INFER
INTERNATIONAL NETWORK FOR ECONOMIC RESEARCH

Working Paper 2020.03

Performance Pay in Insurance Markets: Evidence from Medicare

by

Michele Fioretti

(Department of Economics, Sciences Po)

Hongming Wang

(Center for Global Economic Systems, Hitotsubashi University)

Performance Pay in Insurance Markets: Evidence from Medicare *

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

August 31, 2020

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 U.S. Medicare Advantage market, we find that insurers with higher quality ratings 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 the health of enrollees. Selection inflated the bonus payments and shifted the supply of high-rated insurance to the healthiest counties, hurting the healthcare access of sicker patients 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 and at seminars at the University of Southern California, Sciences Po, and Bar-Ilan 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 actual improvements in quality 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 such as 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 worsen 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 the quality rating, payments, and enrollees.

This paper examines how insurers respond to pay-for-performance by exploiting the introduction of quality bonus payments 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, bonus payments depend on insurance quality through a quality rating that was already available to prospective enrollees before the payment reform. Since the reform shifted insurer payments without affecting consumers' knowledge of the quality rating, we exploit the reform to understand insurers' responses to pay-for-performance and their impact on consumers' access to insurance.

We find that insurers with high-quality ratings before the reform served less risky enrollees after the payment reform. These insurance contracts lowered premiums in healthier, low-risk counties and increased premiums in riskier ones to select healthier enrollees. Risk selection is profitable because the quality rating relies heavily on health outcome measures, but fails to adjust these measures for enrollees' health conditions. In

[1] Medicare provides near-universal health insurance to Americans over the age of 65. The program costs 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 a private insurer in the Medicare Advantage program.

response, selecting insurers inflated the quality rating by avoiding enrollees with more complicated conditions. Due to the selection response, insurance benefits and the supply of high-rated insurance shifted to the healthiest counties, hurting in particular sicker enrollees 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 find that the distribution of risk scores shifted 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 decreased even more in high-rated contracts serving healthier counties before the payment reform – in these "high-selection" contracts, risk scores dropped by 4 percentage points. These effects suggest that the payment reform incentivized high-rated insurers to select enrollees based on risk. We next ask how insurers selected enrollees and why.

To address how selection happened, we examine the pricing strategy of insurers across counties. We find that premiums for prescription drug coverage 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, cost and quality of care 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 using 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 on 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, bonus payments increased by 16% to high-selection contracts.

The selection response has sizeable distributional impacts on enrollees. Since average premiums and enrollee benefits did not differ by the quality rating, premium differences to select healthier enrollees shifted insurance benefits from the sickest to the healthiest enrollees in high-rated insurance. The market share of high-rated contracts increased by more than 17% in the healthiest counties than in the riskiest counties. As insurance benefits and the supply of high-rated insurance shifted to the healthiest counties, the access to generous, high-rated insurance worsened particularly for the riskiest enrollees.

Several aspects of the quality rating contributed to the selection responses. The lack of risk adjustment on health outcomes implies that healthier enrollees are more profitable to insurers. This is because insurers are under-compensated for treating sicker enrollees with predictably worse outcomes. Risk-adjusting the health outcome measures – so that health improvements are relative to the predicted outcomes given risk types – can limit the selection incentive. Second, the biased outcome measures receive the largest weights in the quality rating, which magnifies the financial return to gaming these measures with selection. Since we do not find significant health improvements after adjusting for risk in the selecting contracts, down-weighting the health outcome measures can lower the selection incentives without harming the health of enrollees.

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 effort (Holmstrom and Milgrom 1991, Baker 1992). In relation to the empirical literature on healthcare, previous studies generally find small effects of pay-for-performance on providers (Rosenthal and Frank 2006, Mullen *et al.* 2010), with some evidence of patient selection (Shen 2003, Gupta 2017) and strategic reporting (Gravelle *et al.* 2010) in the case of outcome-based performance measures. We add to this literature by providing the first evidence on how insurers respond to pay-for-performance incentives and the distortionary effects of these responses on prices and consumer access to health insurance.

This paper also relates to the literature on the effects of risk adjustment in health insurance markets (e.g., Newhouse *et al.* 2015, Breyer *et al.* 2011). Ideally, risk adjustment makes different enrollee types equally profitable to insurers. In practice, selection may

still occur post-adjustment over the residual variation in the profitability created by the adjustment formula (Brown *et al.* 2014, Carey 2017, Lavetti and Simon 2018, Geruso and Layton 2018, Geruso *et al.* 2019). This paper suggests that the residual variation is a small price to pay relative to the systematic variation favoring healthier enrollees absent risk adjustments on quality. The endogenous responses to the regulatory formula, however, caution against the pitfalls of pure statistical models of risk adjustment in health insurance markets (Einav *et al.* 2016, Obermeyer *et al.* 2019).

This paper contributes to the literature on the regional disparities in health spending (Skinner, 2011), prices (Cooper *et al.*, 2018), and health outcomes (Dickman *et al.*, 2017) in the U.S. We complement the vast and descriptive evidence by highlighting one particular mechanism affecting disparities, namely the gaming of public subsidies by insurers. This finding suggests the role of supply-side regulations in shaping the regional disparities of healthcare access, and the role for policy to mitigate the inequalities.[2] In Medicare, the design of the payment model not only affects the pass-through of subsidies to enrollees (Duggan *et al.*, 2016, Cabral *et al.*, 2018, Curto *et al.*, 2019) and prices (Decarolis, 2015, Decarolis *et al.*, 2015, Mahoney and Weyl, 2017), but has large distributional consequences across enrollee types hurting in particular the sicker population in the riskiest counties.

The remainder of the paper is organized as follows. Section 2 introduces the background of private Medicare insurance and the Quality Bonus Payment demonstration. Section 3 motivates our empirical analysis with a stylized model of insurer responses to quality bonus payments. Section 4 describes the data. Following the predictions from the model, we examine the effects of bonus payments on risk scores in Section 5 and the pricing responses across counties in Section 6. We inspect the rating design as the source of the selection incentive in Section 7, and address the cost of selection to Medicare and the distributional impacts across space in Section 8. Section 9 concludes.

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 the traditional Medicare, also known as the fee-for-service Medicare, and private Medicare insurance from the Medicare Advantage (MA) market. MA plans cover at least the same medical services as the traditional Medicare and typically provide additional benefits to enrollees. For example, most MA plans also

[2]The effects of supply-side factors on health disparities can also be learned from mover designs that separate such effects from patient demand characteristics (e.g., Finkelstein *et al.*, 2016, 2019, Deryugina and Molitor, 2018).

offer prescription drug coverage. In these plans, consumers will be charged a separate premium for the traditional Medicare benefits (the Part C premium) and a premium for the prescription drug coverage (the Part D premium). Private insurers are regulated by the Centers for Medicare and Medicaid Services (CMS). Insurers offer insurance contracts including a menu of subsidiary insurance plans, in counties within a contract's service area. Premiums and benefit design can vary across plans, but cannot vary by enrollees in the same plan and county.

2.1 Quality Rating in Medicare Advantage

In 2009, CMS introduced the "star rating," which provided consumers with a quality measure of Part C and Part D contracts. The star rating ranks insurance contracts on a scale from 1 to 5 stars, increasing by half-star increments. All subsidiary plans receive the same star ratings as the contract. The star rating observed by consumers is a weighted average of a large number of measure-level star ratings focusing on specific aspects of insurance quality. Measure-level ratings are assigned to contracts based on their percentile rank among all contracts in that measure. All measures received equal weights in 2009-2011. Since 2012, measures of enrollee health outcomes and chronic conditions (the outcome measures) receive 3.0 weights in the overall rating. Measures of customer services (the access measures) receive 1.5 weights. Measures of managed care processes such as preventive care (the process measures) receive 1.0 weights.[3] Upon weighting, nearly half of the overall star rating is determined by the health outcome measures.

Health Outcome Measures. Most of the health outcome measures concern the management of chronic conditions such as diabetes and hypertension. The outcome data come from the Healthcare Effectiveness Data and Information Set (HEDIS), which contains the medical records of enrollees with chronic conditions. According to the HEDIS, a chronic condition is "managed" if its related medical test meets a pre-specified threshold.[4]

The HEDIS outcomes are *not* adjusted for the risk types of enrollees. Specifically, the medical thresholds used to measure chronic conditions are pre-determined; if sicker patients have worse health outcomes, the health outcome measures puts contracts enrolling sicker patients at disadvantage. Thus, health outcome measures do not distinguish

[3] We show the full list of quality measures for MA-PD plans in 2013 in Appendix Table A1 and A2, where we list the weight, the underlying data source, and the period over which data are collected for each measure.

[4] For example, hemoglobin A1c and low-density lipoprotein cholesterol (LDL-C) test results are recorded to monitor the care effectiveness for diabetes patients. 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 available for download with the rating data. The data are accessible at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.

an insurer's quality at disease management from differences in the case-mix of health conditions. A similar problem exists with the Part D outcomes, such as drug safety and adherence measures for patients with diabetes or hypertension, which received 3.0 weights in the overall rating. However, these measures were introduced at different points in time.[5] Therefore, this paper mainly focuses on selection in the HEDIS outcomes.[6]

Other Measures. The HEDIS data also contribute to a large number of process measures in the quality rating. The process measures evaluate care effectiveness through the take-up of screening, functional assessment, and medication reviews.[7] Access measures are drawn from the Consumer Assessment of Healthcare Providers and Systems (CAHPS), where enrollees rate the health plan in terms of getting needed care, complaint resolution, and overall customer service. Importantly, these responses are adjusted for the age, education, and general health status of enrollees.[8]

Star Ratings Refer To Past Years. All quality measures are based on historic data summarizing past performances of the contract. Health outcome measures, the most weighted in the star rating scheme, are delayed by two years between measurement and the entry in the quality rating. For instance, HEDIS measures of diabetes and hypertension control in the 2011 rating are measured from enrollees serviced in 2009. HEDIS process measures are similarly delayed by two years. Access measures of customer satisfaction are the most up-to-date, with the 2011 ratings derived from CAHPS records from early 2010.[9] However, since the 2011 rating is released in the Fall of 2010, the rating does not directly measure the quality of insurance in the 2011 enrollment year, but is heavily influenced by enrollee health outcomes from two years prior in 2009.

2.2 Quality Bonus Payments

Insurance plans rely critically on subsidies from CMS to operate. Curto *et al.* (2019) estimate that subsidies account for over 80% of the cost of covering an enrollee, with the

[5] Specifically, CMS introduced three drug adherence measures in 2012. These measures calculate the share of diabetes, hypertension, and high-cholesterol enrollees taking the prescription as directed. Since 2010, two drug safety measures calculate the share of high-risk patients (e.g., diabetes and hypertension comorbidity) who are prescribed safe medication appropriate for the complication. Similar measures did not exist in the 2009 rating.

[6] In Section 8.1, we detect similar but smaller selection biases for the drug outcomes relative to the HEDIS outcomes (Appendix Table A32).

[7] For diabetes care, for example, contracts are ranked based on the percent of patients who had an eye exam or a kidney function test in the enrollment year.

[8] 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."

[9] The measurement period of all quality measures for 2013 is listed in Appendix Tables A1 and A2.

remaining 20% charged to enrollees.[10] CMS subsidies are competitive. Payments to MA plans are determined by comparing the plan's asking price (bid) with its benchmark set by CMS. The bid (denoted b) reflects the projected cost of an average enrollee in the plan plus an administrative load. For plans bidding below the benchmark (denoted B), the payment equals the plan's bid plus a rebate. The rebate is passed on to enrollees as premium discounts or additional benefits. The payment from CMS is capped at the benchmark, so a plan charges enrollees an extra premium if its bid is greater than its benchmark.

Rewarding High Star Ratings. In an effort to promote value-based payments in healthcare, the Affordable Care Act (ACA) – signed into law in March 2010 – varied benchmarks with the quality rating of contracts. In the ACA model, contracts rated 4.0 stars and above are eligible for a 5% bonus on the benchmarks. Contracts rated below 4.0 stars are not eligible for benchmark bonuses. Formally, we can summarize the ACA model as follows

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

where $\theta^{star} > 1$ adjusts benchmark B according to the star rating. In practice, star ratings in year $t-1$ are used to calculate bonuses for year t.[11] Insurers bidding below the adjusted benchmark still receive a rebate to be passed along to consumers. Rebates are determined by $\gamma^{star} \cdot (\theta^{star} \cdot B - b)$, where γ^{star} increases past the 4.0 quality threshold. Given γ^{star}, rebates depend on insurers' responses to benchmark bonuses through the bid b. Before 2012 when bonus payments were introduced, $\theta^{star} = 1$ and $\gamma^{star} = 75\%$ for all contracts.

Timing. Because of the lag in the measurement of quality, improved insurance star ratings will be rewarded by future bonus payments. Health outcome measures, the most weighted in the star rating, are derived from the health of enrollees two years prior. In addition, bonuses for year t are calculated from $t-1$ star ratings. Together, the payment rules imply a three-year lag between the selection of enrollees in year t and the effect of selection on payments in year $t+3$. Panel (a) of Figure 1 illustrates this for enrollment year 2012, where bonus payments are computed from 2011 star ratings which are in turn derived from health outcomes in 2009.

Quality Bonus Payment Demonstration. The ACA model was not immediately imple-

[10] 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).

[11] This is because year-t ratings are not yet released in June of year $t-1$, when insurers must submit bids for plan offerings in year t. We showcase the timeline of the bidding and enrollment process as well as key policy changes in 2009-2012 in Appendix Figure 1a.

mented in the MA market. Between 2012 and 2014, a different set of payment rules determined the bonus payments to insurers. These rules were introduced in the Quality Bonus Payment Demonstration (QBP), announced by the CMS on November 10th, 2010. Compared to the ACA model, QBP applied less generous bonuses to higher-rated contracts above 4.0 stars, but increased bonuses to these contracts over time as the QBP model blended into the ACA model. After the demonstration ended in 2014, payments fully transitioned to the ACA model in 2015. We summarize θ^{star} and γ^{star} in 2009-2014 and in the first year of ACA (2015) in Table 1.

Despite rich variation in bonuses and rebates within a year, QBP mainly rewarded the past performances of contracts. As illustrated in panel (a) of Figure 1, bonus payments in 2012 are affected by the health of enrollees in year 2009. The three-year delay effectively links enrollees serviced in 2012 (and their outcomes) with payments in 2015, when the implementation of the ACA results in benchmark bonuses only for contracts rated 4.0 stars and above. Specifically, the discrete change at 4.0 stars increases benchmarks by 5% for contracts above the cut-off, and generates incentives to score and remain above the cut-off potentially with the selection of healthier enrollees.[12]

To understand the magnitude of benchmark bonuses introduced by QBP and ACA, we plot the predicted benchmark increases for contracts with different 2009-2010 star ratings in panel (b) of Figure 1.[13] Since 2012, benchmarks increased by less than 1% below the 4.0 star cut-off. For higher-rated contracts above the cut-off, benchmarks increased modestly in 2012-2013 under QBP. By 2014, when QBP benchmark bonuses aligned with the ACA rates for higher-rated contracts, benchmarks increased by 4.1% above the cut-off,[14] implying a $33 increase in payments per enrollee-month compared to the 2009-2010 levels.

