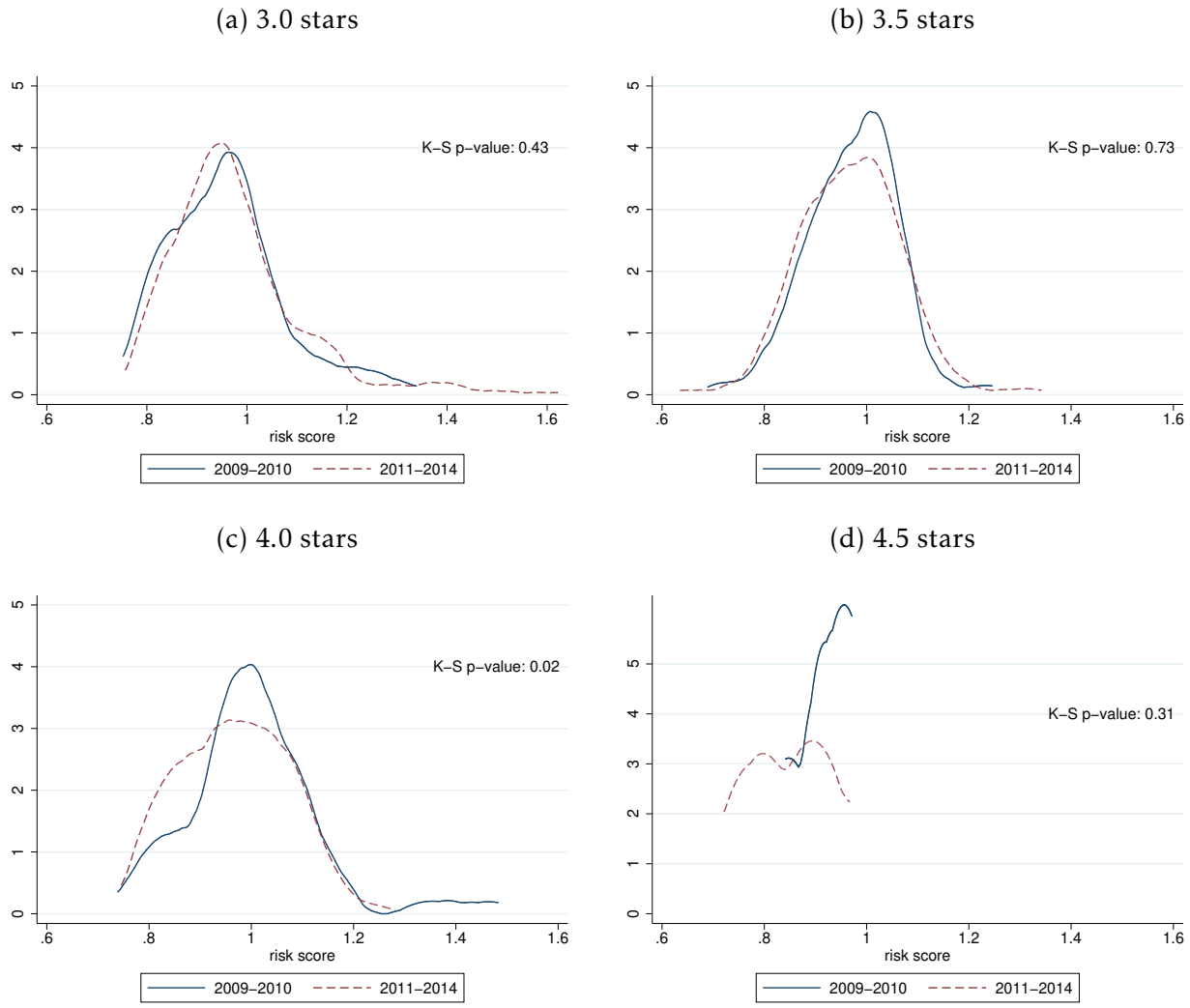
Notes: The figure plots the kernel density of risk scores for high-rated contracts in panel (a), and for low-rated contracts in panel (b). Separate densities are drawn for the before (2009-2010) and after (2011-2014) the payment reform. We test for the null of equal distribution applying the Kolmogorov–Smirnov (K-S) test, and show the p-value next to the density. Risk scores are at the contract level aggregated from plan risk scores weighted by enrollment.

Figure E2: Contract risk scores, kernel density, by star rating and year



Notes: The figure plots the kernel density of risk scores in high-rated contracts in panel (a), and the density of risk scores in low-rated contracts in panel (b). Separate density is drawn for each year. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

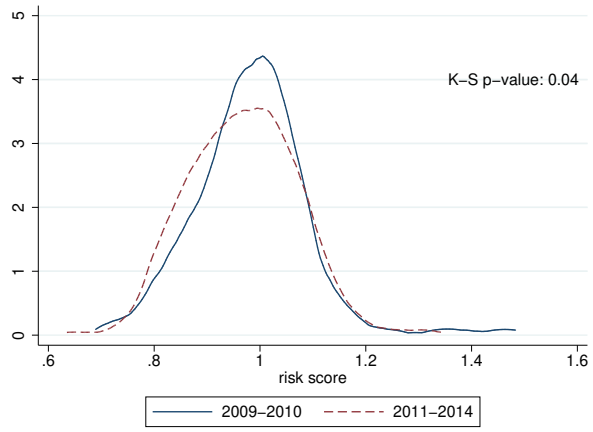
Figure E3: Changes in risk scores over time, baseline ratings computed over both Part C and Part D measures



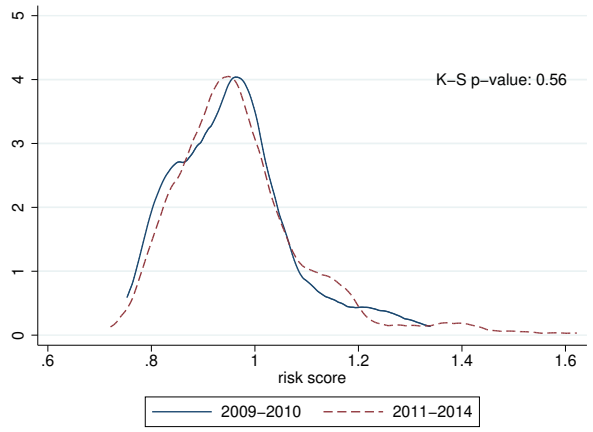
Notes: The figure plots the kernel density of risk scores across baseline star ratings. Different from the main analysis, we define the contract's baseline rating as the maximum overall rating (rather than the maximum Part C rating) in 2009 and 2010. We compute the overall star rating as the average measure-level star rating across all Part C and Part D measures in the year. We contrast the distribution of risk scores before and after the reform for baseline ratings from 3.0 stars to 4.5 stars, and test for the equality of distributions in each case showing the p-value from the Kolmogorov–Smirnov (K-S) test next to the density. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

Figure E4: Changes in risk scores over time, distance to the 4.0-star cut-off

(a) Distance to cut-off ≤ 0.5 stars



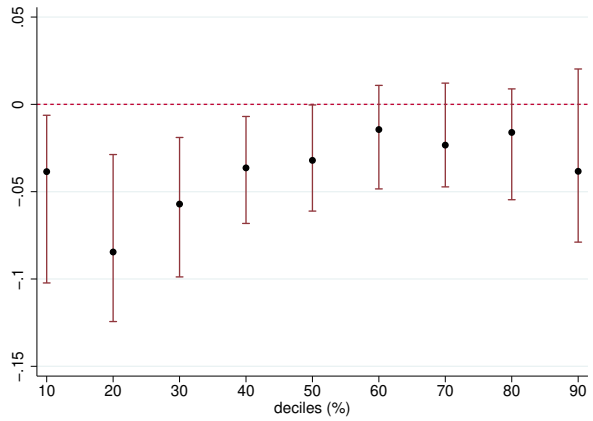
(b) Distance to cut-off: (0.5, 1] stars



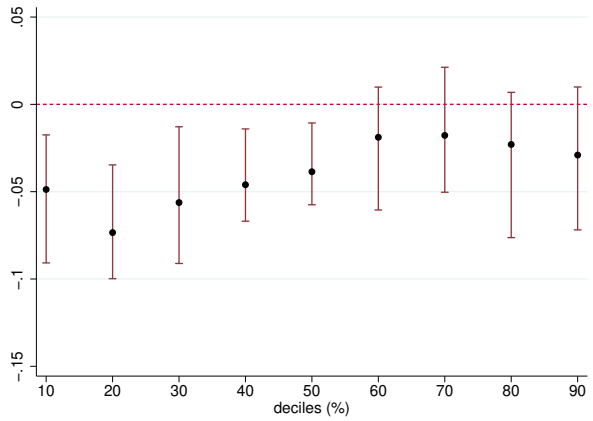
Notes: The figure plots the kernel density of risk scores by the proximity to 3.75 stars, the cut-off for a 4.0-star rating. We construct the continuous rating using the average across Part C and Part D measures, and group contracts comparing the maximum continuous rating in 2009-2010 with the 3.75 cut-off. We plot the distribution of risk scores before and after the reform for contracts less than half-star away from the cut-off in panel (a), and for more distant contracts less than one-star away in panel (b). We test for the equality of distributions in each case showing the p-value from the Kolmogorov–Smirnov (K-S) test next to the density. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

Figure E5: Distributional effects on risk scores, by deciles

(a) Difference-in-Differences



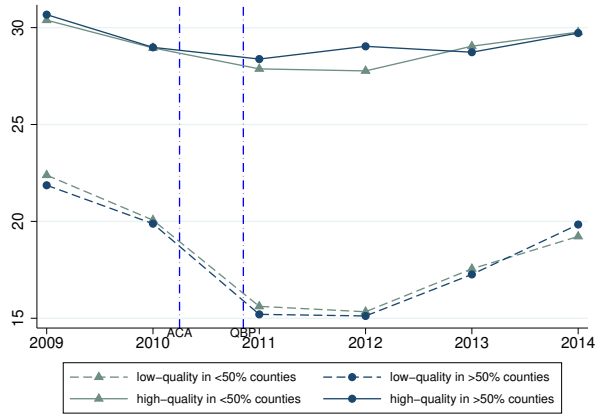
(b) Changes-in-Changes



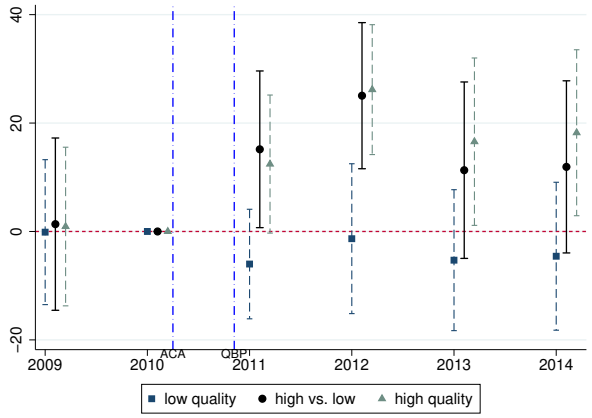
Notes: The figure plots the effect of the payment reform across deciles of risk scores in high-rated contracts. Panel (a) shows estimates from the grouped quantile approach of Chetverikov *et al.* (2016). Panel (b) plots the changes-in-changes estimates following Athey and Imbens (2006). In both cases, plotted 95% confidence intervals are based on the empirical distribution of estimates from 500 replication samples block-bootstrapped by contracts. Appendix Table D4 shows the corresponding point estimates and standard errors of the plotted effects, and compares these effects with the baseline risk scores in high-rated contracts.

Figure E6: Effect on Part D premiums, within-contract differences, event study

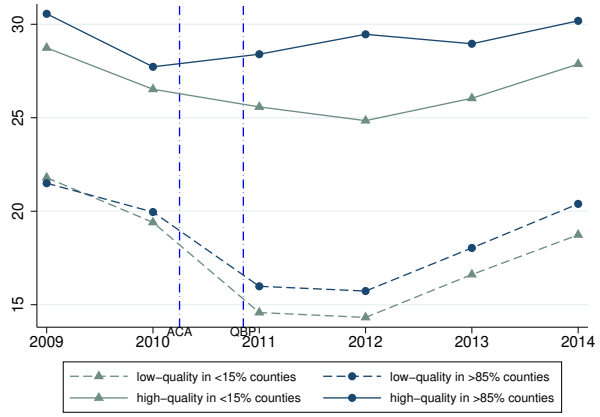
(a) Raw trend



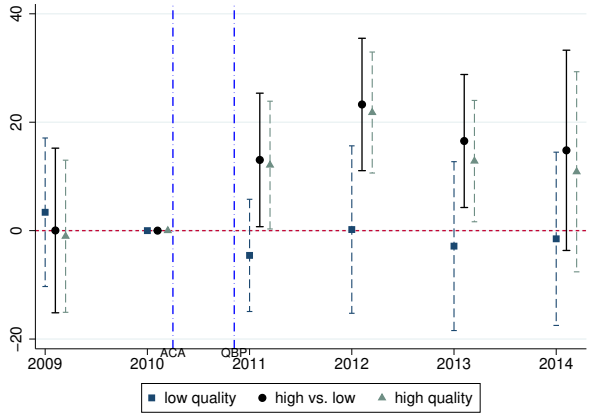
(b) Event study



(c) Raw trend, 15% tails



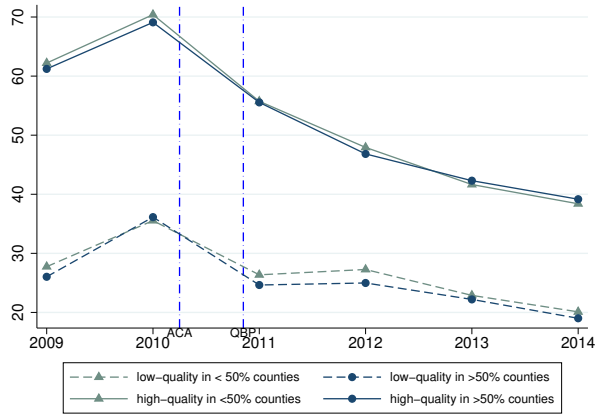
(d) Event study, 15% tails



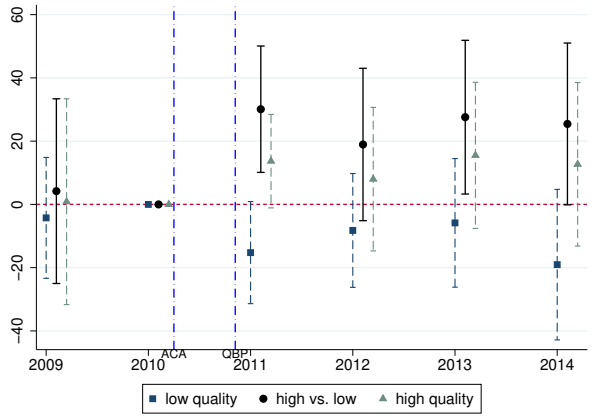
Notes: The figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score, and plot premium levels across 15% tails for an average low-rated contract (dotted lines) and an average high-rated contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure E7: Effect on Part C premiums, within-contract differences, event study

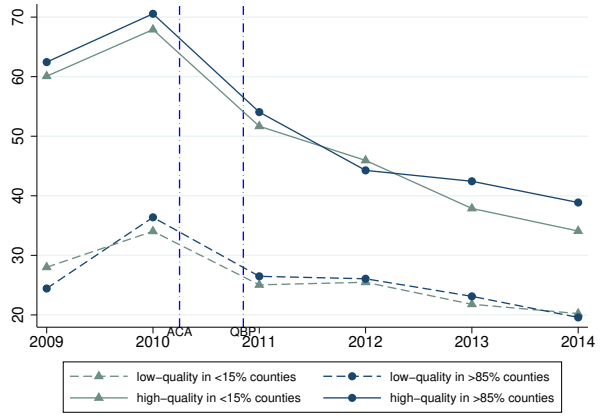
(a) Raw trend



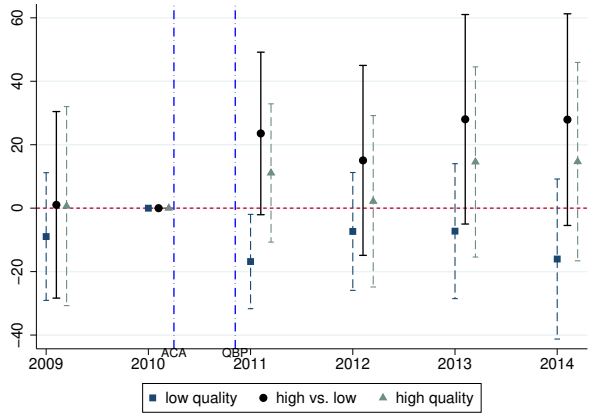
(b) Event study



(c) Raw trend, 15% tails



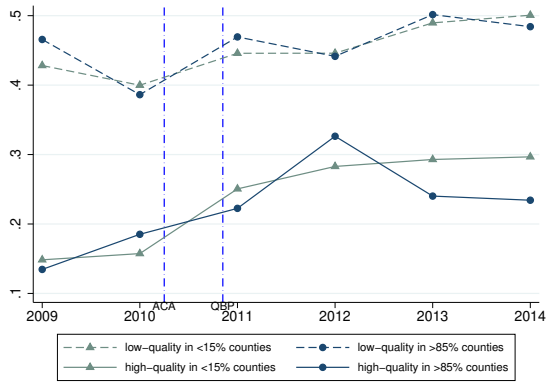
(d) Event study, 15% tails



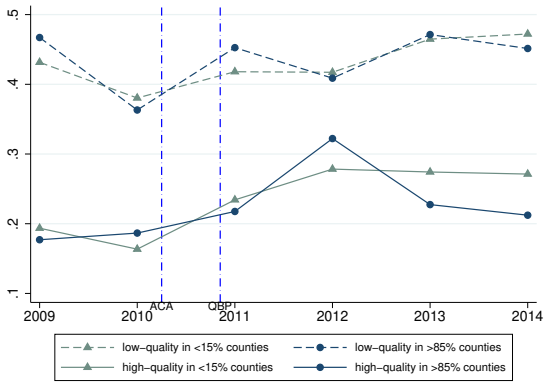
Notes: The figure plots the raw trends of Part C premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score, and plot premium levels across 15% tails for an average low-rated contract (dotted lines) and an average high-rated contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure E8: Effect on zero-premium and zero-deductible plans and enrollment, 15% tails, raw trends

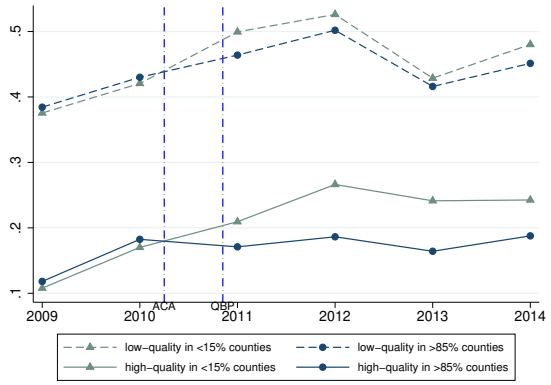
(a) Zero Part C premium



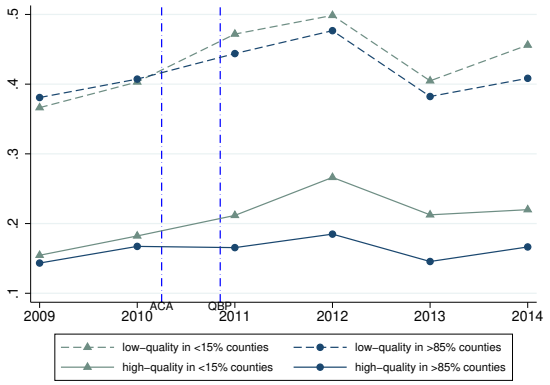
(b) Zero Part C premium, unweighted



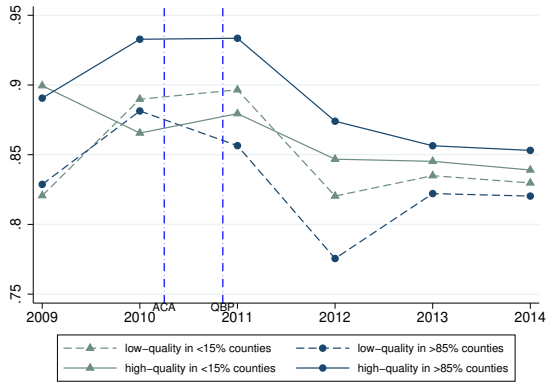
(c) Zero Part D premium



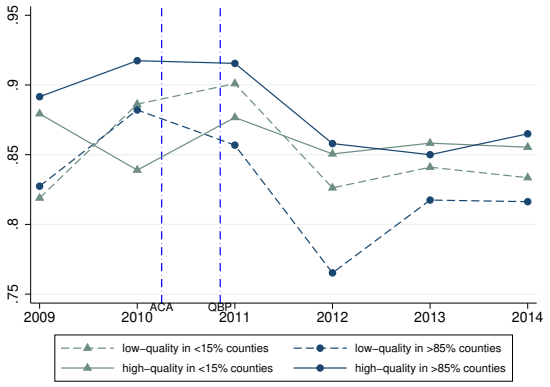
(d) Zero Part D premium, unweighted



(e) Zero drug deductible

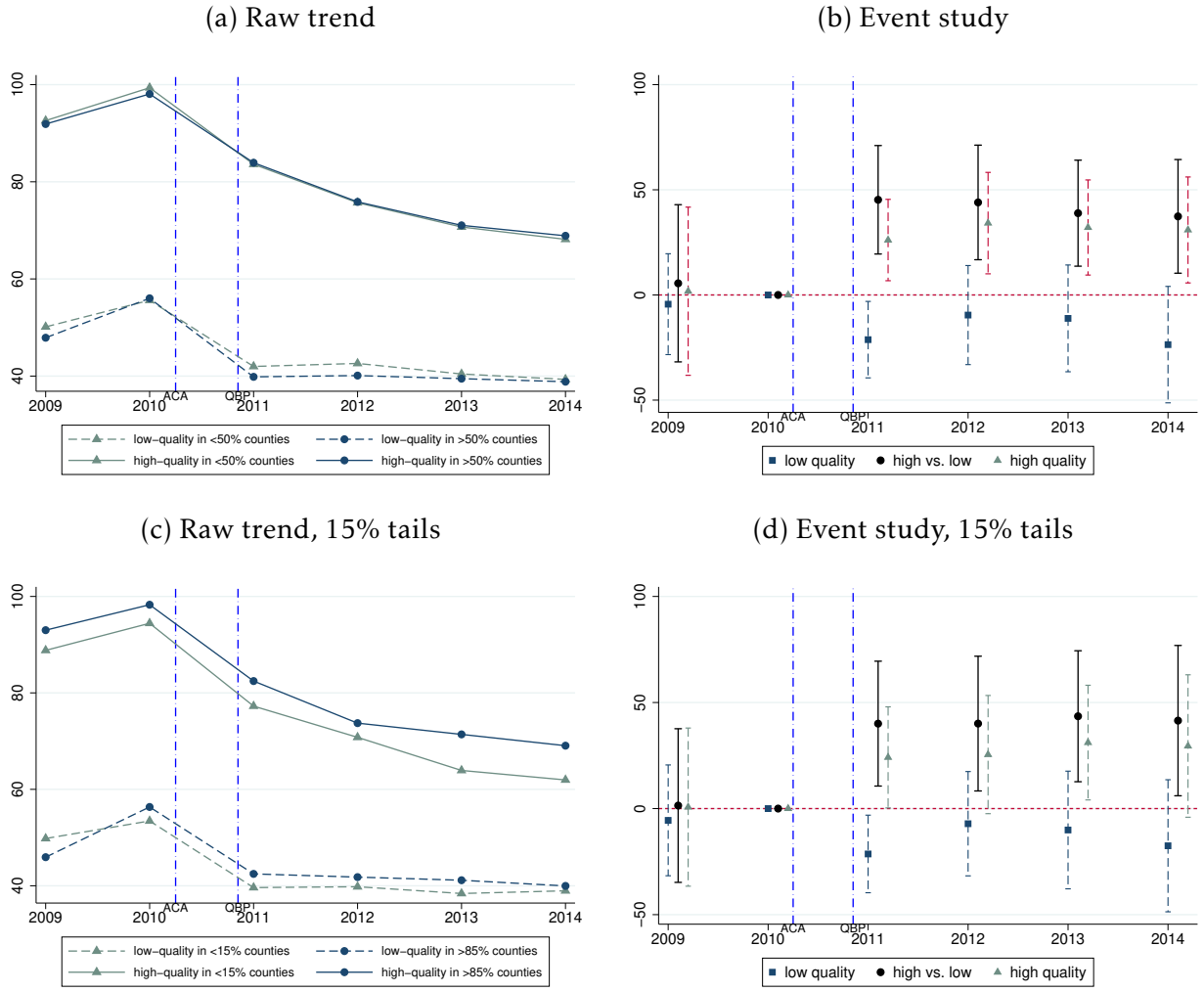


(f) Zero drug deductible, unweighted



Notes: The figure plots the raw trends of zero-premium and zero-deductible plans in the left panels, and similar trends without weighting by enrollment in the right panels. Specifically, outcome variables in the left panels are the percent of zero-premium or zero-drug deductible plans offered by the contract in a contract-county pair, weighted by enrollment. In the right panels, the percent of plans with zero premiums or zero drug deductibles is not weighted by enrollment. We restrict locations to counties in the lower or upper 15% tails of county risk score in the contract's service area, and plot the share of zero-premiums and zero-drug deductible plans across the 15% risk tails for an average low-rated contract (dotted lines) and an average high-rated contract (solid lines).