3 Conceptual Framework

To gain insight into the selection responses to the quality rating, we build a simple two-county model, where the insurer uses premium differences across counties to obtain better ratings and payments. We use the model predictions to guide our empirical analysis.

[12] Appendix Table A3 illustrates the ACA policy variation in bonus rates linking year t star ratings with year $t+3$ payment models. Benchmark bonuses increased discretely from 0% to 5% above the 4.0 star cut-off.

[13] We predict quality-adjusted benchmarks after the payment reform based on the maximum Part C rating in 2009-2010. We apply the corresponding bonus rates to the raw county benchmarks, and use the average benchmark across counties as the predicted benchmark. In the prediction, we restrict counties to those already covered by the insurance contract prior to the payment reform.

[14] The benchmark increase did not exactly match the 5% bonus rate because raw county benchmarks were generally lower since 2012. Benchmark bonuses more than offset the base rate cut for lower-rated contracts, and substantially increased the benchmarks of higher-rated contracts with 4.0-star ratings or above.

Setting. An insurer sells Medicare insurance in two counties. The insurer's revenue depends on the premiums in the two counties (p_1 and p_2) and on the county benchmark B. Under pay-for-performance, benchmark B increases with the insurer's quality rating q. The demand for the contract in county l is given by $s_l = s_l(p_l)$, and the average risk score of contract enrollees is given by $r(p_1, p_2)$.[15]

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 $r(p_1, p_2)$. 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.[16] 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.

The insurer's problem is to maximizes total profits $\sum_{l=1}^{2}(p_l + B - c) \cdot s_l$ by choosing p_1 and p_2.[17] From the first order conditions, the optimal premium in county l solves

$$p_l = c - B + \left(1 + \frac{d\,B}{d\,q} \cdot \frac{\partial q}{\partial r} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_l + s_{-l}}{s_l}\right) \cdot |\varepsilon_l|^{-1}, \tag{2}$$

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

Selection through Prices. Before the payment reform, the optimal premium equals marginal cost plus a mark-up, which is inverse to demand semi-elasticity. After the reform, equation 2 shows that premium also responds to the selection incentives due to a biased quality rating through the term

$$\Delta p_l := \frac{d\,B}{d\,q} \cdot \frac{\partial q}{\partial r} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_l + s_{-l}}{s_l} \cdot |\epsilon_l|^{-1} \tag{3}$$

The selection term is switched on when $\frac{d\,q}{d\,r} \neq 0$. 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 $\frac{d\,q}{d\,r} < 0$, premium will be lower in county one relative to county two if, other things

[15] 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.

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

[17] 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.

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

$$\Delta r = \Delta p_l \frac{\partial r}{\partial p_l} + \Delta p_{-l} \frac{\partial r}{\partial p_{-l}} < 0. \tag{4}$$

Moreover, 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.[18] Specifically,

$$\Delta p_1 - \Delta p_{-l} \propto -\frac{d\,B}{d\,q} \cdot \frac{\partial q}{\partial r} \cdot \left(\Gamma_l^{FFS} - \Gamma_{-l}^{FFS}\right), \tag{5}$$

which states that other things equal, counties with healthier FFS enrollees are more likely to see lower premiums after the reform, relative to counties with riskier FFS enrollees.[19] This is because the risk score is determined jointly from the risk types of enrollees in both counties. For a small shift in the enrollment share across counties, the *level* differences in Γ_l^{FFS} are informative of the risk composition effect on r. Counties with lower Γ_l^{FFS} can more effectively lower the risk score r, and premiums decrease more in these counties.

Empirical Strategies. Equations 4 and 5 predict that contracts eligible for benchmark bonuses may respond by selecting healthier enrollees across markets. In the ACA model, contracts above 4.0 stars are eligible for benchmark bonuses (see Appendix Table A3). To examine the potential selection responses among these contracts, we compare the changes in risk scores above and below 4.0 stars in a difference-in-differences design.[20] To examine the pricing responses predicted in equation 5, we expand our difference-in-differences analysis to focus on premium differences across counties as the selection mechanism in Section 6.

4 Data

Our data come from the administrative registry of all MA-PD plans offered over 2009-2014 (the "Landscape File"). The data contain information on plan characteristics such as pre-

[18] 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.

[19] Omitted derivations are in Appendix C.

[20] We also explore heterogeneous responses among contracts closer to the 4.0 star cut-off, where the selection incentive may be stronger due to the loss of bonus payments below the cut-off.

miums and drug deductibles across 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 construction are available in Appendix D.

We merge this data with the Payment File containing plan payments and plan risk scores to understand how differences in product design and premiums affect the risk types enrolled in the plan. 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 2 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, MA-PD contracts offer 3.4 plans covering over 25 counties in the service area. Most contracts place bids below the benchmark, and enrollees receive an average of $81.04 in rebates per month. A large number of contracts charge zero premiums and zero drug deductibles.

5 Evidence of Risk Selection

In this section, we provide evidence on the selection responses to quality bonus payments in Medicare Advantage. Because the ACA model restricted bonus payments to contracts rated 4.0 stars and above, we estimate the potential selection responses among high-rated contracts (4.0 stars and above) using a difference-in-differences design. We detail the definition of high- and low-rated contracts and present empirical evidence of selection below.

High- and Low-Rated Contracts. We classify high- and low-rated contracts based on the maximum Part C rating in 2009-2010, our baseline period. We focus on Part C ratings because Part D ratings are calculated separately for MA-PD contracts in 2009-2010, whereas regulations of contract quality have traditionally focused on the Part C rating.[21] Specifically, high-rated contracts are those with at least one Part C rating of 4.0 stars or above in 2009-2010, and low-rated contracts are those rated between 3.0 stars and 3.5 stars in 2009-2010.[22] We summarize high- and low-rated contracts in column 3-6 of Table 2. In

[21] For instance, after the introduction of star ratings in 2009, MA contracts receiving Part C ratings below 3.0 stars for three consecutive years are suspended by the CMS. Part D ratings of the same contracts are not subject to the same regulation.

[22] We exclude contracts ever rated less than 3.0 stars in the baseline from our analysis, since the threat of suspension may generate selection incentives that are irrelevant to the bonus payments.

the empirical analysis, we allow for selection responses in 2011, our first post-reform year. This is because ACA was signed into law in March 2010, and insurers had until June 2010 to submit the bid and the benefit design for the 2011 enrollment year. We show evidence of the anticipatory effect examining yearly shifts in the risk score distribution below.

Shifts in the Risk Distribution. We first show evidence that the distribution of risk scores shifted to the lower percentiles in high-rated contracts. Figure 2 plots the kernel density of risk scores by contract rating, before (solid blue line) and after (dashed red line) the payment reform. We see a sharp change in the density of high-rated contracts, where the post density decreased in the middle of the distribution and increased in the lower percentiles.[23] We do not observe similar density shifts for low-rated contracts. Appendix Figure B1 further breaks out the density shifts by year for high- and low-rated contracts. Risk scores followed similar distribution in 2009-2010 for both groups, but shifted to the lower percentiles only for high-rated contracts since 2011, not for low-rated contracts.

Responses by Star Ratings. We further examine heterogeneous responses across baseline ratings in Figure 3. Specifically, we classify contracts by the maximum Part C rating in 2009-2010, and plot the density shifts for each rating from 3.0 stars to 4.5 stars.[24] The risk reduction in high-rated contracts are larger among marginal contracts rated no more than 4.0 stars in the baseline (panel c). For these contracts, risk scores shifted significantly from the middle to the lower percentiles. We observe a smaller and statistically insignificant shift for 4.5 star contracts. We do not find similar shifts for 3.5-star or 3.0-star contracts. Taken together, the selection responses are concentrated among high-rated contracts with 4.0 stars or above in the baseline, consistent with the prediction in Section 3.

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

$$y_{qt}(\kappa) = \beta(\kappa) \cdot high_q \cdot post_t + \alpha_q(\kappa) + \tau_t(\kappa) + \epsilon_{qt}(\kappa), \tag{6}$$

where $high$ indicates high-rated insurance and $post$ indicates the post-reform years (2011 and after). High-rated insurance includes contracts with at least one 4.0-star rating or above in 2009-2010. $\beta(\kappa)$ estimates the shift in the κ-th quantile of risk scores of high-rated contracts after the payment reform. We control for insurance rating, $\alpha_q(\kappa)$, and time fixed effects, $\tau_t(\kappa)$. We show estimates of equation 6 using the group quantile estimator

[23] Formally, we test for the density shift using the Kolmogorov–Smirnov (K-S) test, rejecting the null of equal distribution with a p-value less than 0.01%.

[24] We omit the 5.0 star contracts because very few contracts ever obtained 5.0 star ratings in the baseline.

of Chetverikov *et al.* (2016). The estimator first constructs quantiles $y_{qt}(\kappa)$ by rating-year groups, and then estimates the effects on $y_{qt}(\kappa)$ using standard OLS regressions. We show estimates by deciles of risk scores in panel (a) of Figure 4.[25] In panel (b), we provide complementary evidence from changes-in-changes estimates following Athey and Imbens (2006).[26] Both approaches reveal large and significant reductions in the 20% to 40% of the risk distribution. In these deciles, risk scores dropped by 4-8 percentage points in high-rated contracts, or 4%-9% below the baseline levels (Appendix Table A4). The effects on risk scores in the upper deciles are smaller and statistically insignificant.

High-Selection Contracts. The quantile analysis suggests that the effect of bonus payments on risk scores is highly heterogeneous, 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 in some, but not all, high-rated contracts. To examine the average and heterogeneous effects of the payment reform on high-rated contracts, we estimate the following specification

$$y_{ct} = \beta \cdot treat_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}, \tag{7}$$

where y_{ct} is the risk score of contract c in year t. We include contract (α_c) and year (τ_t) fixed effects, and use $treat$ to indicate different sub-groups of high-rated contracts. β estimates the effect of bonus payments on the risk scores of high-rated contracts indicated by $treat$.

Guided by the model in Section 3, we explore heterogeneous effects for contracts with different fee-for-service risk scores in the service area. Since MA contracts can more effectively select healthier enrollees in counties with lower FFS risk scores (equation 5), contracts more exposed to lower risk types in the service area are potentially better at risk selecting enrollees. Therefore, we calculate the service area risk as the average FFS risk score in the service area for high-rated contracts and consider heterogeneous effects by the median service area risk.

Panel (a) of Figure 5 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 decreased in high-rated contracts serving healthier locations. Panel (c) shows similar patterns across the lower and

[25] We block-bootstrap standard errors clustered by contracts from 500 replications, and plot the empirical 95% confidence intervals around the point estimates.

[26] Changes-in-changes relaxes the functional form assumption on standard difference-in-differences. It estimates quantile treatment effects when the distribution of unobservables does not vary within groups over time (Athey and Imbens, 2006). The latter assumption is not required in Chetverikov *et al.* (2016).

upper 25% of service area risks. We do not detect differential pre-trends. By contrast, the risk score declined in high-rated contracts in the healthiest locations.

Table 3 estimates the heterogeneous effects on high-rated contracts using equation 7. On average, risk scores declined 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 score did not differ meaningfully from low-rated contracts for 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-service risk scores. We next examine pricing responses as the mechanism of selection.

6 How did Insurers Risk Select?

Drawing from the prediction in equation 5, we investigate whether high-rated contracts differentially favored enrollment in healthy counties by lowering premiums in those counties. Specifically, we implement the following tripe-difference design

$$y_{clt} = \beta_0 \cdot risk_{cl} \cdot high_c \cdot post_t + \beta_1 \cdot risk_{cl} \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt}. \quad (8)$$

The variables *high* and *post* identify the high-rated group and the post reform period as in Section 5. The outcome variables are prices varying at the level of contract c, year t, and location l. A contract can offer different plans covering different counties. In each county, we generate contract-level prices from plan premiums and deductibles 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 measure in $risk_{cl}$ as the key independent variable in the analysis.[27] By construction, $risk_{cl}$ varies across locations within contracts and varies across contracts given a location.[28]

[27] To construct the measure, we take the full set of counties covered by a contract, rank them by the baseline FFS risk scores in 2009-2010, derive the median county risk in each contract, and construct the deviation-to-median measure for each county in the service area. Section 6.5 shows the robustness of key results to alternative measures of risk differences within contracts.

[28] Based on the variation in $risk_{cl}$, we cluster standard errors two-way at the level of counties and contracts.

We include contract-county fixed effects α_{cl} to absorb the baseline differences in prices and enrollments across contracts and counties.[29] 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.[30]

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.[31] Other time-varying factors at the contract-location level are harder to control for but can invalidate the design even with parallel pre-trends. For example, if high-rated contracts differentially entered high-bonus counties or exited high-risk counties, then selected service area characteristics would result in biased estimates of the price differences. However, we find little evidence of selection over service area characteristics such as risk and benchmark, mitigating this endogeneity concern.[32]

To illustrate the identifying variation, we show difference-in-differences estimates separately for high- and low-rated contracts, before showing the triple-difference estimates for high-rated contracts. In each case, we assess the validity of the identifying assumption based on raw trends and event study estimates. The assumption in the difference-in-differences setting is that premium differences by county risk scores follow parallel trends in pre-reform years. In the triple-difference setting, it further requires that premium differences between high- and low-rated contracts follow parallel trends in that period.

6.1 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 8 in Table 4. Part D premiums varied significantly

[29] The fixed effects absorb local consumer characteristics which did not vary with the payment reform. Because the payment reform is a supply-side regulation that did not affect consumers' knowledge of the quality rating or the enrollment process, we control for pre-existing consumer characteristics using fixed effects.

[30] When evaluated at the median county risk, $risk_{cl} = 0$ and interaction terms containing $risk_{cl}$ vanish in equation 8. β_2 gives the price change in the median county for high-rated contracts after the reform.

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

[32] Specifically, high-rated contracts did not enter additional counties or change the composition of covered counties based on risk scores or benchmarks. Appendix Table A5 shows the estimates.

with county risk scores in high-rated contracts. For every 10 percentage point increase in the risk score, Part D premiums increased by $1.53 in high-rated contracts (column 3), or by 7.2% above the mean. This effect is very similar to the premium responses estimated separately for high-rated contracts (column 2), and we do not detect similar responses 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.

Figure 6 shows the event study estimates and the raw trends of Part D premiums in high- and low-rated contracts. The event study shows that in both quality, Part D premiums followed similar distribution over county risk scores in 2009 and 2010, and differed similarly between high- and low-rated contracts in 2009 and 2010. After the reform, Part D premiums in high-rated contracts increased by as much as $2.50 per ten percentage point increase in the county risk score, or by 12% above the mean. In low-rated contracts, by contrast, there is no significant difference in premiums across county risk scores throughout the sample period.

To showcase the premium differences on the raw trend (panels a and c), 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 the trends of Part D premiums across binary risk groups for an average high- and low-rated contract. In high-rated contracts, Part D premiums deviated from pre-reform parallel trends and increased particularly in the riskiest counties since 2011. In low-rated contracts, by contrast, Part D premiums followed parallel trends with smaller differences across risk tails.

Part C Premiums. We then examine responses in Part C premiums in Table 5. 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. In Appendix Figure B2, Part C premiums trended similarly in both quality over the sample period, and the event study estimates generally show insignificant differences by county risk scores.