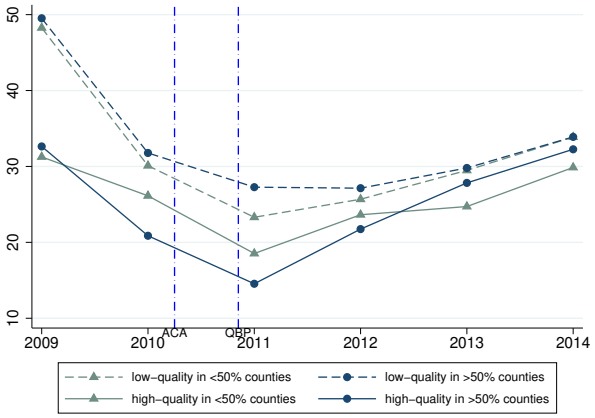
Figure E9: Effect on the total premium (Part C and D), within-contract differences, event study



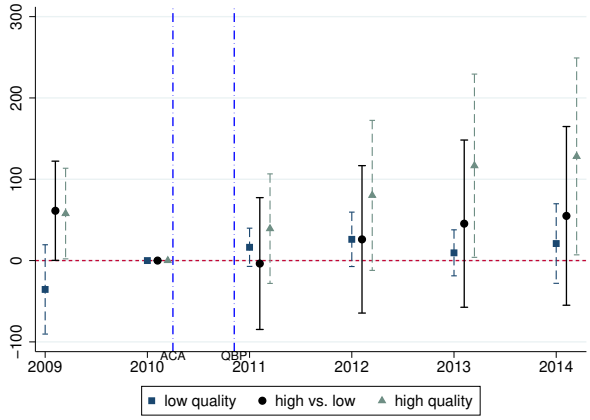
Notes: The figure plots the raw trends of total premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score, and plot premium levels across 15% tails for an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Corresponding event study estimates in panel (b) and (d) show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure E10: Effect on drug deductibles, within-contract differences, event study

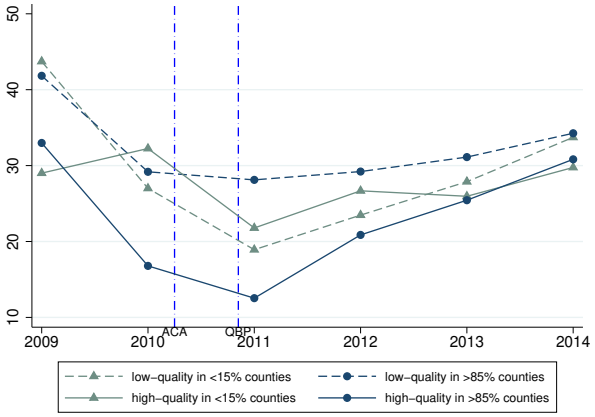
(a) Raw trend



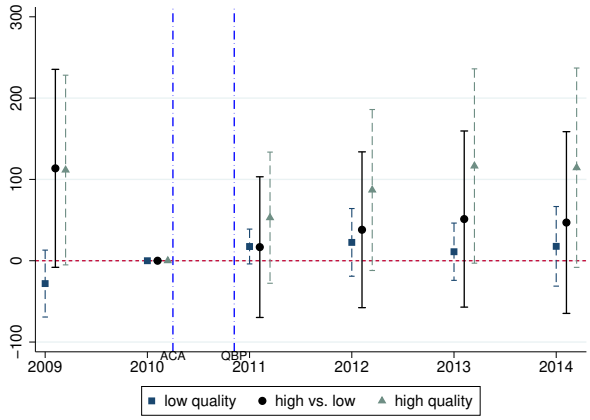
(b) Event study



(c) Raw trend, 15% tails



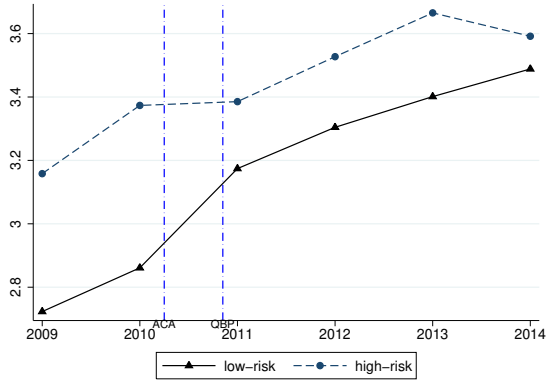
(d) Event study, 15% tails



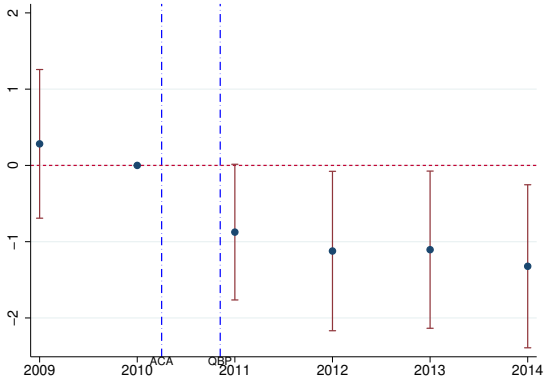
Notes: The figure plots the raw trends of drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the price levels above and below the median risk county within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score, and plot price levels across 15% tails for an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Corresponding event study estimates in panel (b) and (d) show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure E11: Outcome ratings by baseline enrollee risk scores, event study

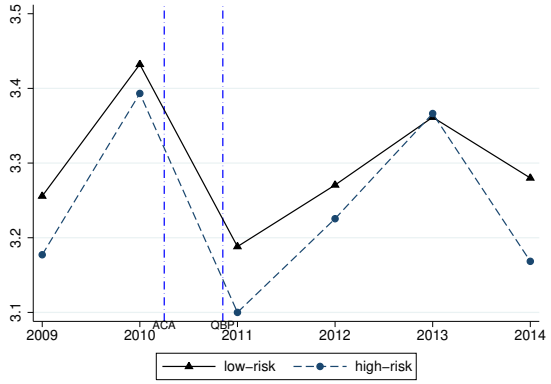
(a) Average outcome rating, raw trend



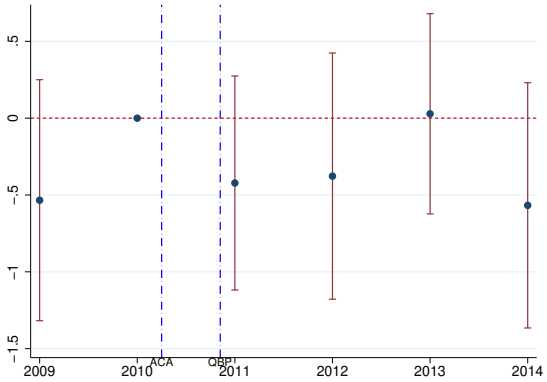
(b) Average outcome rating, event study



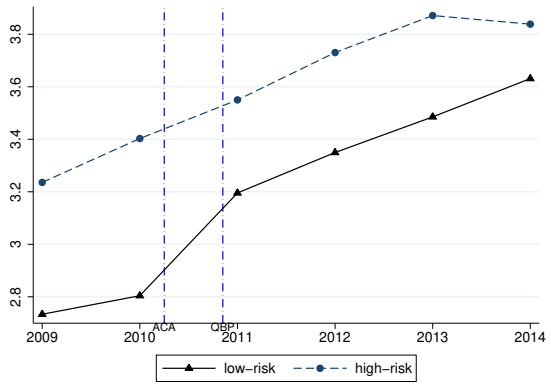
(c) Health improved, raw trend



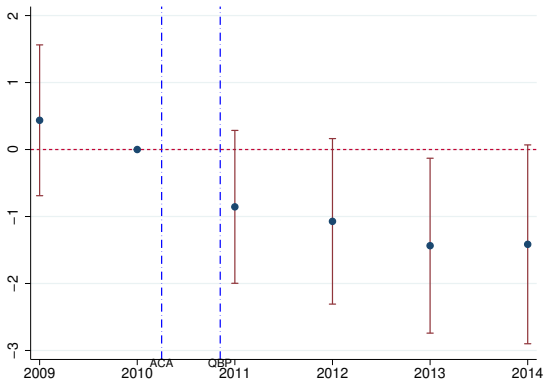
(d) Health improved, event study



(e) Diabetes & blood pressure, raw trend



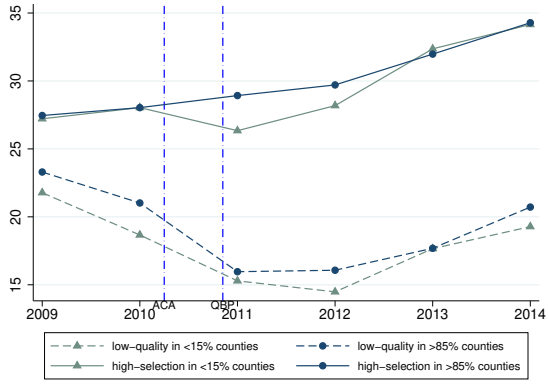
(f) Diabetes & blood pressure, event study



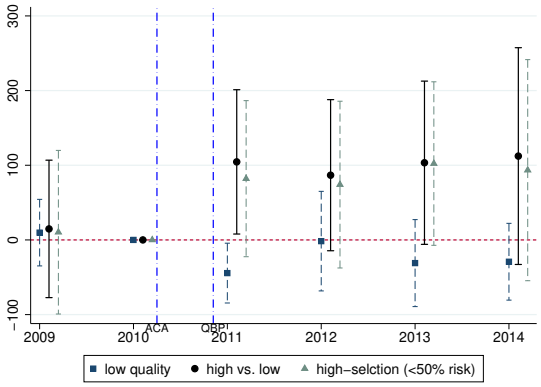
Notes: The figure shows the dynamics of outcome ratings by baseline enrollee risk scores. The raw trends in the left panels plot separate trends for binary groups of contracts above and below the median enrollee risk score (0.97) in the baseline. The right panels show event study estimates from difference-in-differences specifications in the baseline risk score. Panel (a) and (b) look at the average rating of outcome measures. Panel (c) and (d) look at the health improvement measures reported in the Health Outcome Survey (HOS). Panel (e) and (f) look at measures of managing diabetes and blood pressure from the Healthcare Effectiveness Data and Information Set (HEDIS). Event study graphs show 95% confidence intervals based on robust standard errors clustered at the contract level.

Figure E12: Effect on Part D premium, within-contract differences over health-adjusted diabetes prevalence rates, high-selection contracts, event study

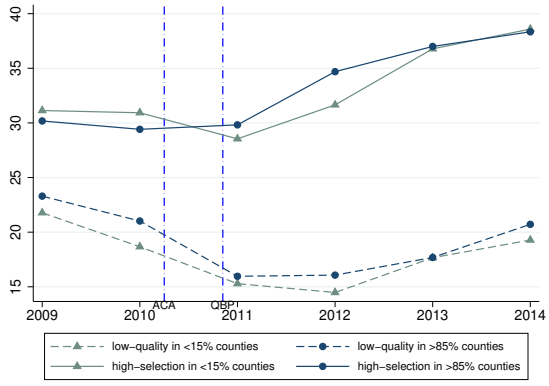
(a) High-Selection (<50% Risk), raw trend



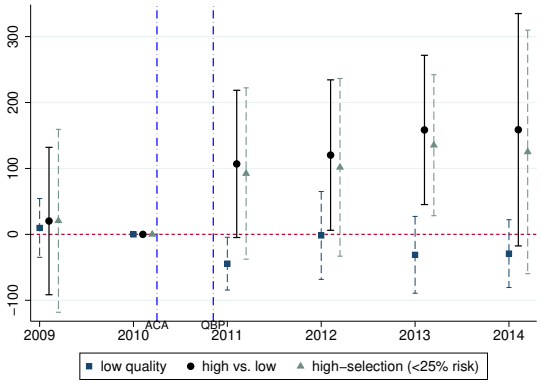
(b) High-Selection (<50% Risk), event study