Zero Premiums. Since a fair number of Medicare Advantage contracts have zero premiums (cf Table 2), we examine the offer of zero-premium plans across service areas as one particular margin of response by insurers. Consistent with the effects on premiums, high-rated contracts significantly increased the offer of plans with zero Part D premiums in low-risk counties, and decreased the offer of such plans in high-risk counties (Appendix Table A6). We do not find similar responses in terms of zero Part C premiums, or by low-rated contracts. Appendix Figure B3 plots the raw trends.

Distributional Effects. Combining Part C and Part D premiums, total premiums increased by $4.05 more in high-rated contracts per ten percentage point risk score in the risk tails (Appendix Table A7, Appendix Figure B4), or 7.6% above the mean. Despite higher premiums in riskier counties, premiums of high-rated contracts stayed roughly constant in the median risk county.[33] This result implies that high-rated contracts may have shifted premiums from low-risk counties to high-risk counties, without changing the premium level of the median enrollee. In support of the distributional effect, we provide additional evidence that at the contract level, the premium paid by an average enrollee did not increase in high-rated contracts after the payment reform.[34] Taken together, the results suggest that high-rated contracts raised premiums in riskier counties and decreased them by a similar amount in healthier counties, without changing the average contract premium.

6.2 Varying Drug Deductibles to Risk Select Enrollees

Drug deductibles did not vary substantially across service areas or ratings. Appendix Table A9 shows estimates from equation 8 using drug deductible as the dependent variable. Both high- and low-rated insurance increased drug deductibles by approximately $3 per ten percentage point risk score. However, raw trends and the event study reveal a significant pre-reform effect in 2009 in high-rated contracts (Appendix Figure B5). Since 2010, the event study estimates indicate rising drug deductibles with county risk scores in high-rated contracts. Due to the noise in the data, we do not pursue differences in drug deductibles as a potential mechanism of the risk composition gain in high-rated insurance and mainly focus on premiums.

6.3 Mechanism

While the premium differences are consistent with the selection of healthier individuals in low-risk counties, a 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

[33] This is indicated by the coefficient before $High \cdot Post$ in column 3 and 6 of Table A7, or coefficient β_2 in equation 8.

[34] We showcase this point using a contract-level difference-in-differences (equation 7), where the outcome variable is the premium paid by an average enrollee in a contract-year. On average, premiums did not increase more in high-rated contracts (Appendix Table A8).

premium differences through the correlation with risk.[35]

Socio-Economic Factors. Appendix Table A10 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 A11 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 had little effects on the premium differences in high-rated contracts (Appendix Table A12).

Market Concentration. Appendix Table A13 explores the role of market concentration in driving the premium differences. Premiums increased with market concentration in high-rated contracts (column 2).[36] However, since more concentrated markets also have healthier enrollees,[37] the effect of market concentration tends to imply higher premiums in lower-risk counties. Moreover, controlling for market concentration increased the premium differences over risk scores (column 5). These results indicate that premium differences are not driven by the competitive effects on prices, but could be constrained by such effects in less concentrated markets.

Provider Quality. We next consider differences in provider costs and quality as alternative drivers of the premium differences over risk scores. 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. We do not detect a consistent

[35] Details of the county characteristics examined here are provided in Appendix D.2.

[36] 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. More concentration at the county level increases premiums in high-rated contracts, but concentration within county-quality pairs has no significant impacts on premiums (column 7-9).

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

differences over either quality in Appendix Table A14.

Provider Cost. We investigate the cost channel exploiting adjustments on fee-for-service (FFS) costs in Appendix Table A15. Premiums did not vary over costs by ratings. In high-selection contracts where risk scores decreased more (column 3), premiums increased with FFS costs. Similar patterns hold when we adjust for the price levels in costs in columns 5-8.[38] Adjusting price-standardized costs by risk scores in columns 9-12 cuts the effect size on high-selection contracts by half and renders the differences insignificant.[39] These results imply that premiums did not vary with local price levels or the practice of care, but varied with costs through the composition of risks across space.

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 in the fee-for-service risk scores. To remove cross-space differences in the coding of risk scores given health, Appendix Table A17 adjusts risk scores with the diagnosis intensity factors developed in Finkelstein *et al.* (2017).[40] Upon adjustment, we find a stronger variation of Part D premiums over risk scores relative to the main results in Table 4. The effects on Part C premiums and drug deductibles remain insignificant. Therefore 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.

6.4 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 8 using rebates as the dependent variable in column 4 of Appendix Table A18. The estimate suggests that rebates increased by $5.63

[38] The adjustment uses national input prices to calculate labor and facility costs, and override local reimbursement rates with a fixed national schedule.

[39] Appendix Table A16 finds similar patterns in the risk tails. Premium differences over costs are greatly reduced once we take out the risk component in costs.

[40] 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.

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 A7). Put together, premium differences account for 72% of the differences in the overall generosity by quality.[41]

In contrast to the significant price differences by county risk scores, average rebates and premiums did not increase in high-selection contracts. We showcase this point by estimating a contract-level difference-in-differences (equation 7) where the outcome variable is the average premium and rebate across markets for a high-rated contract. Despite price differences across percentiles of county risks, average premiums and rebates did not increase more in high-selection contracts (Appendix Table A20). We 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.

6.5 Sensitivity Analysis

Alternative Weights. In the main analysis, we weight plan premiums by enrollment to generate premiums for contracts. The resulting variables capture the joint effect 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 A21 and Appendix Figure B6. 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 A22, Appendix Figure B7).

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 A23. Appendix Figure B8 plots the event study estimates for this set of estimates. We also examine alternative measures of risk tails. Instead of percentiles, Appendix Table A24 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.

[41] Similar calculation for high-selection contracts suggests that premium differences (Appendix Table A19) account for about 65% of the rebate differences between low-rated and high-selection contracts.

7 Why Does the Payment Reform Induce Risk Selection?

Under the lenses of the model, the Quality Bonus Payment demonstration could incentivize selection if some measures in the quality rating are sensitive to enrollee risk types. In this section, we show that health outcome measures are biased against contracts with sicker enrollees. High-selection contracts, on the other hand, improved significantly on these measures mainly through selection.

7.1 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 A25). Here, we explore biases in the health outcome measures as a potential driver of selection.

Cross-Contract Evidence. We apply two empirical strategies to suggest the existence of biases in the health outcome rating due to risk scores. The first strategy exploits the payment reform and the cross-contract differences over baseline risk scores in a difference-in-differences analysis analogous to equation 7. Specifically, we estimate

$$y_{ct} = \beta \cdot risk_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}, \tag{9}$$

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 Table 6 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), on average.[42]

We also estimate separate effects for different types of outcome measures. Measures of self-reported health improvements drawn from survey responses in the Health Outcome Survey (HOS) are minimally correlated with risk scores (column 2). By contrast, measures of diabetes and blood pressure management, drawn from clinical data in the Healthcare Effectiveness Data and Information Set (HEDIS), are significantly and negatively correlated

[42] 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 8.1 we consider the effect of risk scores on all quality measures using an instrumental variable approach.

with risk scores (column 3). The overall correlation between risk scores and outcome ratings is completely driven by the HEDIS measures for chronic conditions.[43]

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 in the case-mix of health condition.[44] 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. Our second empirical strategy examines the relationship between outcome ratings in year t and risk scores in year $t-2$ with the following specification

$$y_{ct} = \beta \cdot risk_{ct-2} + \alpha_c + \tau_t + \epsilon_{ct}. \tag{10}$$

Unlike equation 9, where risk scores are held at the baseline, equation 10 explores how health outcome ratings respond when risk scores change over time. We lag risk scores by two years because outcome ratings rely on enrollees' medical records for two years before the current enrollment year (see Section 2).

If riskier individuals have worse measured outcomes, the negative effect on the outcome rating will appear with a two-year delay. We find supporting evidence 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 A26), and the effect is greater than the average differences by risk scores estimated in Table 6. We do not find similar correlation patterns for low- or high-rated contracts across other lag or lead periods.

7.2 Deterring Enrollees with Chronic Conditions Through Premiums

The selection incentive to improve the health outcome rating implies that premiums should respond to the chronic conditions targeted by the health outcome measures. We inspect such pricing responses here. Adopting the triple-difference design in equation 8, we compare premiums across counties with different diabetes and hypertension prevalence rates. We interact raw prevalence rates with coding-adjusted county risk scores and use the

[43] Appendix Figure B9 plots the raw trends and event study estimates.

[44] 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, Zhang *et al.* (2000) and 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.

health-adjusted prevalence rate as the key independent variable.[45] Therefore, a county is more favorable for risk selection if it has fewer patients with chronic conditions or because its patients have milder conditions.

We focus on diabetes in Appendix Table A27. In high-selection contracts, Part D premiums increased by \$9.47-\$12.44 per ten percentage point increase in the prevalence rate (column 6-7), or by 47%-63% above the average premium. Figure 7 plots the premium differences across high- and low-prevalence counties. Although the raw trends suggest larger premium differences in 2011-2012, the differences over continuous prevalence rates are comparable over the years in the event study. We find similar patterns but smaller magnitudes for hypertension (Appendix Table A28).

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 in the next section to quantify the extent of selection in the health outcome measures.

7.3 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.

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

$$y_{ct} = \alpha_c + \gamma_c \cdot post_t + \beta \cdot risk_{ct-2} + \tau_t + \epsilon_{it}, \tag{11}$$

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 $risk_{ct-2}$. The selection invalidates the ordinary-least-square (OLS) estimate of β. We employ an IV strategy to estimate the effect

[45] Prevalence rates are adjusted downward in counties where patients have milder conditions and fewer complications. We provide more details on the prevalence rates in Appendix D.

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.[46] 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

$$riskiv_{ct-2} = Corr(p_{ct-2}, R_c) = \frac{1}{|N_c|} \sum_{l \in N_c} \frac{(p^l_{ct-2} - \bar{p}_{ct-2}) \cdot (R^l_c - \bar{R}_c)}{\sigma_{p_{ct-2}} \cdot \sigma_{R_c}}, \tag{12}$$

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 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 risks $\sigma_{p_{ct-2}}$ and prices σ_{R_c}, and use the correlation coefficient $Corr(p_{ct-2}, R_c)$ as the instrument $riskiv_{ct-2}$.[47]

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 markets as instruments in the first stage. We isolate premium difference by the health of enrollees using coding-adjusted risk scores for R_c (Finkelstein *et al.*, 2017) in equation 12. We construct additional instruments for cross-county premium differences by diabetes and hypertension prevalence rates based on our results in Section 7.2.[48]

For the instruments to be valid, premium differences should have no direct impacts on the contract's quality rating other than through the risk score $risk_{ct-2}$. This requires that premium differences affected the risk composition of enrollees in the contract, but did not affect unobserved determinants of contract rating in the error term ϵ_{ct}. We examine the

[46] 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 $\gamma_c \cdot post$.

[47] 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.

[48] Premium differences by diabetes prevalence rates are instrumented by $diabiv_{ct-2} = Corr(p_{ct-2}, D_c)$, where D_c is the vector of baseline diabetes prevalence rates in the counties covered by contract c. Similarly, the instrument $hyptiv_{ct-2} = Corr(p_{ct-2}, H_c)$ captures premium differences by hypertension prevalence rates, where H_c is the vector of baseline hypertension prevalence rates in the counties covered by contract c.

plausibility of the exclusion restriction drawing from the analysis in Section 6.3, where we show that premium differences 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 11 in Table 7. The OLS estimates do not indicate significant effects of risk scores on the HEDIS outcomes.[49] Based on these estimates, health outcomes improved by 1.8 percentage points in high-rated contracts after the payment reform (column 2), and by 1.68 percentage points in high-selection contracts (column 4). However, risk selection can bias the OLS estimates towards zero, masking the effect of risk scores on the outcome measures.

To correct for this endogeneity, we instrument contract risk scores, $risk_{ct-2}$, by the premium differences over county risk scores, diabetes prevalence rates and hypertension prevalence rates in Panel B. Consistent with the drop in risk scores, premium differences significantly predict risk scores in high-rated contracts (column 2) and particularly in high-selection contracts (column 4-5).[50] In these contracts, two-stage-least-squares (TSLS) estimates show 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.[51]

We then apply the TSLS estimates to 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 through $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}}$, where ΔRisk is the risk composition gain over 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.[52] Since the TSLS estimates already control for the effect of risk selection on the health outcome measures, we infer the health gains of a standard-risk enrollee from estimates of $\gamma_c \cdot post$ in equation 11. Adjusted for risk, health outcomes improved by a modest $\gamma_c \cdot post = 0.24$ percentage

[49] Recall that the HEDIS outcomes tally the share of diabetes and hypertension patients who have controlled their conditions below specific medical thresholds. We look at the percentage of such enrollees as the dependent variable in Table 7.

[50] We show first-stage estimates in Appendix Table A29. Premium differences by county risk scores strongly and negatively predict contract risk scores in high-rated contracts. Premium differences by diabetes prevalence rates predict higher risk scores for high-rated contract in riskier areas (column 3), but not for high-selection contracts. The joint prediction power is concentrated in high-selection contracts. We further explore the choice of instruments on the implied rating gains of selection in Appendix Table A31.

[51] 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 of high-rated contracts.

[52] 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 5).

points on average for high-selection contracts. The adjustment significantly revises the conclusion from OLS estimates in Panel A. Despite a 1.68 percentage point increase in health outcome measures, selection of healthier enrollees accounted for 86% of the health measure gains in high-selection contracts.[53]

To summarize, high-selection contracts improved the health outcome measures by selecting healthier enrollees. Adjusted for risk, health outcomes improved only modestly in high-selection contracts. In the next section, we calculate the financial gains of selection by estimating the effect of selection on the overall rating linked to payments.

8 Discussion

8.1 The Cost of Selection

In this section, we estimate the effect of selection on the bonus payments by first estimating the selection gains on the quality rating. We then infer the savings in bonus payments after removing the selection gains in the quality rating.

Selection and the Quality Rating. We apply the IV strategy developed in Section 7.3 to estimate the effect of risk selection on the quality rating. We group measures by the weights they receive in the overall rating and estimate the effect of selection on the star ratings of outcome (3.0 weights), access (1.5 weights), and process (1.0 weights) measures using equation 11. We show results for high-selection contracts in Appendix Table A30. Risk scores have large, negative impacts on the outcome rating (column 1-2), where nearly 80% of the rating gain after the payment reform is due to selection.[54] Applied to the risk composition gain over low-rated contracts from 2010 to 2012, these estimates imply that selection increased outcome ratings by 0.46 stars in high-selection contracts,[55] and increased the overall rating by 0.23 stars.[56] Ratings on the lower-weighted access and process measures are minimally affected by risk scores.

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

[54] The selection effect is comparable to but different from the 90% calculated in Section 7.3 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 include all measures receiving 3.0 weights in the health outcome category, whereas in Section 7.3 we focused only on the three HEDIS measures.

[55] Since our counterfactual analysis relies heavily on the effect of selection on the health outcome ratings, we show that the effect is robust to alternative choices of instruments in Appendix Table A31.

[56] This is because health outcome measures account for 50% of the overall star rating in high-selection contracts. Within health outcome measures, HEDIS outcomes are most sensitive to risk scores (Appendix Table A32), followed by drug related outcome ratings. Together, these measures explain all of the selection effect on the outcome rating. Self-reported health improvements are not affected by risk scores.

To construct counterfactual ratings absent the selection response, we first recover the underlying continuous star rating of each contract in 2014.[57] From the continuous rating, we subtract the selection gains due to the risk score change from 2010 to 2012.[58] The adjusted rating holds the risk composition fixed at 2010, and removes the effect of selected risk scores since 2011 on the quality rating.

We compare risk-adjusted star ratings with the initial rating in panel (a) of Figure 8. The horizontal axis groups high-selection contracts by the initial star rating in 2014. The vertical axis shows the percent of enrollees experiencing a change in the star rating, conditional on the initial rating. Adjusted for risk, the quality rating drops to 3.0 stars for around half of the enrollees in 4.0-star contracts and all enrollees in 3.5-star contracts. The affected contracts are marginal high-rated contracts below 4.0 stars on the continuous scale in 2014. 98% of enrollees in the 3.5-4.0 star range were in marginal high-rated contracts. Even small increases in the risk score would downgrade their contracts to 3.5 stars and below. By contrast, risk adjustment has smaller impacts in the 4.5-5.0 range, where 96% of enrollees are in contracts rated above 4.5 stars on the continuous scale. Adjusted for risk, all contracts in the 4.5-5.0 range still maintain at least a 4.0-star rating.

Selection and Payments. The new star rating changes the generosity of bonus payments to insurers (Table 1).[59] To determine payments to insurers at the risk-adjusted rating, we assume that insurers respond by aligning bids to the new benchmark while keeping rebates to enrollees unchanged. We show empirical support for this assumption by estimating a near-zero pass-through of benchmark bonuses to enrollees in high-selection contracts (Appendix Table A33).[60] We therefore infer counterfactual bids by inverting equation 1 holding rebates at the pre-adjustment level.[61]

We then calculate the counterfactual payments under the risk-adjusted quality rating. The difference in bonus payments indicates the extent of overpayments in the current rating that fails to adjust for enrollee risks. Panel (b) of Figure 8 shows that overpayments are largest in the 3.5-4.0 star range, where an average contract gained $25 in bonus payments due to selection. Risk adjustment would downgrade the star ratings of these contracts to

[57] The continuous rating is a weighted average of star ratings across measures. The weighted average is then rounded to the nearest half star as the overall star rating. More information is in Section 2.

[58] We use the risk composition change relative to low-rated contracts to estimate the selection gains. We obtain similar results by constructing counterfactual risk scores for each high-selection contracts using the synthetic control method (Abadie *et al.*, 2010).

[59] Since the 2014 star rating determines payments in 2015, we re-calculate 2015 payments using the risk-adjusted rating.

[60] Specifically, we find that high-selection contracts submitted higher bids after the introduction of benchmark bonuses, narrowed the distance between bids and benchmarks, and kept the rebates to enrollees unchanged. Appendix Figure B11 plots the raw trends of the bidding adjustment.

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

3.0-3.5 stars. Because benchmark bonuses no longer apply below 4.0 stars under the ACA model, crossing the 4.0-star cut-off significantly increased the bonus payments to marginal high-quality contracts. Although similar incentives exist for contracts in the 4.5-5.0 range, overpayments are small ($0.59) because none of the contracts in this range fell below 4.0 stars upon risk adjustment.[62]

To evaluate the size of the overpayment, we compare it with the benchmark bonus to high-selection contracts. Including the bonus, high-selection contracts received $87 more in benchmarks in 2015 relative to low-rated contracts.[63] Selection raised the bonus payments to high-selection contracts by $12, or 14% of the benchmark bonus in 2015. In marginal high-rated contracts (3.5-4.0 stars), selection increased payments by 29% of the benchmark bonus.[64]

8.2 Distribution of Quality Ratings Across Counties

We next explore the distributional implications of the selection responses on the market share of high-rated contracts across counties. Because premiums decreased significantly in the healthiest counties covered by high-rated contracts, the market share of high-rated contracts tends to decrease with county risk scores moving from the healthiest to the riskiest counties. We examine this implication by estimating the distribution of high-rated contracts across county risk scores using the following specification

$$\begin{aligned} y_{clt} = {} & \beta_0 \cdot risk_l \cdot high_c \cdot post_t + \beta_1 \cdot risk_l \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta_3 \cdot high_c \cdot risk_l \\ & + \beta \cdot X_{lt} + \alpha_c + \gamma_l + \tau_t + \epsilon_{clt}, \end{aligned} \tag{13}$$

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. We therefore examine the changes in market share y_{clt} as risk score increases from the healthiest to the riskiest counties in $risk_l$.[65] We allow market shares to differ by rating in β_3, and control for the effect of bonus payments across counties in β_1 and on high-rated insurance in β_2. β_0 then estimates the effect of bonus payments on the market share of high-rated insurance across

[62] We find similar overpayments using the synthetic control method (Appendix Figure B10), which gives identical results in the 3.5-4.0 star range, and slightly lower overpayments above 4.5 stars.

[63] The estimate is the event study coefficient for 2015 in an extended analysis of contract benchmarks using the difference-in-differences model in equation 7.

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

[65] This is different from equation 8 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 in equation 13. We control for the same set of county variables in X_{lt}.

county risk scores.

Figure 9 plots the raw trends of market shares on the left panels and the event study estimates from equation 13 on the right panels. The raw trends focus on counties in the lower and upper 15% of risk scores, where market shares diverged markedly by rating. At the contract level (panel a), market shares of high-rated contracts increased in the lowest-risk counties (gray lines) and decreased in the riskiest counties (blue lines). Panel (c) looks at the overall market share of high- and low-rated contracts, showing a similar divergence in the market shares of high-rated contracts across the risk tails (solid lines). Across all counties, Appendix Table A34 estimates that a 10 percentage point increase in the risk score 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). We examine robustness and estimate the distributional effects on premiums in Appendix E.

The shifts in the spatial distribution of high-rated insurance are the result of a supply shock on insurer revenues. Consumer knowledge of the quality rating and preferences for quality are not directly affected by the payment reform. Thus, insurer responses to payment incentives are the primary driver behind the diverging market shares of high-rated insurance across risk tails. Selecting insurers directed bonus payments and the supply of high-rated insurance to the healthiest counties, worsening the access to high-rated insurance in the riskiest counties. Within high-rated contracts, insurance benefits shifted from the sickest to the healthiest enrollees, worsening the financial protection of insurance to the sickest enrollees. Therefore, across counties and contracts, insurer selection increased premiums and lowered benefits for sicker patients in the riskiest counties, hurting in particular the healthcare access of the more vulnerable population.

8.3 Improving the Quality Rating

The selection response suggests that the raw HEDIS data are not directly suitable for comparing health outcomes across contracts. When the case-mix of health conditions differ, the HEDIS outcomes bias the quality rating against contracts with sicker enrollees. To remove the bias, CMS may consider risk-adjusting the medical thresholds used in the health outcome measures. The adjustment accounts for predictable outcome differences due to differences in the health status of enrollees. If sicker patients have worse outcomes in expectation, then measuring health improvements relative to the risk-adjusted thresholds can limit the selection incentives on insurers.[66]

[66] When linked to payments, the adjusted rating compensates insurers for predictable outcome differences due to the risk of enrollees. A similar idea underlies the cost prediction for capitated payments.

CMS may also consider adjusting the incentive structure across different measures of the quality rating. Currently, health outcome measures account for 25% of all quality measures, but receive three times the weights as the process measures and twice the weights as the access measures. Upon weighting, health outcomes account for 50% of the final rating linked to payments. The high-powered incentives can induce both investments and the gaming of these measures. We find that selecting contracts gained significantly from enrolling healthier individuals but, adjusted for risk, health improvements were small in selecting contracts. Therefore, down-weighting the health outcome measures is likely to limit the selection responses without harming the health of enrollees.

9 Conclusions

This paper studies how suppliers respond to regulations linking government subsidies to the quality of social services, and the implications for consumers. We show that the presence of an efficiency-equity trade-off complicates the design of performance-based incentives in the social sector. Our results on the Medicare Advantage market suggest that policy efforts to improve the quality of social services can have unintended consequences on enrollee access to quality insurance. In particular, failing to correct for selected risk types in the quality rating worsened the access to high-rated insurance in the riskiest counties, without substantially improving insurance quality. Our findings illustrate that supply-side responses to the payment incentives can worsen the geographic disparity in service quality, undoing the purpose of the reform. Legislators should carefully consider the distributional consequences of pay-for-performance to effectively improve the quality and accessibility of social services.

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.

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.

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

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

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. (2015). Medicare Part D: Are insurers gaming the low income subsidy design? *American Economic Review*, **105** (4), 1547–80.

— 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. (2015). *Subsidy design in privately-provided social insurance: Lessons from medicare part d*. Tech. rep., National Bureau of Economic Research.

Deryugina, T. and Molitor, D. (2018). *Does when you die depend on where you live? Evidence from Hurricane Katrina*. Tech. rep., National Bureau of Economic Research, National Bureau of Economic Research, No. w24822.

Dickman, S. L., Himmelstein, D. U. and Woolhandler, S. (2017). Inequality and the health-care system in the USA. *The Lancet*, **389** (10077), 1431–1441.

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

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

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.

— and Layton, T. J. (2018). Upcoding: Evidence from Medicare on squishy risk adjustment. *Available at SSRN 2612913*.

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.

Gupta, A. (2017). Impacts of performance pay for hospitals: The readmissions reduction program, Mimeo, Wharton.

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.

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

Mahoney, N. and Weyl, E. G. (2017). Imperfect competition in selection markets. *Review of Economics and Statistics*, **99** (4), 637–651.

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., 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.

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, **366** (6464), 447–453.

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.

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

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

Zhang, Q., Safford, M., Ottenweller, J., Hawley, G., Repke, D., Burgess, J., Dhar, S., Cheng, H., Naito, H. and Pogach, L. M. (2000). Performance status of health care facilities changes with risk adjustment of hba1c. *Diabetes Care*, **23** (7), 919–927.

Tables and Figures

Table 1: Bonus adjustments on benchmarks and rebates

	Star Rating					
Year	≤2.5	3.0	3.5	4.0	4.5	5.0
	Benchmark Bonus $\theta^{star} = 1 + \%$					
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. after the payment reform ended in 2014, the ACA payment model took effect. 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 2: 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 D.

Table 3: 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)
Treat	high-rated	high-rated + risk		high-rated + risk	
		<median	>median	<25%	>75%
Control		low-rated		low-rated	
y mean	0.97	0.97	0.97	0.97	0.97
$R\hat{2}$	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 heterogeneous effects of the payment reform on the risk score of high-rated contracts. Column 1 shows the average effect on high-rated contracts. Column 2-5 shows heterogeneous effects for contracts with different service area risks. Column 2 and 3 divide high-rated contracts by the median service area risk (0.975), and estimate separate effects for those below (column 2) and above (column 3) the median. Column 4 and 5 focus on high-rated contracts in the lower and upper 25% tails of service area risks. Column 4 estimates the effect 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. Standard errors clustered at the level of contracts in the parenthesis.

Table 4: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table 5: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table 6: 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 effect of baseline enrollee risk scores on the outcome ratings. The difference-in-differences estimates compare 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 level of contracts in the parenthesis.

Table 7: 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

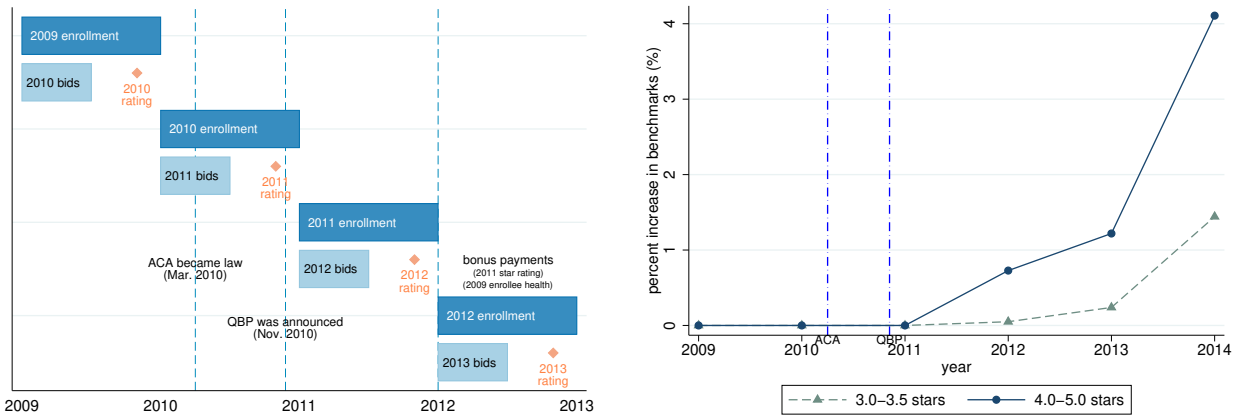
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

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. 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 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 outcome measures, we infer risk-adjusted health 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 health 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 level of contracts in the parenthesis.