(c) High-Selection (<25% Risk), raw trend



(d) High-Selection (<25% Risk), event study

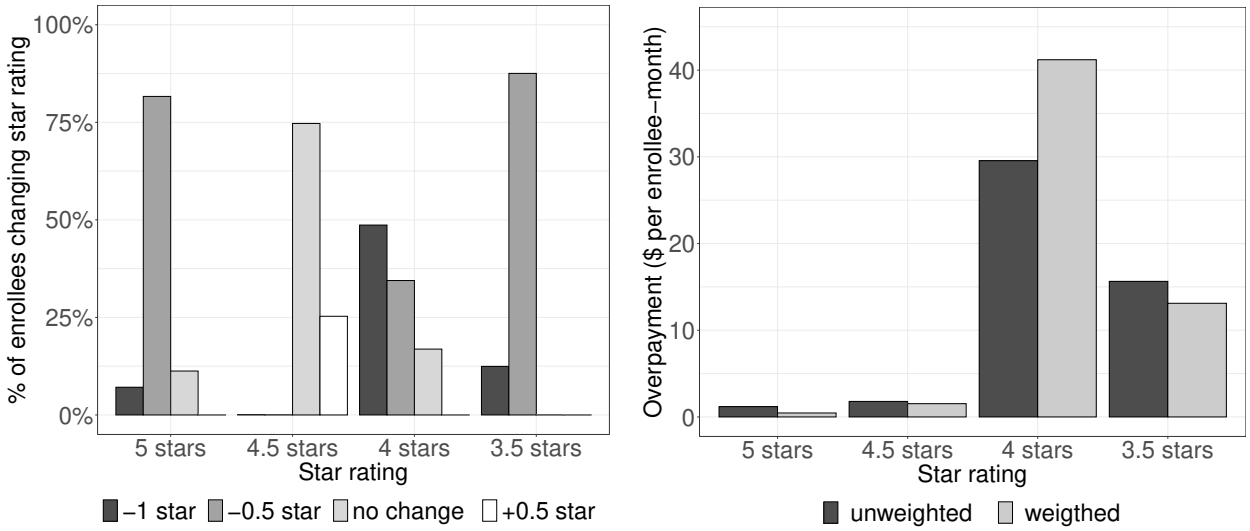


Notes: The figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract differences over county differences in health-adjusted diabetes prevalence rates in the right panels. The health-adjusted prevalence rate multiplies the raw prevalence rate by the coding-adjusted county risk score. We restrict within-contract locations to counties in the lower and upper 15% of baseline prevalence rates in the contract's service area. The raw trends plot the price levels across the 15% tails within an average low-rated contract (dotted lines) and an average high-selection contract (solid lines) below the median service area risk (0.975) in panel (a), and below the 25th percentile (0.902) in panel (c). Corresponding event study estimates in panel (b) and (d) show the within-contract differences over county differences in continuous prevalence rates. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure E13: Effects of selection on the quality rating and overpayments, synthetic control

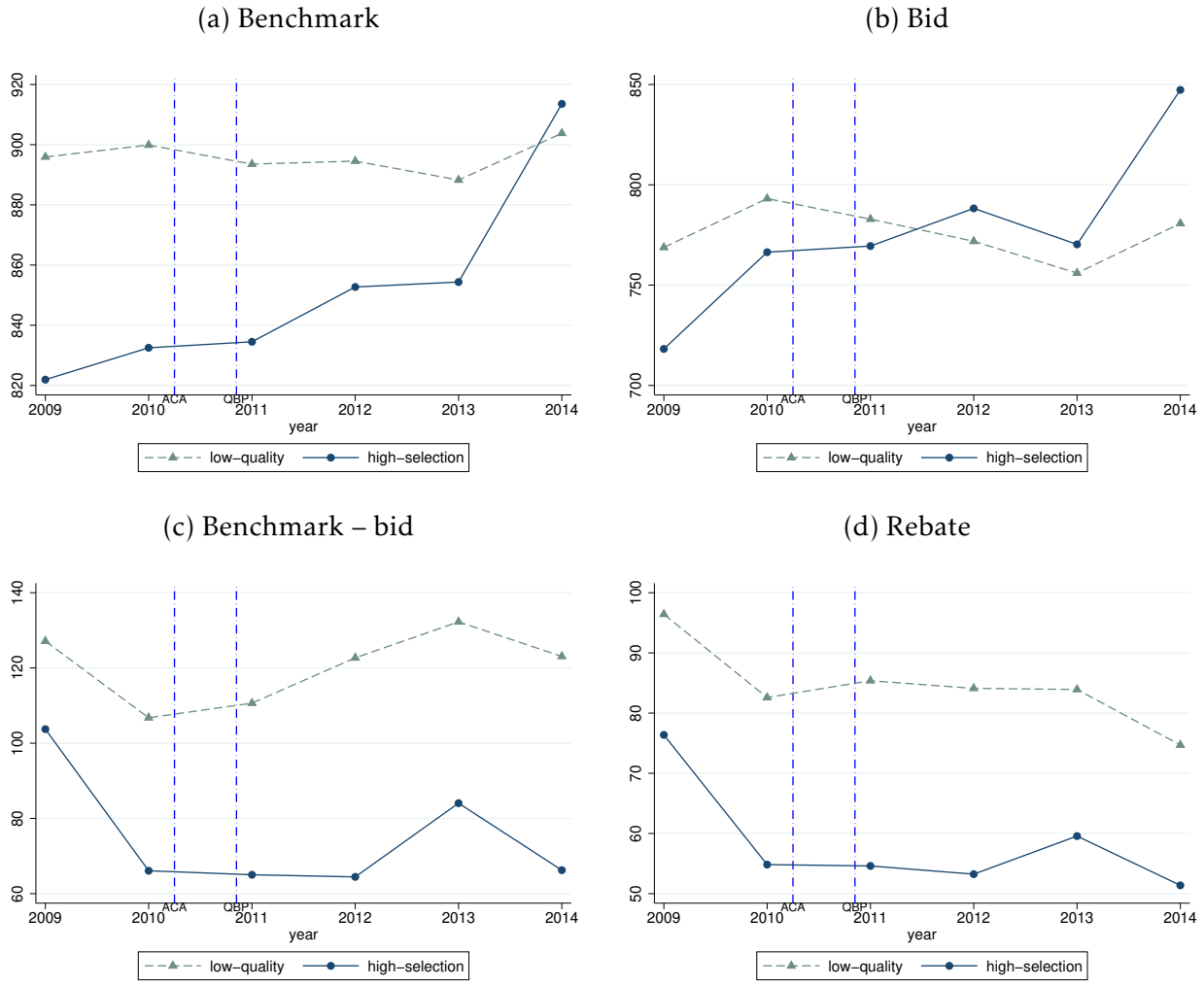
(a) Share of enrollees with star rating change absent selection

(b) Overpayments due to selection



Notes: The figure shows the effect of risk adjustment on the overall star ratings of high-selection contracts in panel (a) and on the payments to these contracts in panel (b). For different star ratings in 2014 (horizontal axis), panel (a) shows the percentage of enrollees receiving lower (by 1 star or 0.5 star) or higher (by 0.5 star or unchanged) star ratings upon adjustment for selected risk scores. The adjustment holds the risk composition at the 2010 level (corresponding to 2012 rating), and re-calculates the star rating discarding the effect of selected risk scores since 2011. Different from the main analysis, we estimate the risk score change for each high-rated contract using a weighted average of low-rated contracts as the synthetic control (Abadie *et al.*, 2010). Based on the changes in panel (a), panel (b) shows changes in 2015 payments by the 2014 star rating. We assume that contracts receiving a downgrade (upgrade) in the star rating adjust bids downward (upward) relative to the new benchmarks so that rebates to enrollees remain unchanged. The assumption is supported by our empirical analysis of bidding and pricing strategies by high-selection contracts after the payment reform. Overpayments are the amount saved when the effect of selected risk scores since 2011 is removed from the star rating. We show overpayments by 2014 star ratings with and without weighting by enrollment.

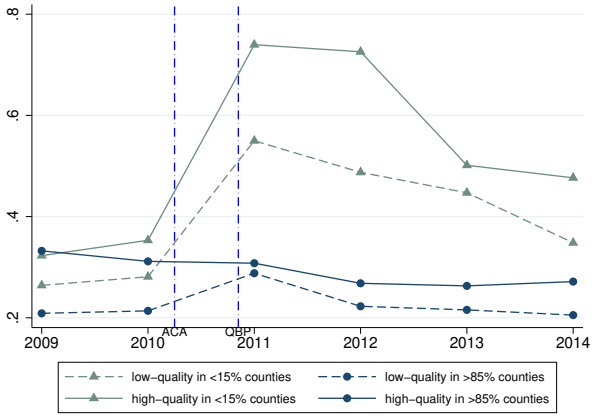
Figure E14: Effect on benchmarks, bids, and rebates, raw trends



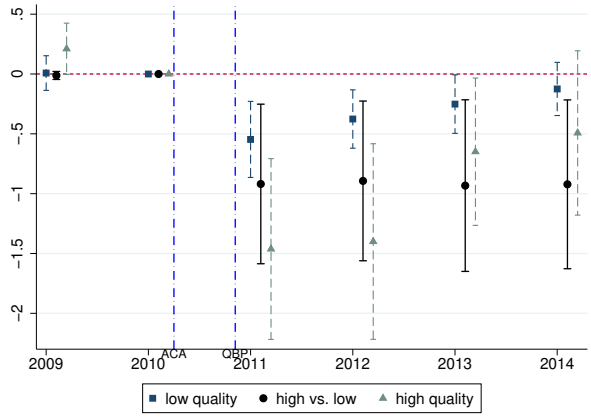
Notes: The figure shows the raw trends of benchmarks (panel a), bids (panel b), the difference between benchmarks and bids (panel c), and rebates (panel d) for high-selection contracts and low-rated contracts. All variables are at the level of contracts aggregated from plan variables weighted by enrollment. All prices are for a standard-risk enrollee. Benchmarks and rebates are inclusive of bonus adjustments.

Figure E15: Effect on market shares, cross-county differences, event study

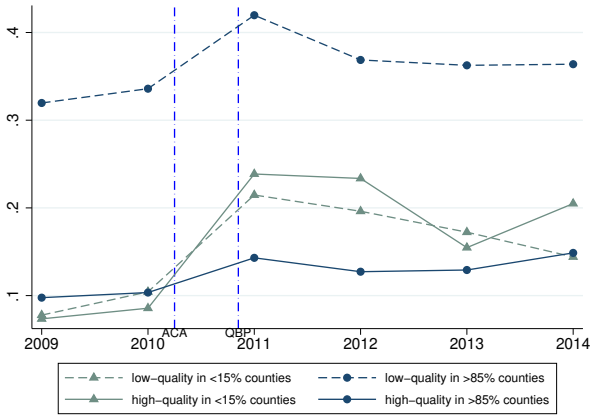
(a) Contract-County-Year, raw trend, 15% tails



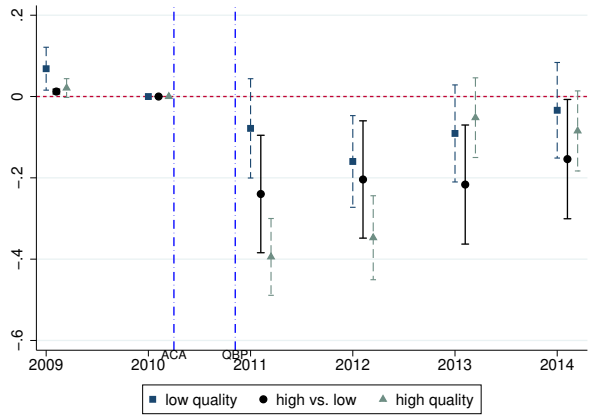
(b) Contract-County-Year, event study, All Counties



(c) Rating-County-Year, raw trend, 15% tails



(d) Rating-County-Year, event study, All Counties



Notes: The figure shows the cross-county differences in market shares of contracts in panel (a) and (b), and by high and low quality ratings (across 4.0 stars in the baseline) in panel (c) and (d), where we examine changes in the overall market share of high- and low-rated contracts using a balanced panel of county-years. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels in panel (b), and based on robust standard errors clustered by counties in panel (d).

F Sensitivity Analysis

F.1 Alternative Enrollment Weights

In the main analysis, we weight plan premiums by enrollment to generate premiums for contracts. The resulting variables capture the joint effects of insurer price-setting and enrollment responses to prices. Alternatively, to isolate premium differences due to insurer price-setting, we construct premiums taking simple averages across plans, and find similar effects across county risk scores in Appendix Table F1 and Appendix Figure F1. We further examine the sensitivity of premium differences to outliers by using the median plan price as the contract price. The median price shows similar differences across county risk scores as in the main analysis (Appendix Table F2, Appendix Figure F2).

F.2 Alternative Risk Measures

We show the robustness of results to alternative measures of risk differences across counties. Although the main analysis uses the deviation-to-median measure, we find similar differences over risk scores using the deviation-to-mean measure in Appendix Tables F3. Appendix Figure F3 plots the event study estimates for this set of estimates. We also examine alternative measures of risk tails. Instead of percentiles, Appendix Table F4 looks at risk tails defined in terms of standard deviations from the mean. We find larger differences in Part D premiums across the more remote risk tails.