Figure 1: Star rating computation and its implications for benchmarks

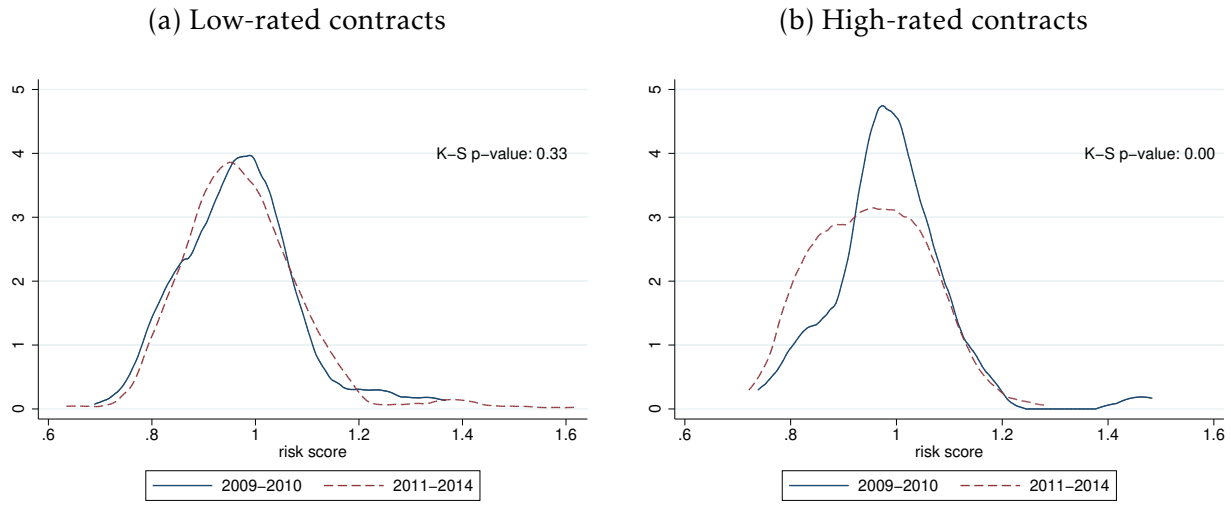
(a) Timeline of bidding, enrollment, and star rating disclosure, 2009-2012

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



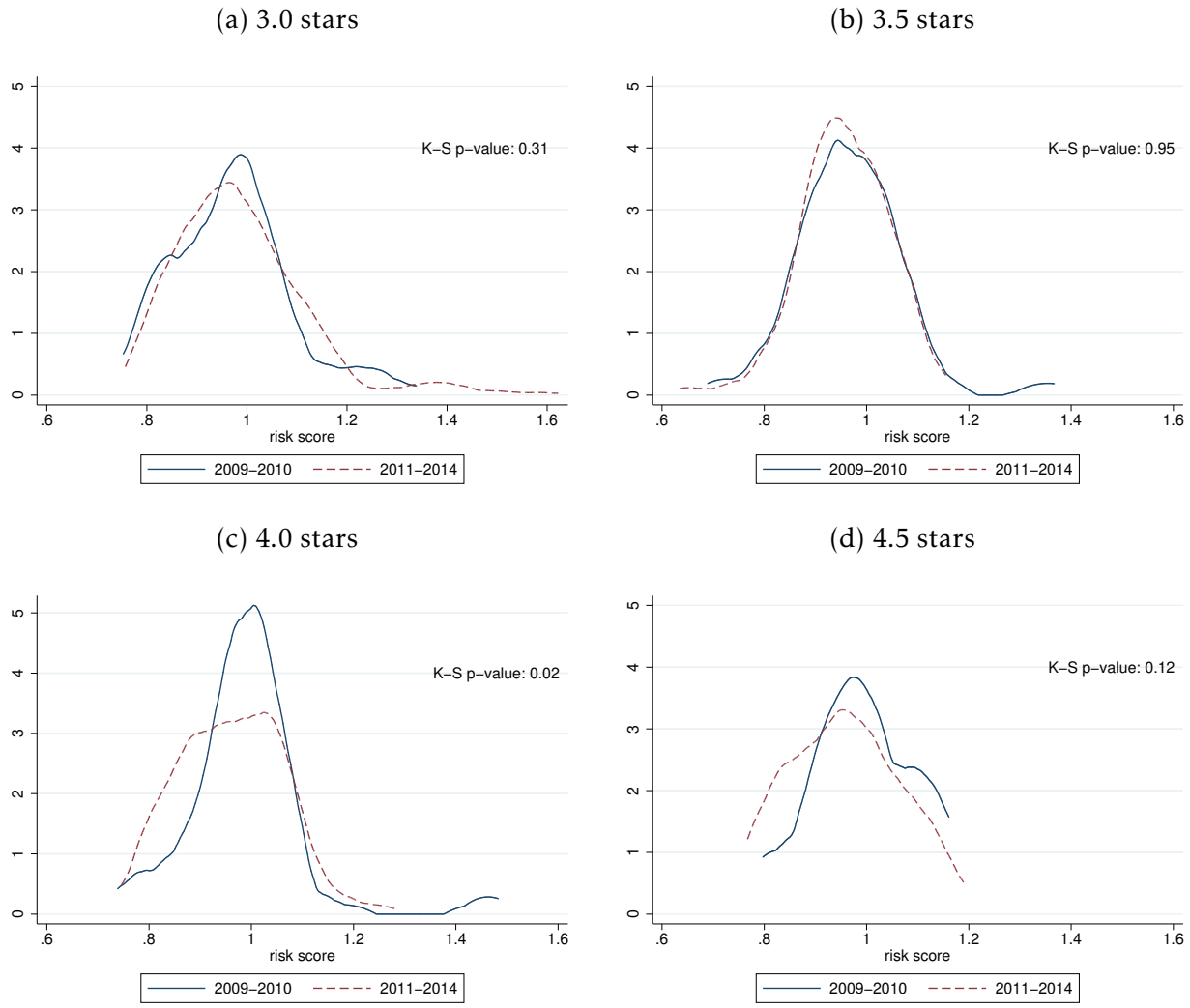
Notes: The figure in panel (a) plots the timeline of Medicare Advantage (MA) enrollment, plan bidding, and star rating disclosure for enrollment years 2009-2012. Insurers submit bids by the June of year $t-1$ for their plan offerings in year t. The star rating for year t is released in the Fall of year $t-1$. Since the bidding occurs before the release of year-t rating, quality bonuses – effective since enrollment year 2012 – are calculated based on the star rating of year $t-1$. Therefore 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. The figure also marks key policy change dates including the passage of ACA in March 2010 and the announcement of QBP revising the ACA model for 2012-2014 in November 2010. The figure in 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. We apply the bonus rates to the raw county benchmarks, and average the bonus-adjusted benchmarks across counties to predict contract benchmarks. In the prediction, we restrict counties to those already covered by the insurance contract prior to the payment reform. We compare the predicted benchmarks with the pre-reform benchmarks and plot the percent increase after the payment reform.

Figure 2: Effect on risk scores, kernel density



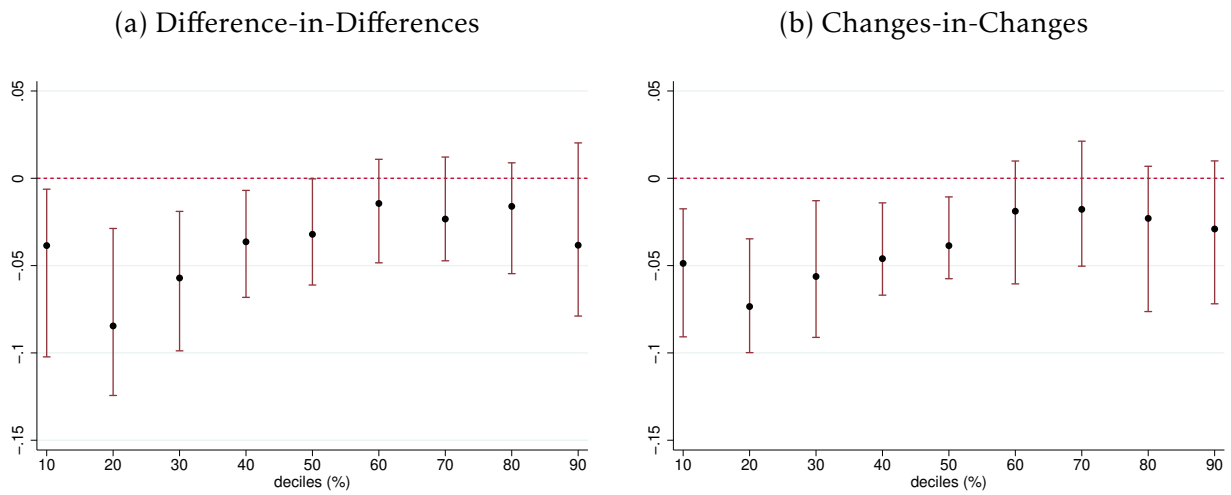
Notes: The figure plots the kernel density of risk scores for high-rated contracts in panel (a), and for low-rated contracts in panel (b). Separate densities are drawn for the before (2009-2010) and after (2011-2014) the payment reform. We test for the null of equal distribution applying the Kolmogorov–Smirnov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

Figure 3: Effect on risk scores by the baseline rating, kernel density



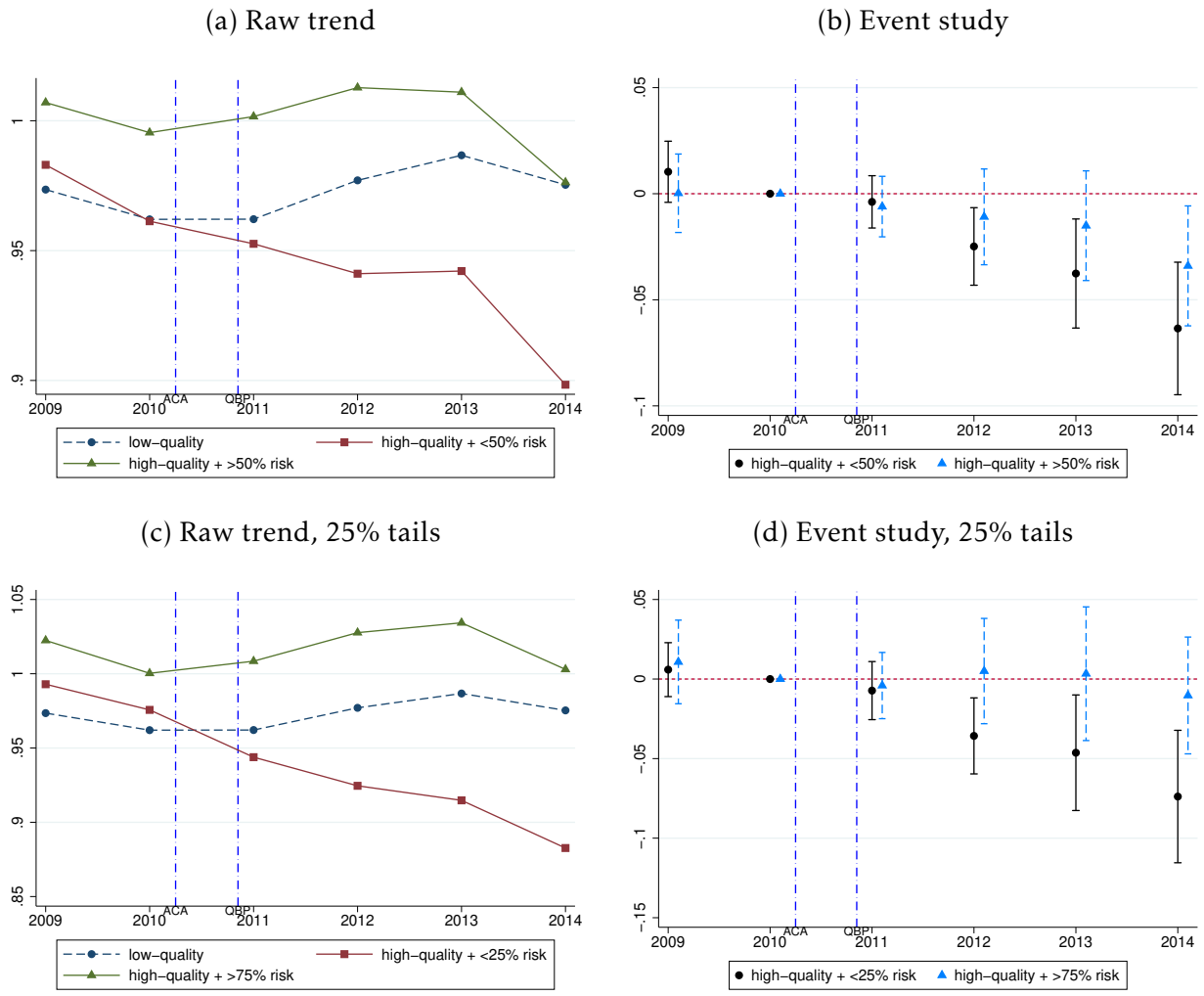
Notes: The figure plots the kernel density of risk scores by the baseline rating of contracts. For each rating from 3.0 stars to 4.5 stars, the figure compares the density of risk scores before and after the payment reform, and tests for the null of equal distribution applying the Kolmogorov–Smirnov (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.

Figure 4: Distributional effects on risk scores, by deciles



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 A4 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.

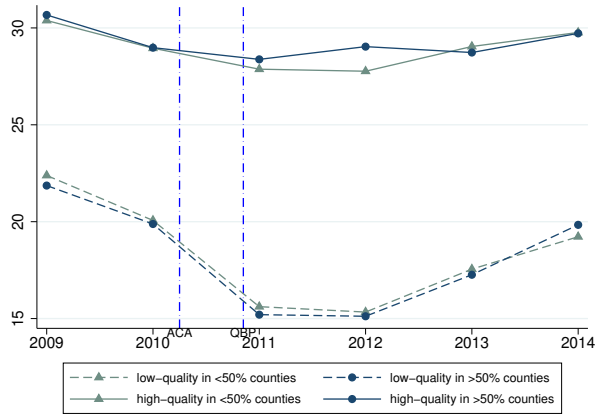
Figure 5: Effect on risk scores, by service area risks, event study



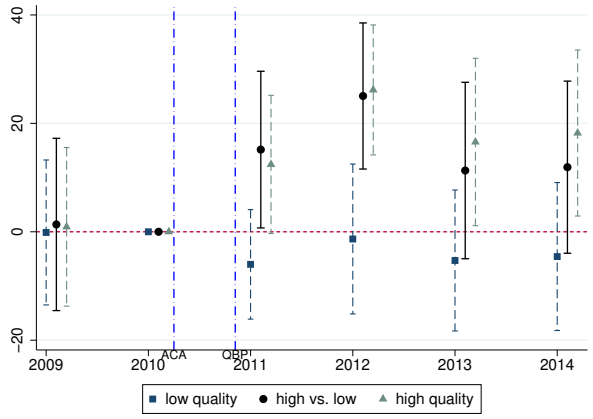
Notes: The figure shows the heterogeneous effects on high-rated contracts with different service area risks. Panel (a) shows the raw trends of risk scores for high-rated contracts above and below the median service area risk (0.975) and for low-rated contracts. Panel (b) shows the event study estimates 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 level of contracts in panel (b) and (d).

Figure 6: Effect on Part D premiums, within-contract differences, event study

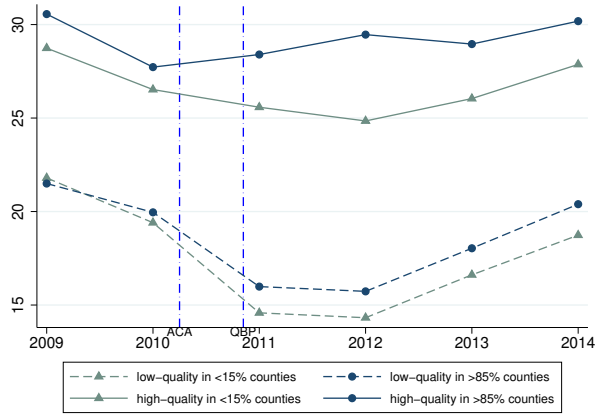
(a) Raw trend



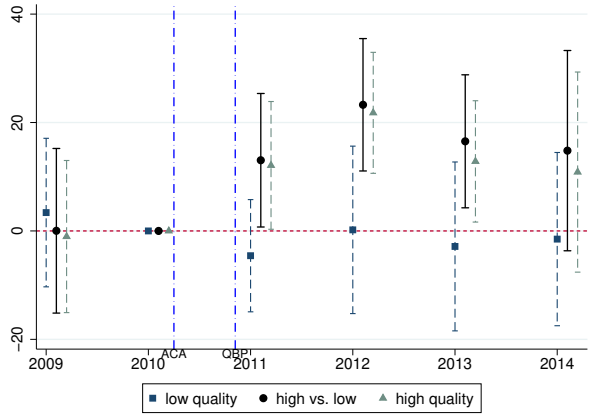
(b) Event study



(c) Raw trend, 15% tails



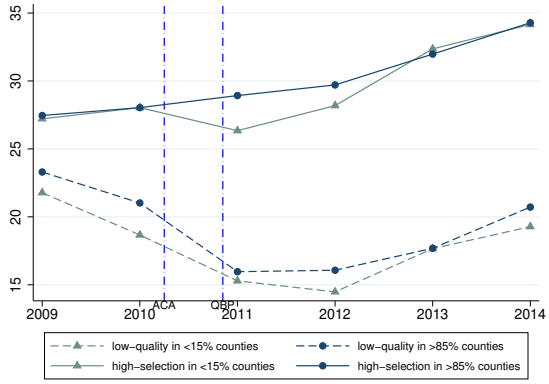
(d) Event study, 15% tails



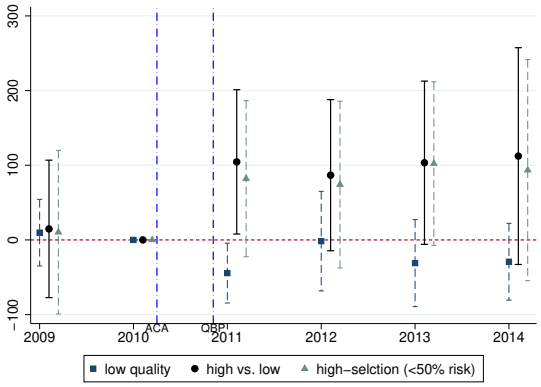
Notes: The figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated 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 standard errors clustered two-way at the level of counties and contracts.