Table F1: Effect of the payment reform on premiums and drug deductibles, within-contract differences, unweighted by enrollment

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			21.82			19.51***			-15.68
			(13.21)			(7.00)			(43.52)
Risk · Post	-10.85	12.86	-10.56	-4.97	19.60***	-3.93	62.04**	73.44*	66.65**
	(7.78)	(12.49)	(7.69)	(4.86)	(6.33)	(4.85)	(25.30)	(42.80)	(25.78)
High · Post			-6.74*			1.16			-11.85
			(3.68)			(2.16)			(11.80)
Counties		all			all			all	
Contracts	low	high	all	low	high	all	low	high	all
y mean	27.79	50.96	34.14	19.31	29.89	22.21	32.07	28.36	31.05
R^2	0.76	0.80	0.79	0.74	0.66	0.74	0.68	0.50	0.64
N	14,861	5,611	20,472	14,861	5,611	20,472	14,861	5,611	20,472

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

Notes: The table shows the within-contract differences in Part C premiums (column 1-3), Part D premiums (column 4-6), and drug deductibles (column 7-9) over county risk scores. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. We first show difference-in-differences estimates for low- and high-rated contracts, respectively, before showing the differential effects on high-rated contracts. We include all counties in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table F2: Effect of the payment reform on median premiums and drug deductibles, within-contract differences

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			26.87**			18.51**			-1.32
			(12.83)			(7.58)			(46.39)
Risk · Post	-15.96*	12.00	-15.82*	-4.70	18.62***	-3.77	38.26	63.15	42.61*
	(9.18)	(11.76)	(8.97)	(5.24)	(6.17)	(5.18)	(23.86)	(46.72)	(24.33)
High · Post			-8.30**			0.82			-13.73
			(4.13)			(2.13)			(10.07)
Counties		all			all			all	
Contracts	low	high	all	low	high	all	low	high	all
y mean	27.19	50.32	33.53	19.42	29.88	22.29	29.82	23.91	28.20
R^2	0.73	0.82	0.77	0.73	0.67	0.73	0.67	0.53	0.64
N	14,861	5,611	20,472	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: The table shows the within-contract differences in premiums and drug deductibles over county risk scores. Different from the main analysis, we aggregate plan prices to the contract-county level using the median plan price. We restrict within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. Column 1-2 show the difference-in-differences estimates of Part C premium in low- and high-rated contracts, respectively. Column 3 shows the triple-differences estimate on the differential variation in high-rated contracts. Column 4-6 (7-9) repeat the analysis for Part D premium (drug deductible). All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table F3: Effect of the payment reform on Part C premiums, within-contract differences, deviation to mean

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			16.45			17.29*			-30.01
			(17.68)			(9.41)			(54.87)
Risk · Post	-8.02	3.66	-10.16	-3.82	16.66**	-3.18	34.14**	26.43	38.43**
	(7.46)	(17.29)	(7.04)	(5.50)	(8.35)	(5.28)	(15.81)	(54.02)	(16.52)
High · Post			-9.18**			2.39			-15.57
			(4.27)			(1.99)			(9.76)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	25.84	49.48	32.23	18.05	27.99	20.74	29.27	25.49	28.25
R^2	0.77	0.84	0.81	0.75	0.70	0.75	0.70	0.65	0.69
N	4,393	1,633	6,026	4,393	1,633	6,026	4,393	1,633	6,026

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

Notes: The table shows the within-contract differences in Part C premiums (column 1-3), Part D premiums (column 4-6), and drug deductibles (column 7-9) over county risk scores. Differences in risk scores are measured as the deviation to the mean county risk in the service area, as opposed to the deviation-to-median measure in the main analysis. We show differences across county risks for low- and high-rated contracts, respectively, before showing the differential effect on high-rated contracts. We restrict locations to counties in the lower and upper 15% of county risks in the contract's service area. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Table F4: Effect on premiums and drug deductibles, within-contract differences, standard deviations from mean county risk

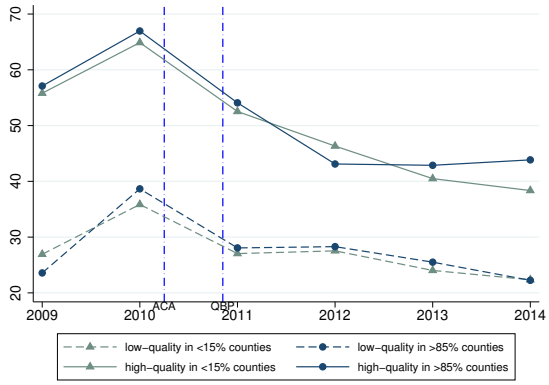
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
Risk · High · Post			10.78			15.39*			20.52**
			(6.99)			(8.98)			(10.05)
Risk · Post	-3.05	11.44*	-2.32	-8.82	15.26**	-6.90	-11.18	24.11**	-8.23
	(5.67)	(5.80)	(5.48)	(7.11)	(7.19)	(7.14)	(9.27)	(10.66)	(9.18)
High · Post			2.65			3.37			2.81
			(2.08)			(2.12)			(2.74)
Counties	deviation to mean > s.d.			deviation to mean > 1.5 s.d.			deviation to mean > 2 s.d.		
Contracts	low	high	all	low	high	all	low	high	all
y mean	17.59	28.98	20.66	17.45	28.96	20.28	19.64	22.89	20.39
R^2	0.75	0.69	0.75	0.76	0.69	0.76	0.76	0.80	0.76
N	4,386	1,615	6,001	1,787	583	2,370	637	192	829

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

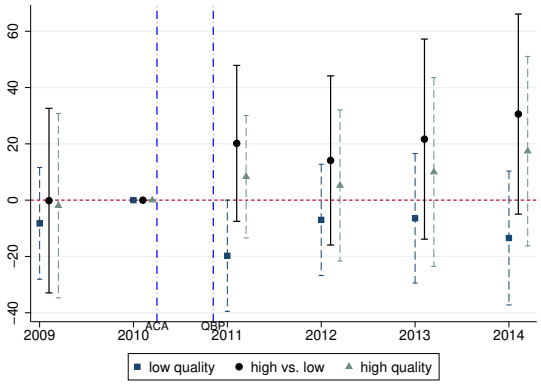
Notes: The table shows the within-contract differences in Part D premiums over county risk scores, where we include counties more than one standard deviation away from the mean county risk in column 1-3, more than 1.5 standard deviations away in column 4-6, and more than 2 standard deviations away in column 7-9. We show the difference-in-differences estimates for low- and high-rated contracts respectively, before showing the triple-difference estimate on high-rated contracts. All regressions control for contract-county fixed effects. Two-way clustered standard errors at the contract and county levels in parenthesis.

Figure F1: Effect on premiums and drug deductibles, within-contract differences, event study, unweighted by enrollment

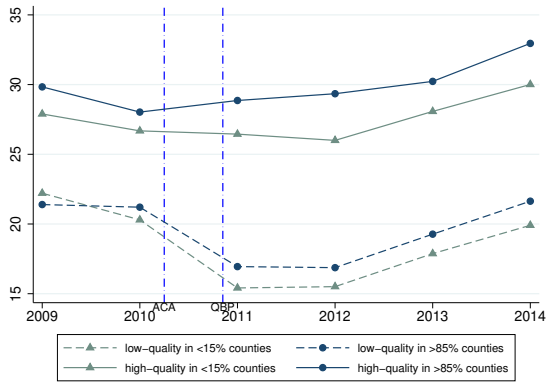
(a) Part C premium, raw trend



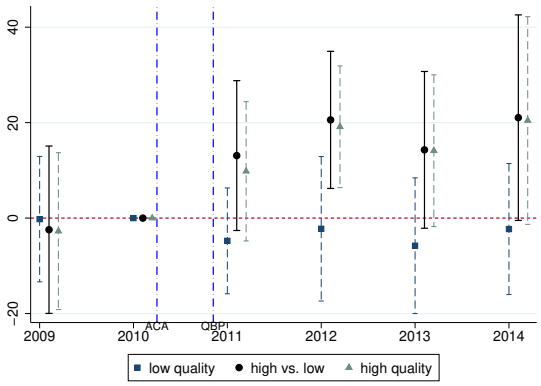
(b) Part C premium, event study



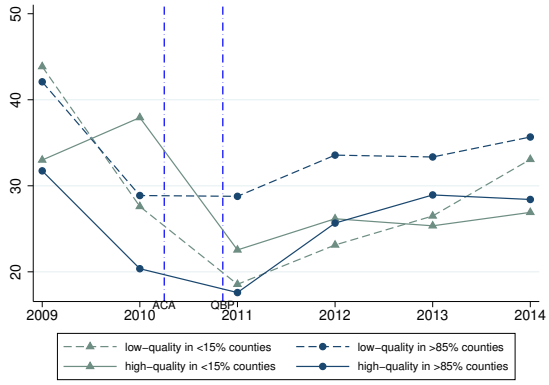
(c) Part D premium, raw trend



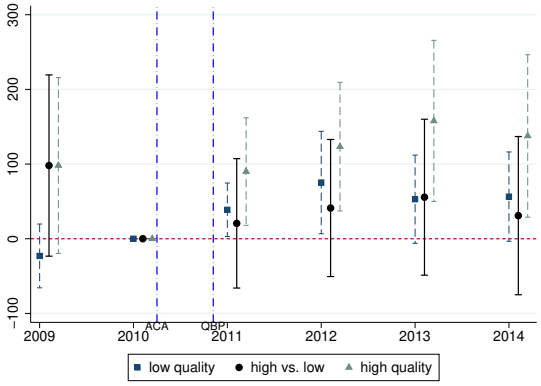
(d) Part D premium, event study



(e) Drug deductible, raw trend



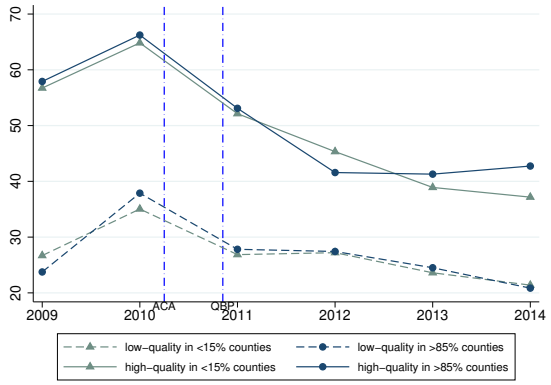
(f) Drug deductible, event study



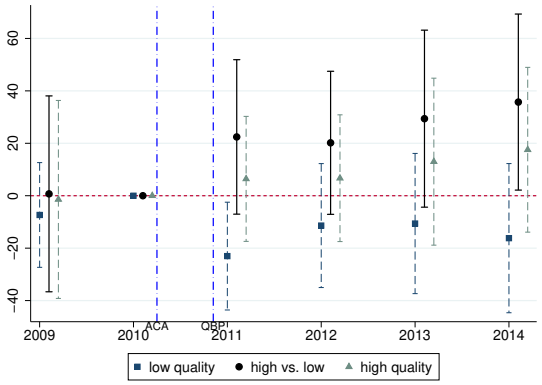
Notes: The figure plots the raw trends of premiums and drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. Different from the main analysis, we aggregate plan prices to the contract-county level taking simple averages. We restrict locations to the lower and upper 15% tails of county risk scores in the contract's service area. The raw trends plot the price levels across the 15% risk tails within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Corresponding event study estimates in the right panels show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure F2: Effect on median premiums and drug deductibles, within-contract differences, event study

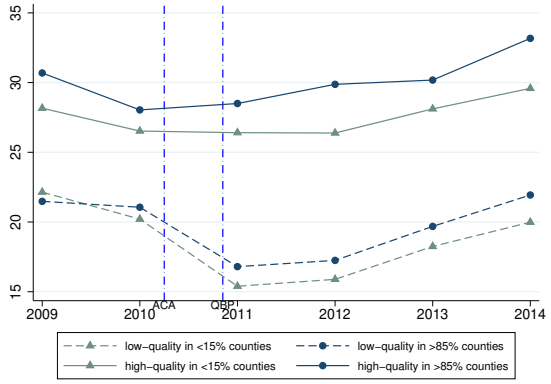
(a) Part C premium, raw trend



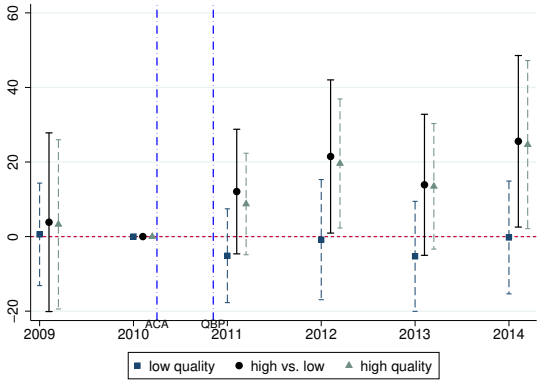
(b) Part C premium, event study



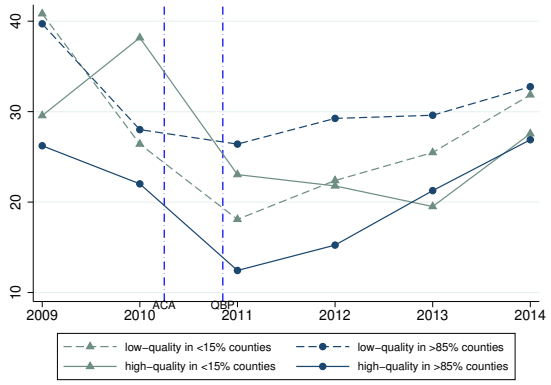
(c) Part D premium, raw trend



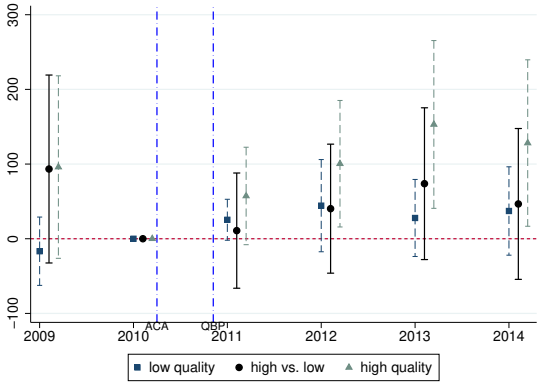
(d) Part D premium, event study



(e) Drug deductible, raw trend



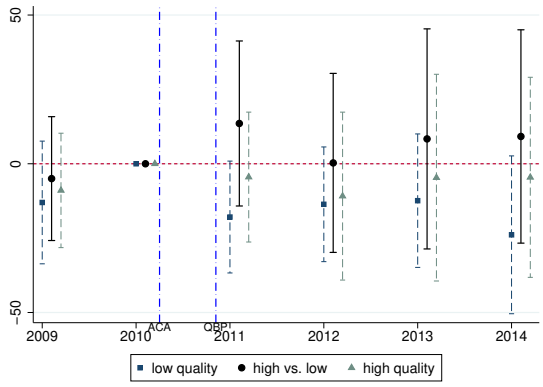
(f) Drug deductible, event study



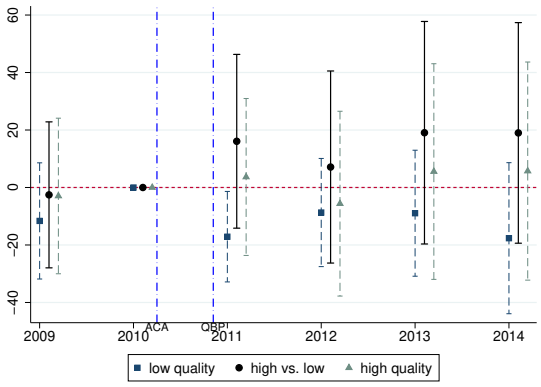
Notes: The figure plots the raw trends of premiums and drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. Different from the main analysis, we aggregate plan prices to the contract-county level using the median plan price. We restrict locations to the lower and upper 15% tails of county risk scores in the contract's service area. The raw trends plot the price levels across the 15% risk tails within an average low-rated contract (dotted lines) and an average high-rated contract (solid lines). Corresponding event study estimates in the right panels show the within-contract differences over continuous risk scores. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

Figure F3: Effect on premiums and drug deductibles, within-contract differences, event study, deviation to mean

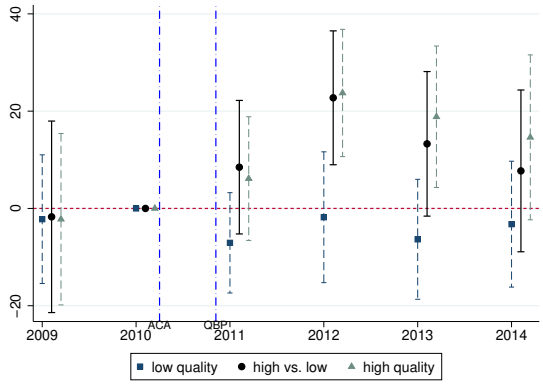
(a) Part C premium



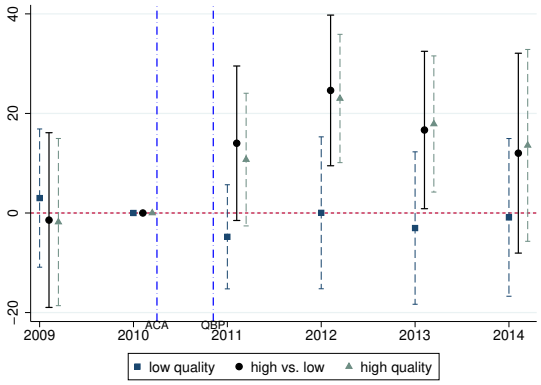
(b) Part C premium, 15% tails



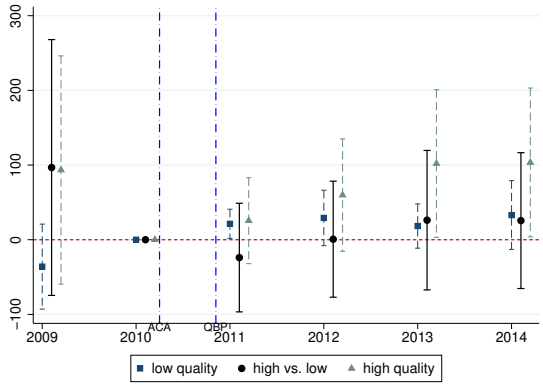
(c) Part D premium



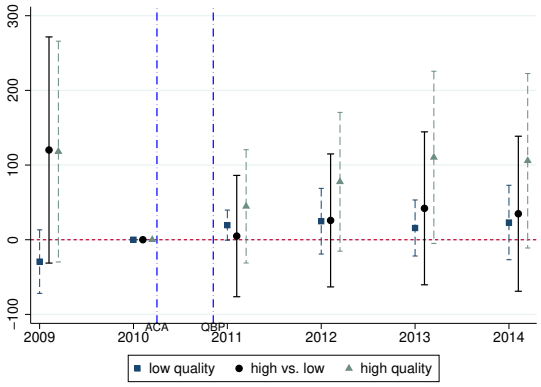
(d) Part D premium, 15% tails



(e) Drug deductible



(f) Drug deductible, 15% tails



Notes: The figure plots the event study estimates of the within-contract differences over county risk scores. We focus on Part C premiums in panel (a)-(b), Part D premiums in panel (c)-(d), and drug deductibles in panel (e)-(f). County differences in risk scores are measured as the deviation to the mean county risk in the service area, as opposed to the deviation-to-median measure in the main analysis. The right panels restrict within-contract locations to the lower and upper 15% of county risk scores in the contract's service area. Plotted 95% confidence intervals are based on robust two-way clustered standard errors at the contract and county levels.

G Enrollment Responses to Premium Changes

We provide additional evidence on the selection mechanism focusing on the enrollment responses to premiums. We first infer marginal risk types from the change in contract risk scores in Appendix Figure G1. New enrollees in high-rated contracts have an average risk score of 0.90, lower than the risk score of average enrollees in the contract (0.96 cf Table 1). Moreover, new enrollees are significantly healthier in high-rated contracts than in low-rated contracts (p-value=0.02). Next, we exploit the premium differences across counties to examine the enrollment responses to selection. Specifically, we estimate the following equation for enrollment

$$s_{clt} = \eta_0 \cdot p_{clt} + \eta \cdot X_{lt} + \rho_{cl} + \theta_t + \omega_{clt}, \tag{G1}$$

where s_{clt} is the enrollment share in county l relative to the total enrollment in contract c and year t. We hence use equation G1 to examine enrollment responses within contracts across counties. To focus on marginal enrollees, we instrument premiums p_{clt} based on the selection mechanism estimated in our difference-in-differences strategy (equation 5). Specifically, we estimate the following first stage for premiums p_{clt}

$$p_{clt} = \beta_0 \cdot risk_{cl} \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt}, \tag{G2}$$

where $risk_{cl}$ is the deviation from the median county risk score in contract c, and $post_t$ indicates year 2011 and after. The instrument $risk_{cl} \cdot post_t$ exploits pre-existing differences within contracts across county risk scores as predictors of premiums after the payment reform. Under the assumption that premiums would have followed parallel trends across county risk scores absent the reform, the instrument isolates premium differences generated by the selection incentive in the payment model. We regress enrollment s_{clt} on the

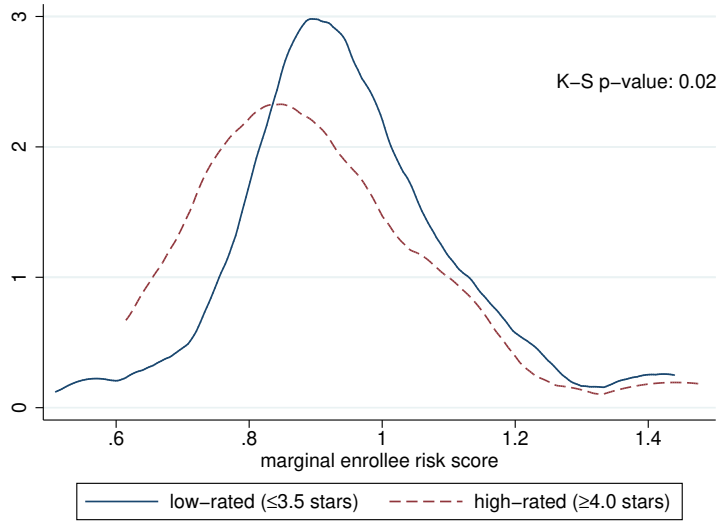
predicted premium $\hat{p}_{clt}$ in the second stage

$$s_{clt} = \gamma_0 \cdot \hat{p}_{clt} + \gamma \cdot X_{lt} + \psi_{cl} + \phi_t + \varepsilon_{clt}, \tag{G3}$$

where γ_0 estimates the enrollment responses of marginal enrollees.

We show estimates of $\hat{\gamma}_0$ for the Part D premium and the total premium of high-rated insurance in column 1-2 of Appendix Table G1. Additionally, we examine enrollment responses within county l across contracts in column 3-4. Across specifications, the implied demand elasticity ranges from -2 to -3 for Part D premiums, and from -3 to -4 for total premiums. These elasticities are comparable to existing estimates in Part D (e.g., Lucarelli *et al.* 2012, Decarolis *et al.* 2020, Starc and Town 2020), with plan-specific elasticities ranging from -2 to -6. Thus, enrollment responses imply similar demand elasticities as those in the literature, lending support to the selection mechanism through premiums.

Figure G1: New enrollee risk scores, kernel density



Notes: The figure plots the kernel density of new enrollee risk scores in low- and high-rated contracts. New enrollee risk scores are inferred from the changes in contract risk scores and enrollments after the payment reform. We test for the equality of distributions applying the Kolmogorov–Smirnov (K-S) test, showing the p-value on the top-right corner.

Table G1: Enrollment Responses to Premiums in High-Rated Contracts

	(I)	(II)	(III)	(IV)
Part D Premium	-0.003**		-0.10*	
	(0.001)		(0.053)	
Total Premium		-0.002**		-0.053*
		(0.001)		(0.028)
Enrollment	within-contract shares		county log enrollment	
Elasticity	-1.73	-2.85	-2.91	-4.25
First-stage F-stat	9.20	6.82	6.85	6.46
y-mean	0.062	0.062	8.32	8.32
N	5,611	5,611	5,660	5,660

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

Notes: The table estimates enrollment responses to premiums in high-rated contracts. The outcome variable is enrollment in contract c in county l and year t. We instrument premiums building on our difference-in-differences specification in equation 5. Specifically, we estimate the following first-stage equation for premium p_{clt}

$$p_{clt} = \beta_0 \cdot risk_{cl} \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt},$$

where $risk_{cl}$ in the deviation from the median county risk score in contract c, and $post_t$ indicates year 2011 and after. The instrument $risk_{cl} \cdot post_t$ isolates pre-existing differences in county risk scores as a predictor of premiums after the payment reform. We then use the predicted premium $\hat{p}_{clt}$ to estimate the enrollment responses in the second stage

$$s_{clt} = \gamma_0 \cdot \hat{p}_{clt} + \gamma \cdot X_{lt} + \psi_{cl} + \phi_t + \varepsilon_{clt},$$

where y_{clt} is the enrollment in county l relative to the total enrollment in contract c in year t. We hence estimate enrollment responses within contracts across county premiums in column 1-2. Additionally, we examine enrollment responses within county l across contracts in column 3-4. We compute the implied demand elasticity for each specification in the table. Robust Two-way clustered standard errors at the contract and county levels in parenthesis.

H Distributional Impacts: Additional Evidence

We provide additional evidence on the distribution of high-rated insurance across county risk scores. We first examine how the premium differences within high-rated contracts may impact the market share of insurance across counties in the upper and lower 15% of risk scores. We focus on the risk tails because premium differences within contracts would tend to decrease (increase) premiums in the healthiest (riskiest) counties. In the intermediate range, premiums can either increase or decrease depending on the ranking of the county's risk score within contracts and the distribution of contracts across counties.

Table H1 estimates the impacts on premiums using the specification in equation 10. We estimate effects separately for the lower and the upper risk tail because high-selection contracts are highly concentrated in the lower risk tail and account for only 10% of the high-rated insurance in the upper risk tail. Consistent with stronger premium differences within high-selection contracts, premiums of high-rated insurance decreased significantly with county risk scores in the lower risk tail (column 2), but not in the upper risk tail (column 5). Across risk tails (Table H2), high-selection contracts increased premiums by \$7.6 for a ten percentage point increase in the county risk score (column 2), or by 7.7% above the mean.

Table H2 examines the market shares of high-selection contracts in column 4-6. The increase in premiums reduced the market share of high-selection contracts by 9.3 percentage points for a ten percentage point increase in risk score, or by 7.7 percentage points differentially compared to the low-rated contracts. These estimates suggest that moving from the 15th to the 85th percentile of county risk score would increase premiums of high-selection contracts by \$11.98 and reduce market shares by 14.65 percentage points after the payment reform.[37] Figure H1 shows the event study.