Figure 7: Effect on Part D premium, within-contract differences over health-adjusted diabetes prevalence rates, high-selection contracts, event study

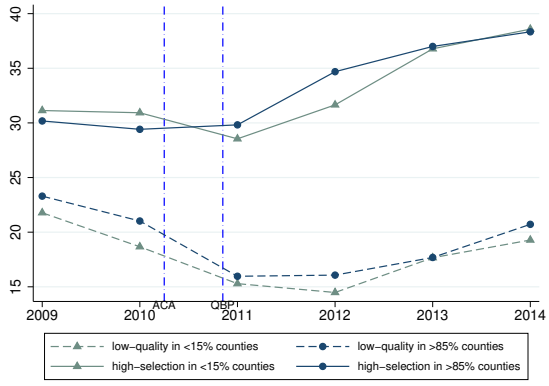
(a) High-Selection (<50% Risk), raw trend



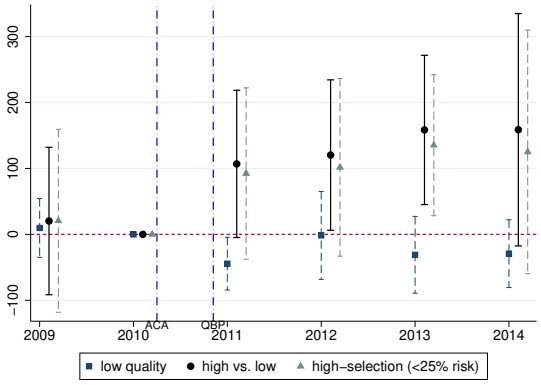
(b) High-Selection (<50% Risk), event study



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



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



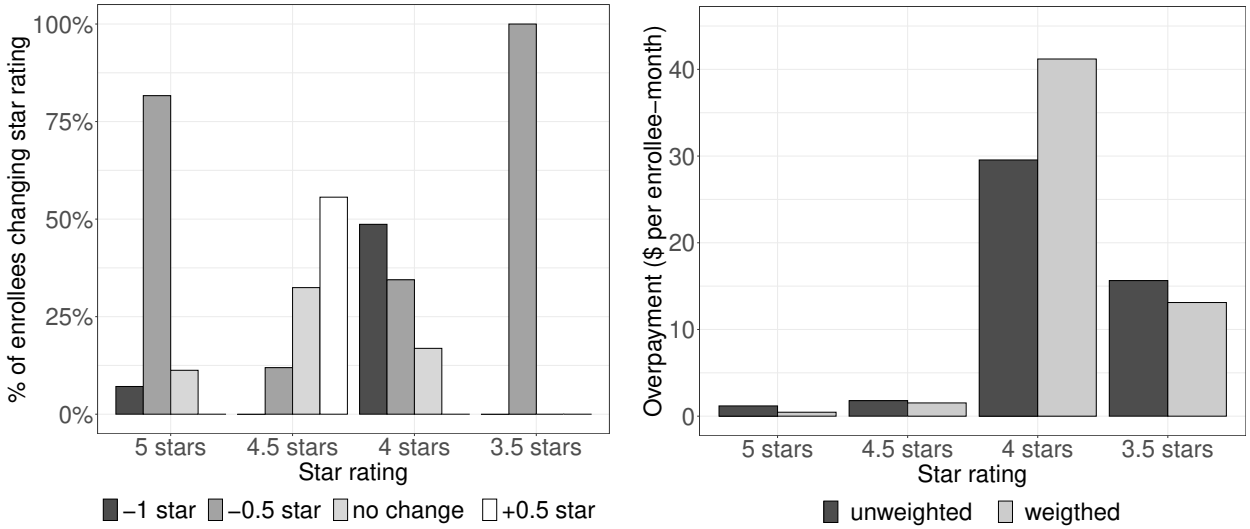
Notes: The figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract differences over county differences in health-adjusted diabetes prevalence rates in the right panels. The health-adjusted prevalence rate multiplies the raw prevalence rate by 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 standard errors clustered two-way at the level of counties and contracts.

Figure 8: Effects of selection on the quality rating and overpayments

(a) Share of Enrollees with Star Rating Change

Absent Selection

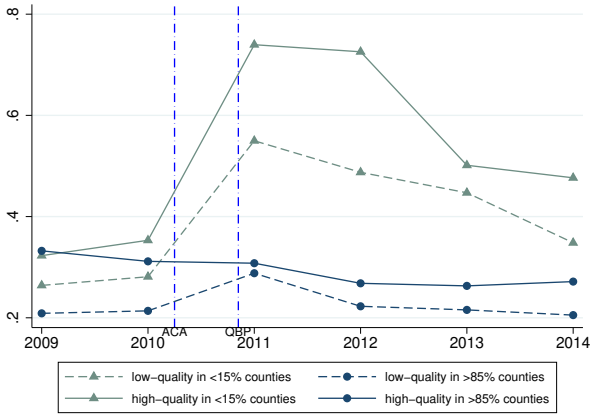
(b) Overpayments due to Selection



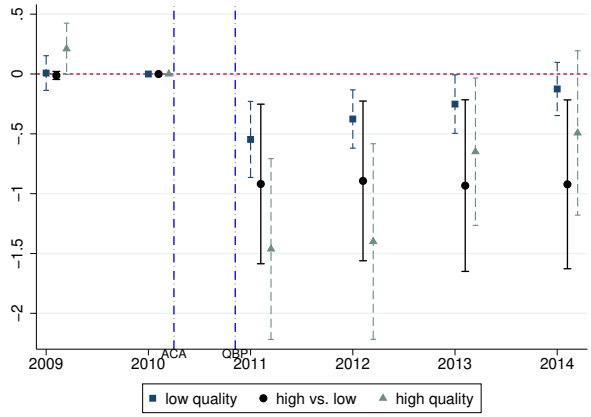
Notes: The figure shows the effect of adjusting risk selection on the overall star ratings of high-selection contracts in panel (a) and on the payments to these contracts in panel (b). Panel (a) plots for each overall star rating level in 2014 (horizontal axis) the percentage of enrollees receiving lower (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. 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. 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 9: Effect on market shares, cross-county differences, event study

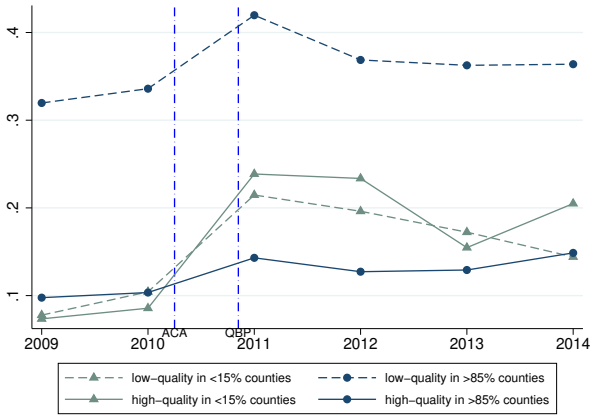
(a) Contract-County-Year, raw trend, 15% tails



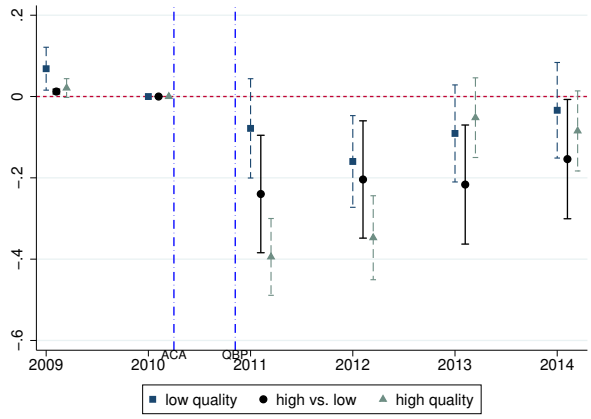
(b) Contract-County-Year, event study, All Counties



(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 standard errors clustered two-way at the level of contracts and counties in panel (b), and based on robust standard errors clustered by counties in panel (d).

Online Appendix

A Additional Tables

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

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

Notes: The table lists the Part C measures in the 2013 quality rating, the weight of the measure in the final rating, the type (outcome, process, or access) of the measure based on the weight, the data source, and the measurement period.

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

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

Notes: The table lists the Part D measures in the 2013 quality rating, the weight of the measure in the final rating, the type (outcome, process, or access) of the measure based on the weight, the data source, and the measurement period.

Table A3: 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 shows the bonus rates in year $t+3$ for contracts with different star ratings in year t. Contracts failing to maintain the star rating are subject to greater loss of bonus payments if bonus rates are higher at the original star rating. This summarizes the incentive to risk-select healthier enrollees when healthier enrollees improve the star rating. The table showcases the selection incentive during the Quality Bonus Payment Demonstration (QBP) in 2012-2014. In the empirical analysis, we also allow for anticipatory effects, which may happen if insurers respond immediately to the passage of ACA in March, 2010 and begin selecting healthier enrollees in year 2011. In this case, the selection incentive in 2011 is determined by the ACA payment model and identical to the incentive structures for 2012-2014 in the table. Since the ACA model was announced prior to the payment reform and was phased-in over 2012-2014 before implemented fully in 2015, the incentive structure of the ACA model was known and anticipated by insurers.

Table A4: 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, standard errors in the 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 A5: 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. Standard errors clustered at the level of contracts in the parenthesis.

Table A6: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A7: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A8: Effect of the payment reform on premiums and rebates

	(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 level of contracts in the parenthesis.

Table A9: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A10: 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 4). All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A11: 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 4). All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A12: 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 4. 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A13: 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 4). All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A14: 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 4). All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A15: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A16: 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 4). All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A17: 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 5), Part D premiums (Table 4), and drug deductibles (Appendix Table A9), 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A18: 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 standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A19: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A20: 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 level of contracts in the parenthesis.

Table A21: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A22: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A23: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A24: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A25: Weight increase and the composition of measures 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 to the overall rating, corresponding to a weight increase from 1.0 to 3.0 in 2012. Column 3-4 estimate the contribution of access measures, where 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 level of contracts in the parenthesis.

Table A26: 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 level of contracts in the parenthesis.

Table A27: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A28: 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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A29: 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 level of contracts in the parenthesis.

Table A30: 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 level of contracts in the parenthesis.

Table A31: 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 level of contracts in the parenthesis.

Table A32: 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 level of contracts in the parenthesis.

Table A33: 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 level of contracts in the parenthesis.

Table A34: Effect of the payment reform on market shares, across county risk scores

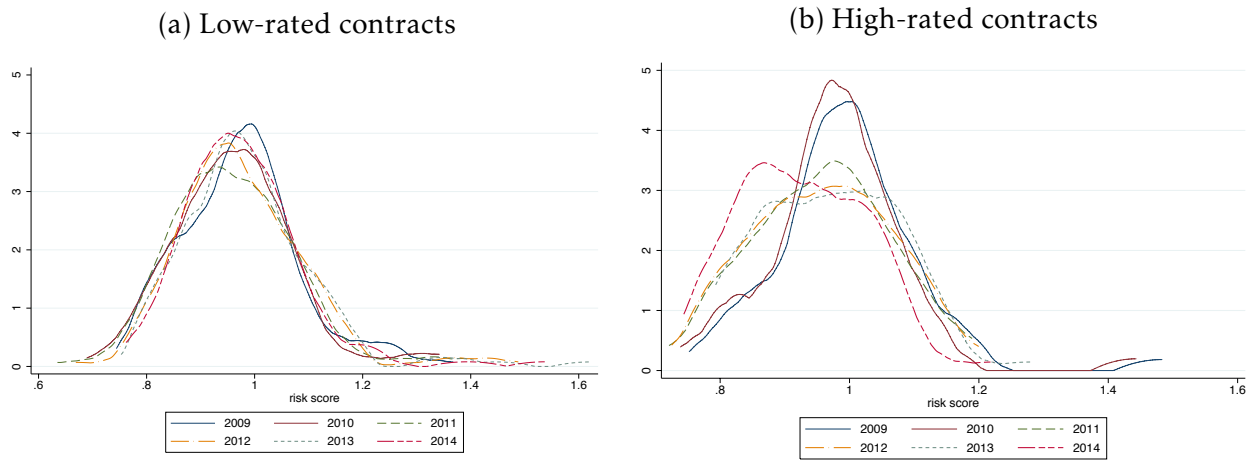
	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-0.90**			-0.19**
			(0.35)			(0.074)
Risk · Post	-0.38***	-1.18***	-0.36***	-0.14***	-0.24***	-0.095*
	(0.12)	(0.34)	(0.11)	(0.050)	(0.045)	(0.050)
High · Post			0.88**			0.15**
			(0.36)			(0.072)
Risk · High			0.65*			-0.83***
			(0.38)			(0.079)
Observations	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 on the market shares of Medicare Advantage contracts across county risk scores. Column 1-3 estimates the effects on contract market shares using equation 13. Robust standard errors clustered two-way at the level of contracts and counties in the 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 level of contracts in the parenthesis. Market shares (y mean) are lower in column 4-6 due to the incidence of zero market shares in the balanced panel.