[37] Specifically, risk score increases by $1.028 - 0.87046 = 0.15754$ from the 15th to the 85th percentile, implying larger premiums by $0.15754 \cdot \$76.02 = \$11.98$ and lower market shares by $0.15754 \cdot 0.93 = 14.65\%$ for high-selection contracts.

Table H1: Premium differences across county risk scores, lower and upper 15% risk tails

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · Post	63.81	121.40***	119.98***	-25.18**	32.15	-60.07
	(39.89)	(35.37)	(35.99)	(11.48)	(46.35)	(105.61)
Counties		<15% risk			>85% risk	
Contracts	low	high	high-select	low	high	high-select
y mean	53.01	98.57	102.26	41.50	73.50	66.63
R^2	0.83	0.82	0.79	0.81	0.76	0.91
N	1,233	1,035	959	3,909	1,019	96

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

Notes: The table estimates the premium differences across county risk scores separately in the lower 15% risk tail (column 1-3) and the upper 15% (column 4-6). Two-way clustered standard errors at the contract and county levels in parenthesis.

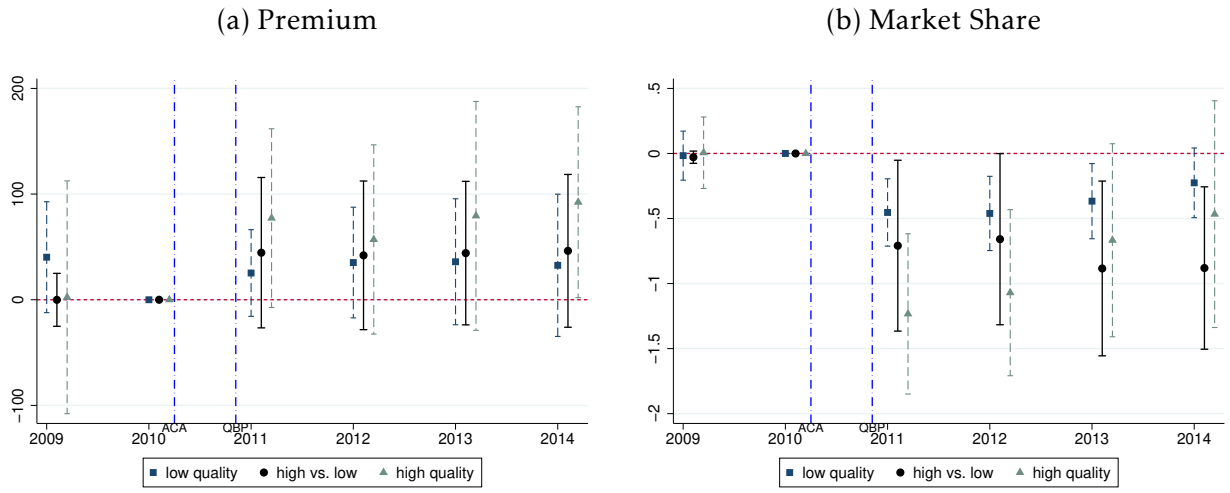
Table H2: Effect of bonus payments on premiums and market shares across the 15% risk tails

	(I)	(II)	(III)	(IV)	(V)	(VI)
	Premium			Market Share		
Risk · High · Post			45.17			-0.77**
			(37.80)			(0.32)
Risk · Post	11.91	76.02**	14.36	-0.39***	-0.93**	-0.40***
	(17.62)	(34.27)	(18.02)	(0.11)	(0.32)	(0.11)
High · Post			-37.95			0.80**
			(38.58)			(0.33)
Risk · High			-77.85			0.26
			(49.23)			(0.57)
Counties		15% tails			15% tails	
Contracts	low	high-select	(2) vs. (1)	low	high-select	(5) vs. (4)
y mean	44.26	99.02	53.24	0.27	0.52	0.31
R^2	0.82	0.81	0.84	0.70	0.64	0.69
N	5,143	1,055	6,251	5,143	1,055	6,251

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

Notes: The table estimates the differences in premiums (column 1-3) and market shares (column 4-6) over county risk scores across the 15% risk tails. We restrict high-rated contracts to high-selection contracts below the median service area risk (0.975) of high-rated insurance. Two-way clustered standard errors at the contract and county levels in parenthesis.

Figure H1: Effects on premiums and market shares across the 15% risk tails



Notes: The figure plots the event study estimates on premiums (panel a) and market shares (panel b) across county risk scores in the 15% risk tails. We restrict high-rated contracts to high-selection contracts below the median service area risk (0.975) of high-rated insurance. 95% confidence intervals are plotted based on robust two-way clustered standard errors at the contract and county levels.

H.1 Robustness

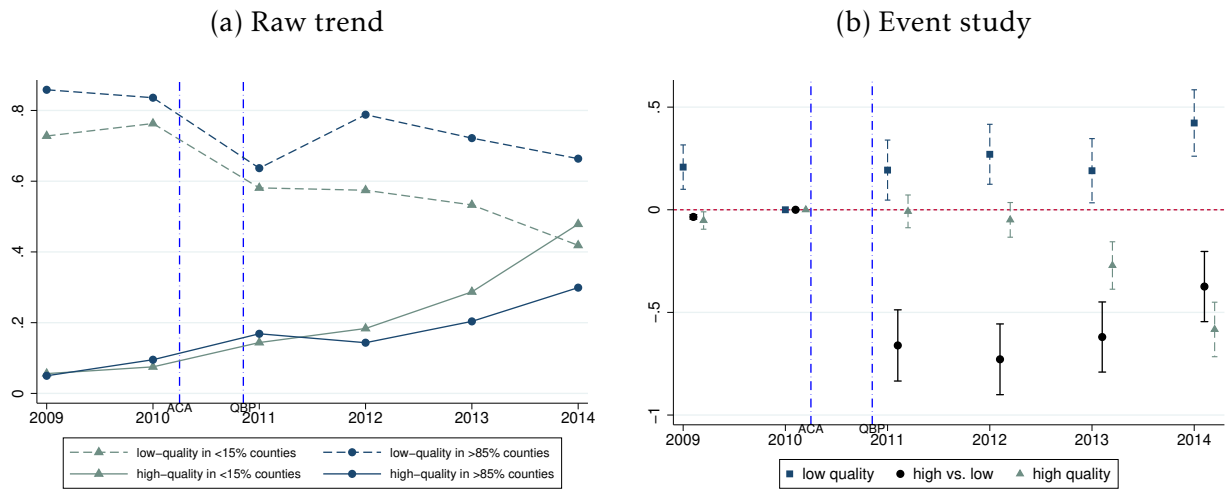
We show robustness of the results on high-rated market shares in two ways. First, since we distinguish contracts by the baseline quality rating, actual distribution of quality rating may differ from our estimates if ratings improved differentially across counties. We show that similar divergence in high-rated market shares occurs when the quality rating is based on the contemporaneous rating. We also extend our sample to include all Medicare Advantage contracts in the Landscape Files. In the full sample, we find similar divergence in high-rated market shares in the risk tails, driven by greater growth rate of high-rated insurance in the healthiest counties. These results support the finding that the payment reform worsened the regional disparity in the access to high-rated insurance in Medicare.

H.1.1 Contemporaneous Star Ratings

Figure H2 shows the market share changes for high- and low-rated insurance based on the contemporaneous rating. In the lower 15% risk tail (gray lines), the market share of

high-rated insurance increased and leveled with the market share of low-rated insurance over the sample period. In high-rated insurance (solid lines), market shares followed parallel trends in 2009-2010 but increased differentially in the lower risk tail after the payment reform. Table H3 estimates the changes in market shares across county risk scores in column 4-6.

Figure H2: Effects on market shares, contemporaneous star rating, 15% risk tails



Notes: The figure plots the raw trends of the market share changes in high- and low-rated insurance in panel (a) and the event study estimates in panel (b). High-rated insurance includes contracts rated 4.0 stars or above in the contemporaneous rating. We construct market shares for a balanced panel of rating-county-years and assume zero market shares for county-years with masked enrollment. We show market share changes in the lower and upper 15% of county risk scores. 95% confidence intervals are plotted based on robust standard errors clustered at the county level.

H.1.2 Full Sample of MA Contracts

For a comprehensive view of the quality rating distribution in the Medicare Advantage market, we construct market shares including all contracts listed in the Landscape Files. The full sample is different from the estimation sample in that Regional Preferred Provider Organization (PPO) plans, Part-C only plans, and contracts with missing quality ratings for payment purposes are retained in the full sample. High-rated insurance includes contracts rated 4.0 stars or above in the contemporaneous rating. We construct market

Table H3: Effects on market shares, contemporaneous quality rating

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-0.71***			-0.51***
			(0.11)			(0.12)
Risk · Post	0.28***	-0.21**	0.39***	0.17**	-0.10**	0.29***
	(0.069)	(0.047)	(0.070)	(0.079)	(0.052)	(0.079)
High · Post			1.03***			0.83***
			(0.11)			(0.12)
Risk · High			-0.29***			-0.35***
			(0.088)			(0.096)
Counties		all			15% tails	
Contracts	<4.0 stars	≥4.0 stars	all	<4.0 stars	≥4.0 stars	all
y mean	0.70	0.18	0.44	0.69	0.17	0.43
R^2	0.49	0.55	0.39	0.49	0.54	0.39
N	17,236	17,236	34,508	5,060	5,060	10,144

Notes: The table estimates the effect of bonus payments on the market shares of high- and low-rated insurance across county risk scores. We distinguish across quality ratings using the contemporaneous rating, and classify contracts with a 4.0 star rating and above as high rated. We then aggregate contract market shares to the rating level in a balanced panel of rating-county-years where counties with masked enrollment data in some but not all years receive zero market shares for missing enrollments. We show market share changes across the full sample of counties in the balanced panel in column 1-3, and restrict the sample to counties in the 15% risk tails in column 4-6. Robust standard errors clustered at the county level in parenthesis.

shares of high-rated insurance for county-years with at least one Medicare Advantage contract listed in the Landscape Files.

Table H4 estimates the market share changes in high-rated insurance across the 15% risk tails: a ten percentage point increase in the county risk score would reduce high-rated market share by 1.6 percentage points (column 3). This effect is driven by high-rated plans in the lower risk tail, whose market share increased substantially. A ten percentage point reduction in the county risk score would increase the high-rated market share by 8.4 percentage points (column 1). Moving from the lower to the upper 15% risk tail would decrease the high-rated market share by 3.7 percentage points (column 4) after the payment reform.

Panel (a) of Figure H3 compares the market share of high-rated insurance in the lower (gray line) and upper (blue line) risk tails. The widening gap across risk tails is driven by a larger growth rate of high-rated insurance in the healthiest 15% counties. By 2014, high-rated market share in the healthiest counties surpassed that in the riskiest counties by over 20 percentage points (panel b). Prior to the payment reform, market shares in both risk tails stayed on close and parallel trends. After a temporary drop in 2011 due to the revision of the star rating,[38] market shares diverged across risk tails and increased at a faster rate in the healthiest counties. In 2014, the market share of high-rated insurance was 46.5% in the healthiest 15% counties and 25.4% in the riskiest counties.

The maps in Figure H4 portray the differential growth rates across county risk tails. Counties with the lowest risk scores are concentrated in the North West, South West, and parts of the Mid West (panel a), where high-rated market shares increased by a median of 16% in 2009-2014 and increased even more in the healthiest 7.5% of counties (panel b). By contrast, high-rated market shares increased by a median of 5% in the riskiest counties in the South and the coastal areas, and decreased markedly particularly (up to -28%) in the riskier Southern counties.

H.2 Availability of High-Rated Insurance Across Counties

We next examine selective entry and plan offerings across counties as an additional mechanism of selection. For instance, high-rated insurers may choose to exist riskier counties or reduce plan offerings in these counties. Comparing across contracts, Appendix Table D6 indicates no differential changes in the service area risk score or plan offering between high- and low-rated insurance. Comparing across counties, Appendix Figure H5 illustrates the plan offering of high-rated insurance in the risk tails, plotting the share of

[38] The 2011 rating is the first rating computed from both Part C and Part D measures. The revised rating requires a larger number of measure ratings, some of which could not be computed for small-enrollment contracts based on historic data. The disruption does not affect high-rated contracts in the estimation sample (Figure H2), where market shares decreased similarly in riskier counties after the payment reform (column 5 of Appendix Table H3).

Table H4: Effects on high-rated market shares, contemporaneous star rating

	(I)	(II)	(III)	(IV)
Risk · Post	-0.84***	-0.073	-0.16***	-0.037***
	(0.27)	(0.12)	(0.048)	(0.013)
Risk	continuous FFS risk			>85%
Counties	<15%	>85%	15% tails	15% tails
y mean	0.19	0.13	0.16	0.16
R^2	0.50	0.65	0.54	0.54
N	2,410	2,639	5,049	5,049

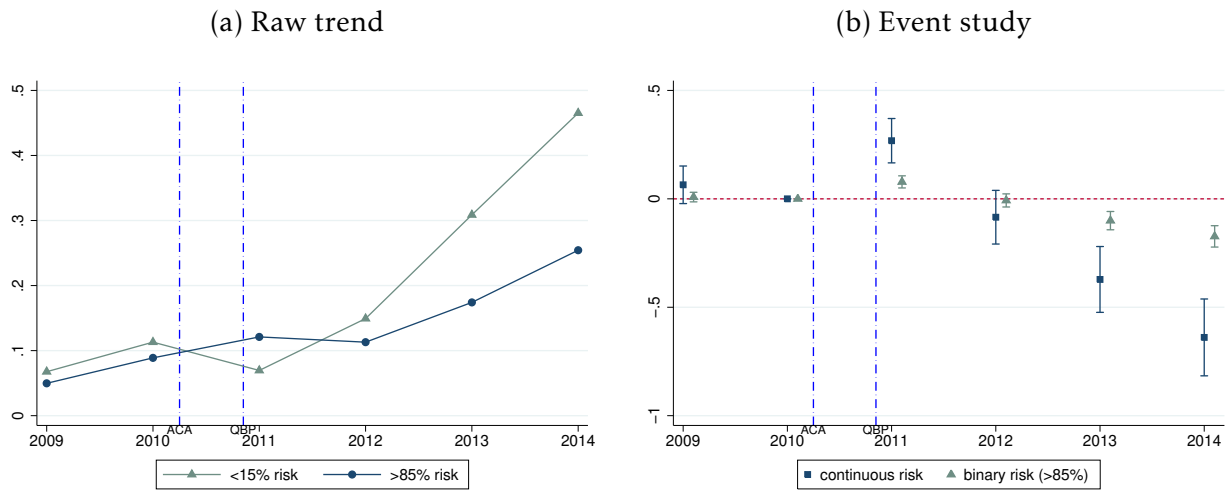
Notes: The table shows the effect of bonus payments on the market share of high-rated insurance across county risk scores. High-rated insurance includes contracts rated 4.0 stars or above in the contemporaneous rating. We construct market shares of high-rated insurance for county-years with at least one MA contract listed in the Landscape Files. We focus on counties in the lower 15% of county risk scores in column 1, in the upper 15% in column 2, and across the 15% risk tails in column 3. In column 4, the Risk variable is a binary indicator of the upper risk tail, with the estimate indicating the change in high-rated market share when risk score increased from the lower to the upper 15%. Robust standard errors clustered at the county level in parenthesis.

counties with at least one high-rated insurance plan in panel (a) and the share with at least two high-rated plans in panel (b). In both cases, plan offering increased on roughly parallel trends across risk tails after the payment reform. These patterns suggest that the availability of high-rated plans did not drive the market share changes across risk tails.

H.3 Market Share Gains and County Characteristics

We also ask whether the growth of high-rated insurance across space was associated with county characteristics in addition to risk scores. Based on our analysis of the selection mechanism in Section 4.2, the premium differences would impact market shares across

Figure H3: Effects on high-rated market shares, contemporaneous star rating, 15% tails



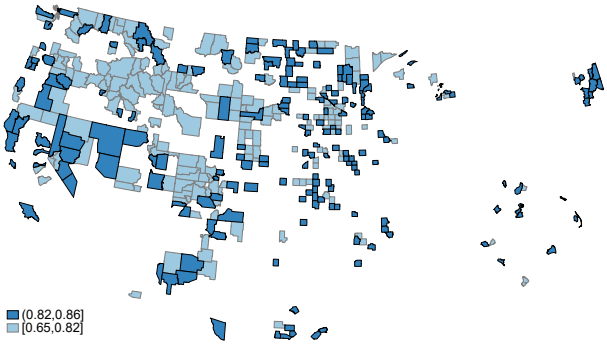
Notes: The figure plots the raw trends (panel a) and the event study estimates (panel b) of high-rated market shares across county risk scores in 2009-2014. High-rated insurance includes contracts rated 4.0 stars or above in the contemporaneous rating. We construct the market share of high-rated insurance for county-years with at least one MA contract listed in the Landscape Files. The raw trends in panel (a) show high-rated market shares in the lower and the upper 15% risk tails. The event study in panel (b) plots the yearly differences in market shares across continuous county risk scores on the left, followed by the discrete change in market shares for an increase in county risk scores from the lower to the upper 15%. 95% confidence intervals are plotted based on robust standard errors clustered at the county level.

county risk scores, but not across alternative characteristics such as income, healthcare spending, or indicators of care quality. We consider these alternative characteristics as contributors of the high-rated market share in Table H5. Specifically, we regress each characteristics on the FFS risk score, and relate the residual variation to the growth of high-rated market share using the Spearman rank correlation. The correlation coefficient shown in Table H5 indicates the extent to which higher growth counties also rank higher for a given characteristic. Across columns, counties with larger growth of high-rated insurance also have lower FFS risk scores and higher insurer concentration in the baseline, but are not associated with higher incomes, healthcare spending, or quality.[39] These results support our finding that the growth of high-rated insurance in the healthiest counties is driven by insurer selection of healthy enrollees in these counties.

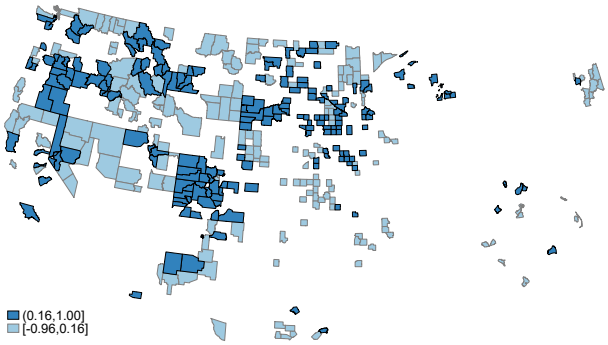
[39]In particular, the association with risk scores and insurer concentration increases in counties with the largest growth of high-rated insurance, but the association with alternative characteristics remains small and statistically insignificant.

Figure H4: Distribution of county risk scores and changes in the high-rated market share, 15% risk tails

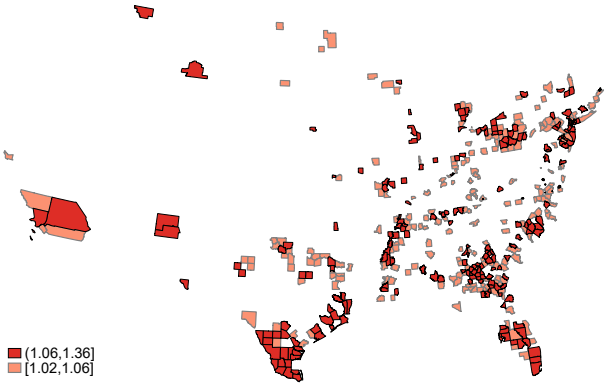
(a) County risk scores, healthiest 15% counties



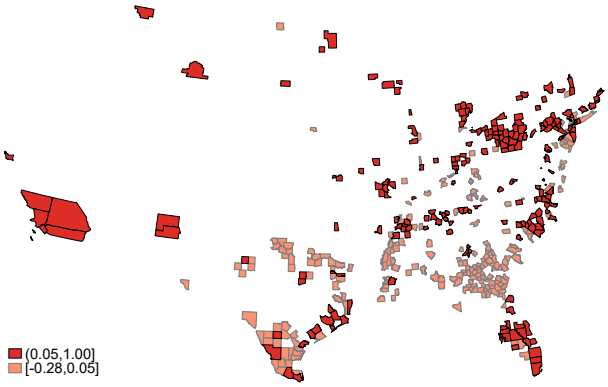
(b) Changes in the high-rated market share, healthiest 15%



(c) County risk scores, riskiest 15% counties



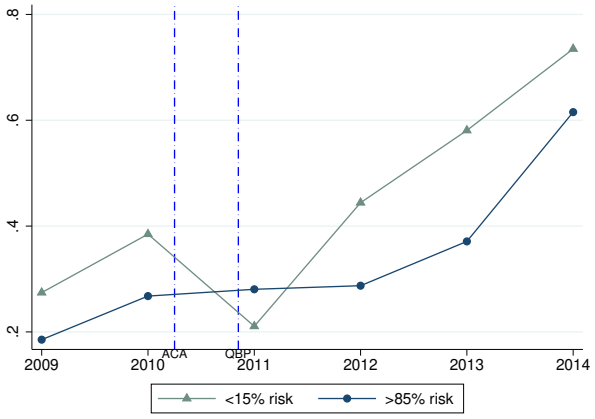
(d) Changes in the high-rated market share, riskiest 15%



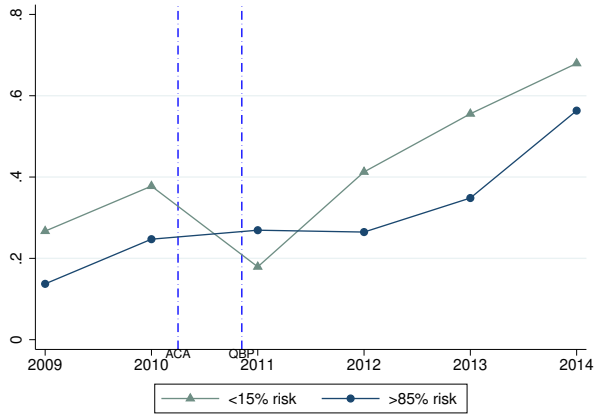
Notes: The figure plots the cross-county distribution of fee-for-service risk scores in the healthiest and the riskiest 15% counties in the left panels, and the changes in the market share of high-rated insurance in the risk tails in the right panels. We calculate the market share of high-rated insurance based on the contract's contemporaneous star rating (4.0 stars or above), and plot the increase in 2013-2014 market shares compared to the baseline in 2009-2010 across counties. Lighter colors indicate lower county risk scores in 2009-2010 in the left panels whereas darker colors indicate greater increases in high-rated market shares in the right panels. We choose the cut-point for colors based on the median of the outcome variable in each panel.

Figure H5: Effects on plan availability across counties, 15% risk tails

(a) At least one high-rated plan



(b) At least two high-rated plans



Notes: The figure plots the availability of high-rated insurance across the riskiest and the healthiest 15% counties in 2009-2014. In panel (a), we measure availability using the share of counties with at least one insurance plan rated 4.0 stars or above in the contemporaneous rating. In panel (b), we use the share of counties with at least two insurance plans rated 4.0 stars or above in the contemporaneous rating. We plot separate trends for the healthiest and the riskiest 15% counties in each panel.

Table H5: Growth of high-rated market shares and county characteristics at baseline, Spearman rank correlations

	(I)	(II)	(III)	(IV)
FFS risk score	-0.23	-0.30	-0.35	-0.45
Income	0.12	0.21	0.11	0.08
FFS spending	-0.06	-0.01	0.14	0.20
Re-admission	-0.11	-0.15	-0.06	-0.18
HHI	0.11	0.17	0.34	0.51
ΔHigh-Rated Share	>0	>0.30	>0.50	>0.75
N	1,760	752	377	161

Notes: The table shows the Spearman rank correlations between the increase in high-rated market shares in 2009-2014 and county characteristics in the baseline (2009-2010). We partial out the variation of FFS risk scores from the non-risk characteristics, and relate the residual variations to the growth of high-rated insurance across counties using the Spearman rank correlation. Across columns, we examine characteristics in counties with larger increases of high-rated market shares.

References in the Online Appendix

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.

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.

Curto, V., Einav, L., Levin, J. and Bhattacharya, J. (2019). Can health insurance competition work? Evidence from Medicare Advantage, Mimeo, Harvard University, Stanford University, Stanford University, Stanford University.

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., Polyakova, M. and Ryan, S. P. (2020). Subsidy design in privately provided social insurance: Lessons from medicare part d. *Journal of Political Economy*, **128** (5), 1712–1752.

Fang, H., Keane, M. P. and Silverman, D. (2008). Sources of advantageous selection: Evidence from the medigap insurance market. *Journal of political Economy*, **116** (2), 303–350.

Finkelstein, A., Gentzkow, M., Hull, P. and Williams, H. (2017). Adjusting risk adjustment–accounting for variation in diagnostic intensity. *The New England Journal of Medicine*, **376** (7), 608.

Han, T. and Lavetti, K. (2017). Does part d abet advantageous selection in medicare advantage? *Journal of health economics*, **56**, 368–382.

Lucarelli, C., Prince, J. and Simon, K. (2012). The welfare impact of reducing choice in medicare part d: A comparison of two regulation strategies. *International Economic Review*, **53** (4), 1155–1177.

Newhouse, J. P., 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.

Olives, C., Myerson, R., Mokdad, A. H., Murray, C. J. and Lim, S. S. (2013). Prevalence, awareness, treatment, and control of hypertension in United States counties, 2001–2009. *PloS One*, **8** (4), e60308.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. and Sobek, M. (2019). IPUMS USA: Version 9.0 [dataset].

Sorbero, M. E., Paddock, S. M., Damberg, C. L., Haas, A., Kommareddi, M., Tolpadi, A., Mathews, M. and Elliott, M. N. (2018). Adjusting medicare advantage star ratings for socioeconomic status and disability. *Am J Manag Care*, **24** (9), e285–e291.

Starc, A. and Town, R. J. (2020). Externalities and benefit design in health insurance. *The Review of Economic Studies*, **87** (6), 2827–2858.