B Additional Figures

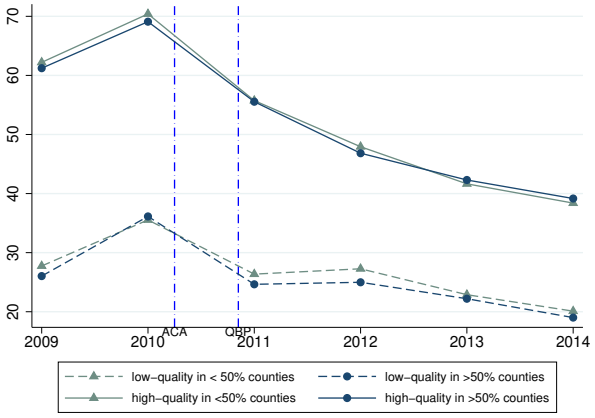
Figure B1: Contract risk scores, kernel density, by star rating and year



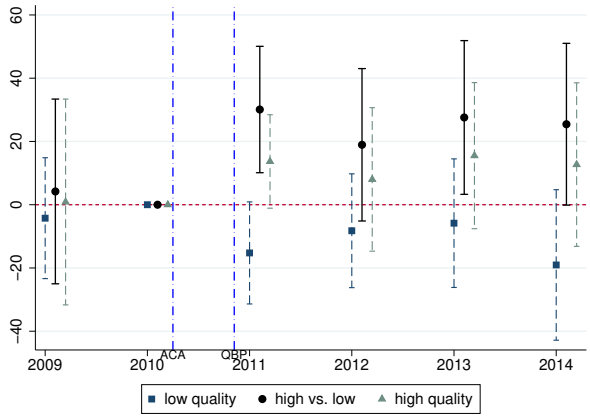
Notes: The figure plots the kernel density of risk scores in high-rated contracts in panel (a), and the density of risk scores in low-rated contracts in panel (b). Separate density is drawn for each year. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

Figure B2: Effect on Part C premiums, within-contract differences, event study

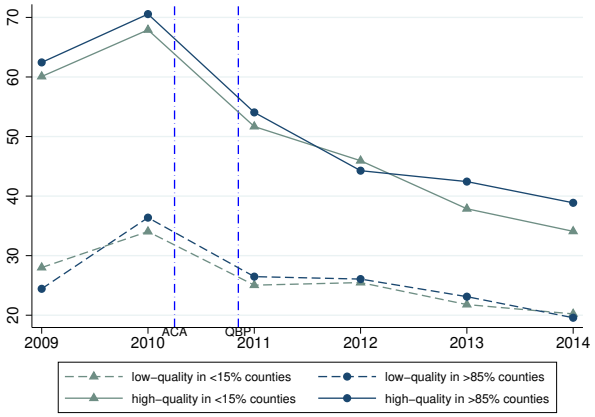
(a) Raw trend



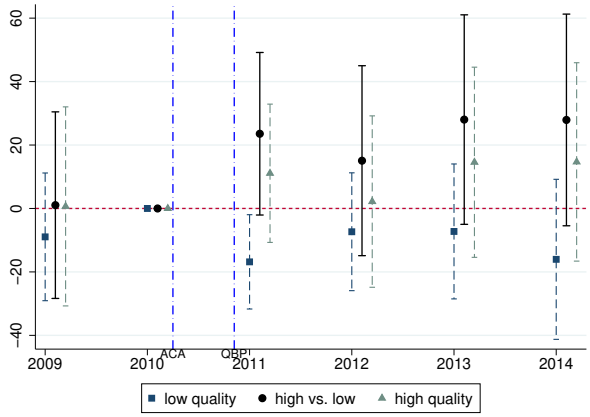
(b) Event study



(c) Raw trend, 15% tails



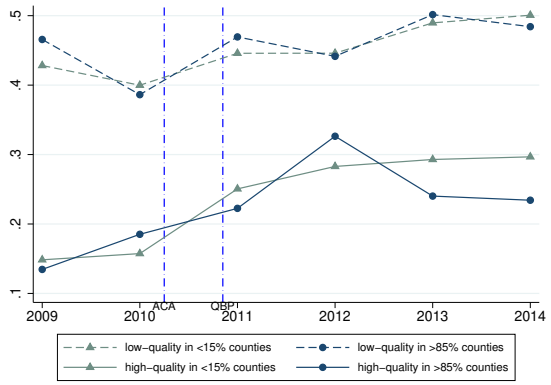
(d) Event study, 15% tails



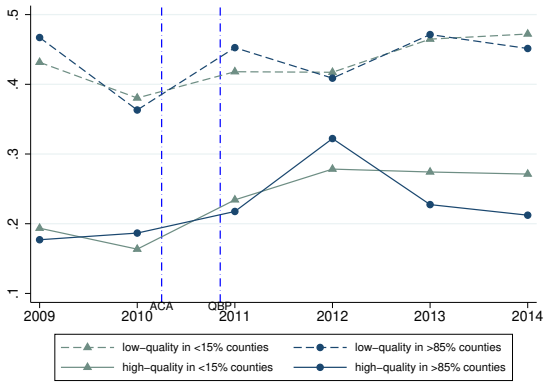
Notes: The figure plots the raw trends of Part C premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated 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 standard errors clustered two-way at the level of counties and contracts.

Figure B3: Effect on zero-premium and zero-deductible plans and enrollment, 15% tails, raw trends

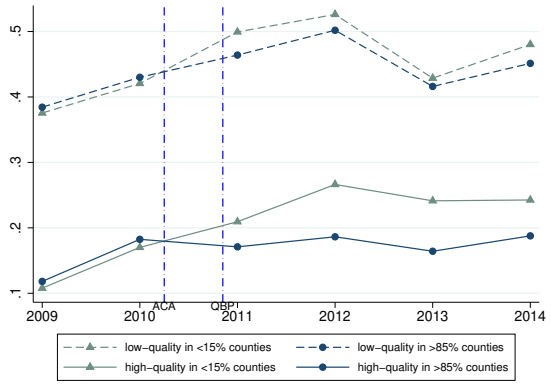
(a) Zero Part C premium



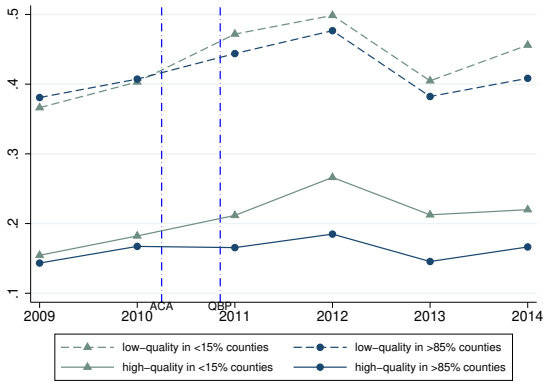
(b) Zero Part C premium, unweighted



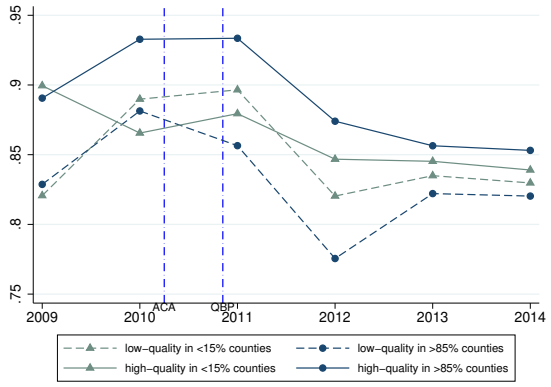
(c) Zero Part D premium



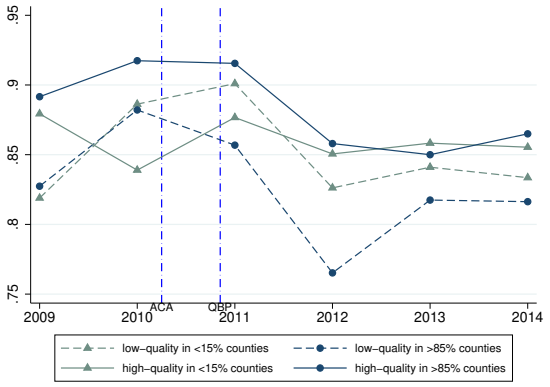
(d) Zero Part D premium, unweighted



(e) Zero drug deductible

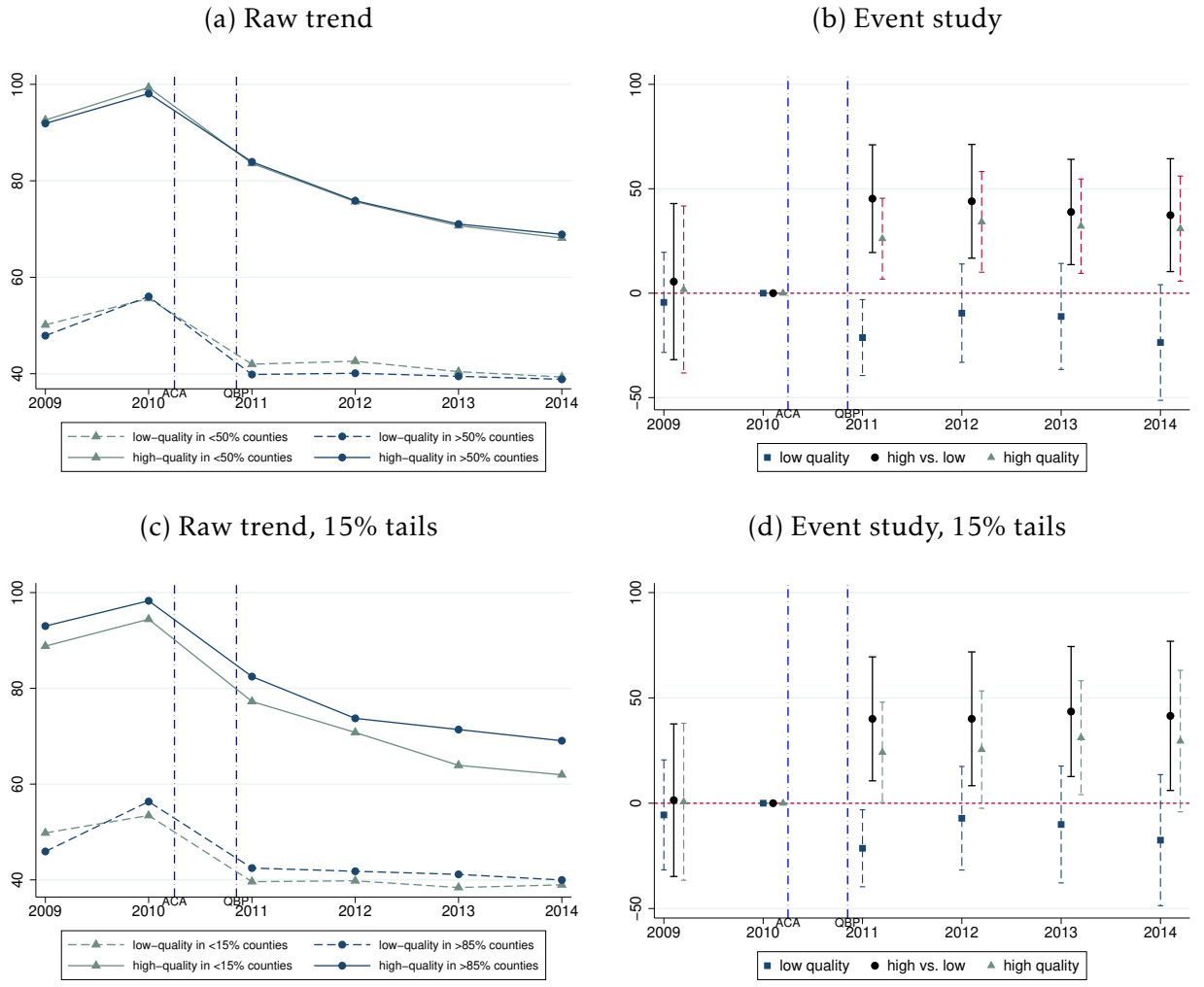


(f) Zero drug deductible, unweighted



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).

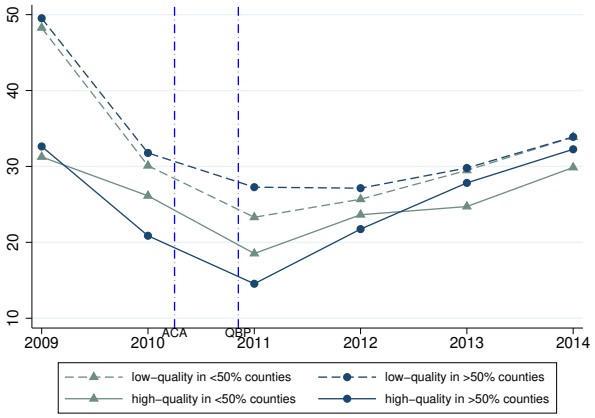
Figure B4: Effect on the total premium (Part C and D), within-contract differences, event study



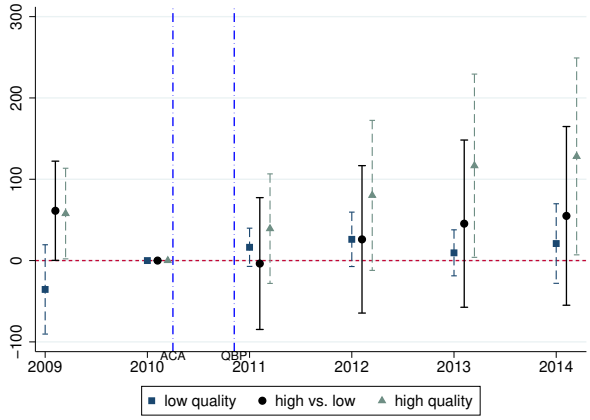
Notes: The figure plots the raw trends of total premiums in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-rated 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 standard errors clustered two-way at the level of counties and contracts.

Figure B5: Effect on drug deductibles, within-contract differences, event study

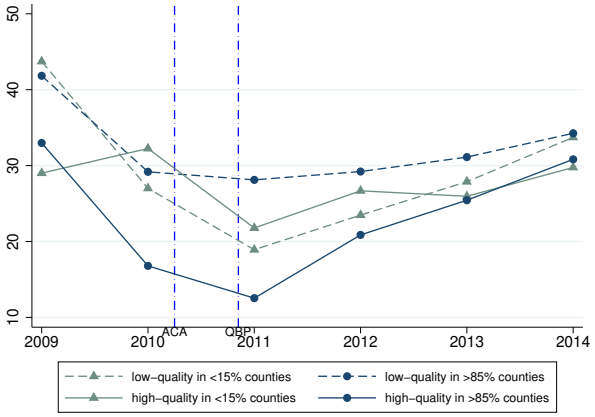
(a) Raw trend



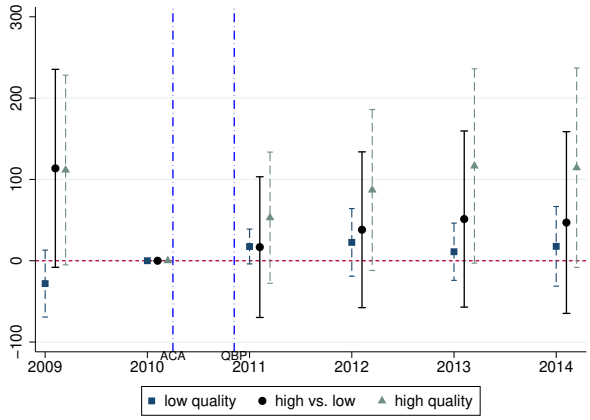
(b) Event study



(c) Raw trend, 15% tails



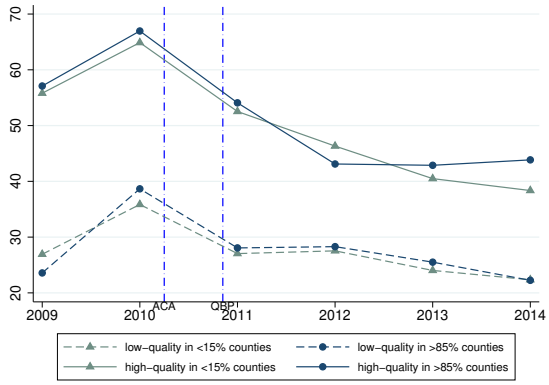
(d) Event study, 15% tails



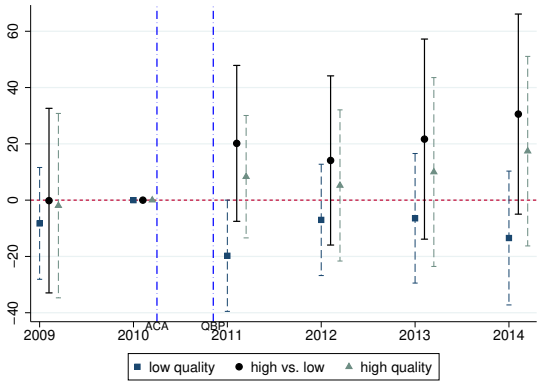
Notes: The figure plots the raw trends of drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. The raw trends in panel (a) plot the price levels above and below the median risk county within an average low-rated 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 standard errors clustered two-way at the level of counties and contracts.

Figure B6: Effect on premiums and drug deductibles, within-contract differences, event study, unweighted by enrollment

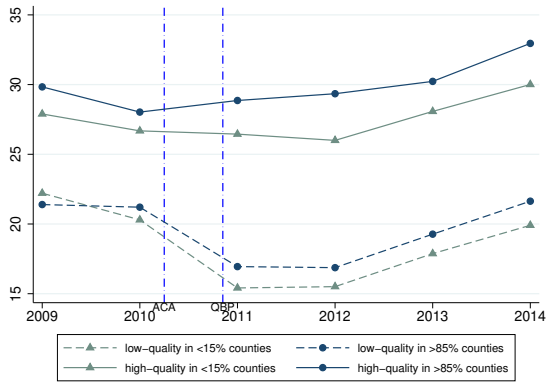
(a) Part C premium, raw trend



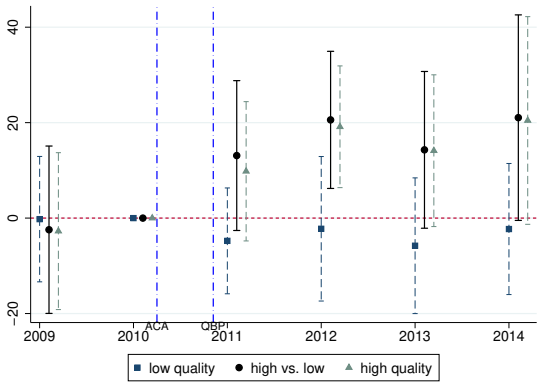
(b) Part C premium, event study



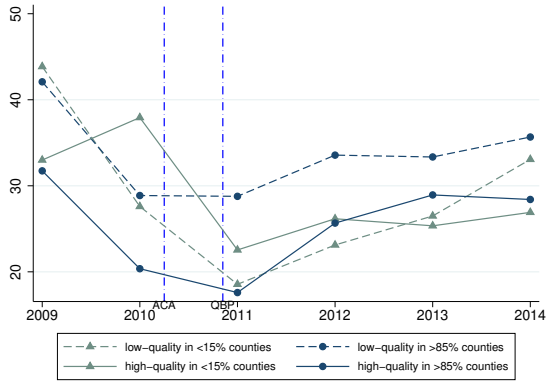
(c) Part D premium, raw trend



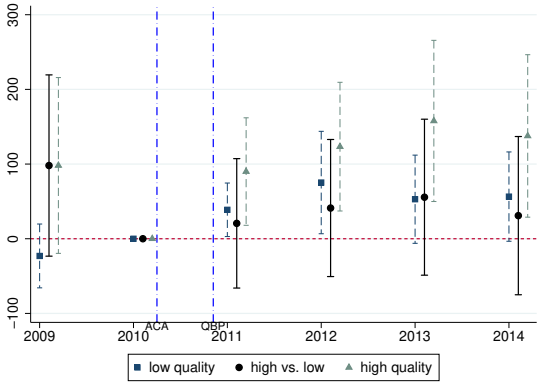
(d) Part D premium, event study



(e) Drug deductible, raw trend



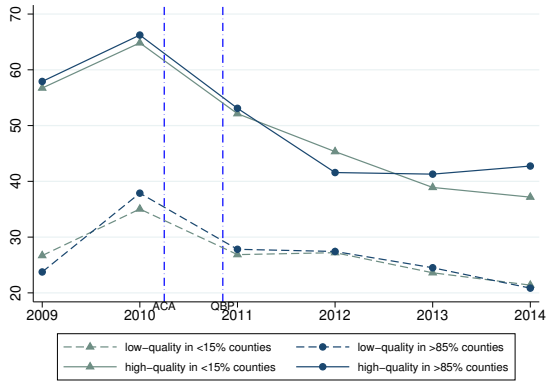
(f) Drug deductible, event study



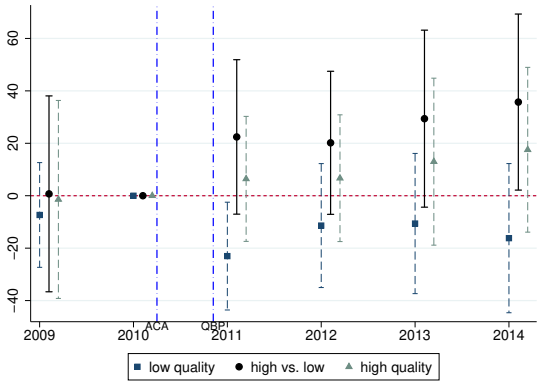
Notes: The figure plots the raw trends of premiums and drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. Different from the main analysis, we aggregate plan prices to the contract-county level taking simple averages. 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 standard errors clustered two-way at the level of counties and contracts.

Figure B7: Effect on median premiums and drug deductibles, within-contract differences, event study

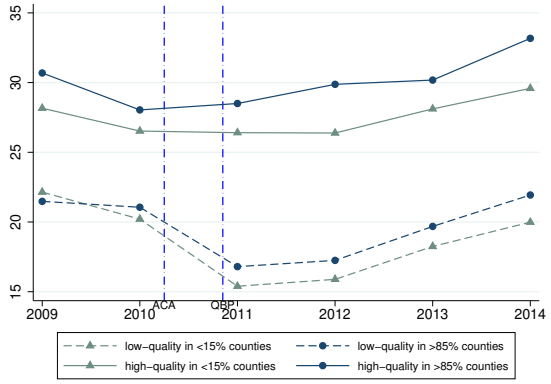
(a) Part C premium, raw trend



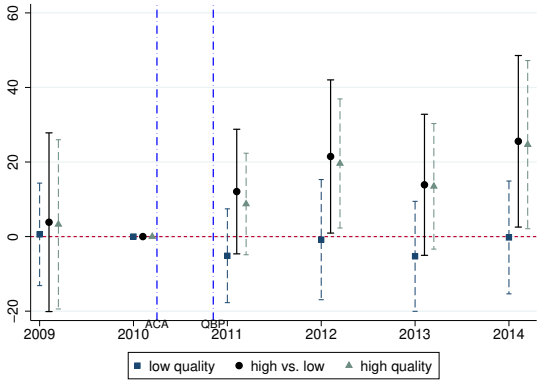
(b) Part C premium, event study



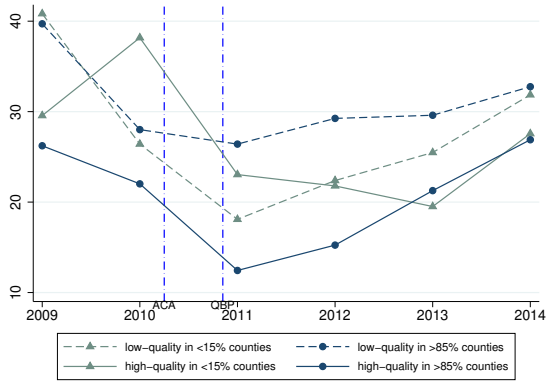
(c) Part D premium, raw trend



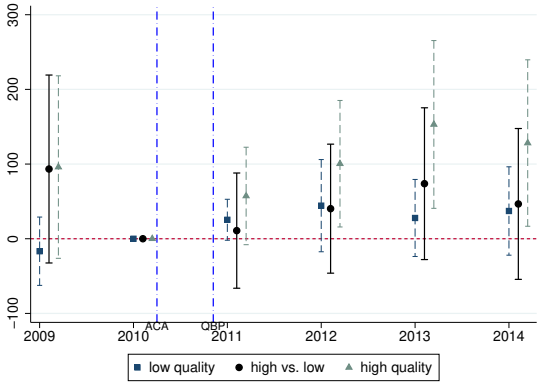
(d) Part D premium, event study



(e) Drug deductible, raw trend



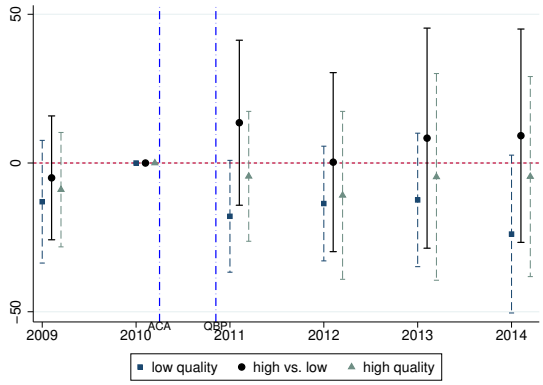
(f) Drug deductible, event study



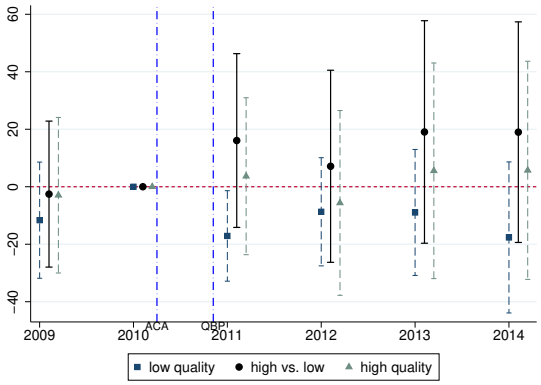
Notes: The figure plots the raw trends of premiums and drug deductibles in the left panels and event study estimates of the within-contract differences over county risk scores in the right panels. Different from the main analysis, we aggregate plan prices to the contract-county level using the median 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 standard errors clustered two-way at the level of counties and contracts.

Figure B8: Effect on premiums and drug deductibles, within-contract differences, event study, deviation to mean

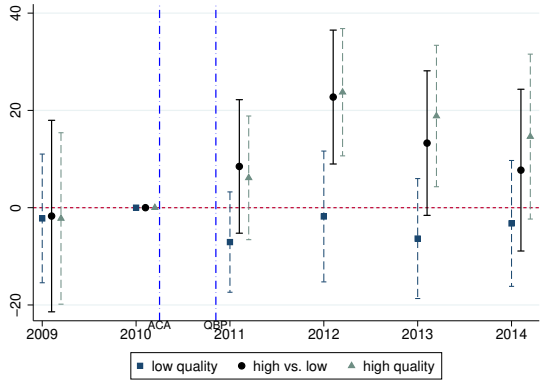
(a) Part C premium



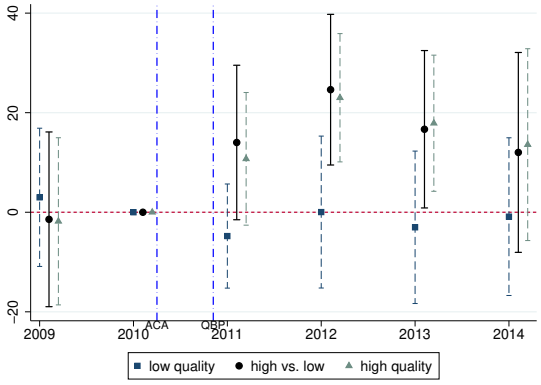
(b) Part C premium, 15% tails



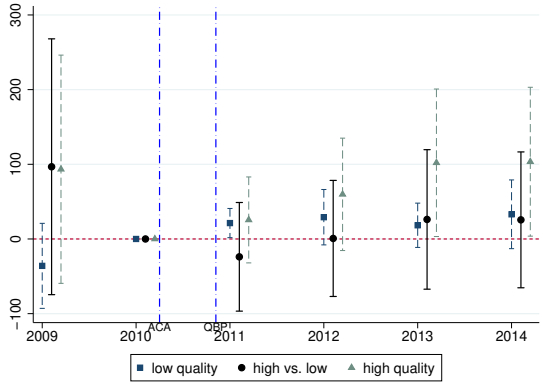
(c) Part D premium



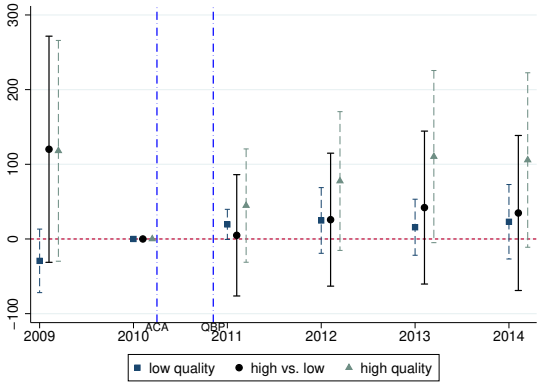
(d) Part D premium, 15% tails



(e) Drug deductible



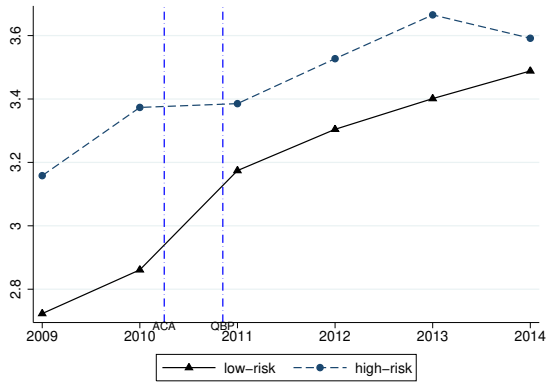
(f) Drug deductible, 15% tails



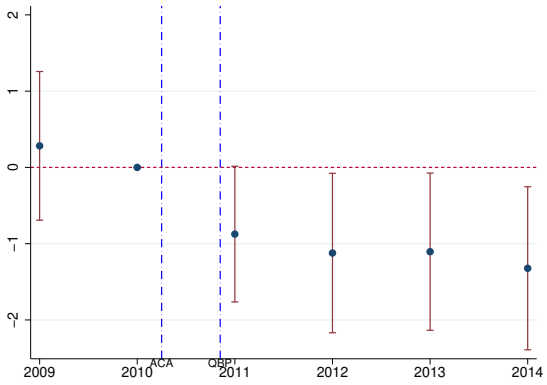
Notes: The figure plots the event study estimates of the within-contract differences over county risk scores. We focus on Part C premiums in panel (a)-(b), Part D premiums in panel (c)-(d), and drug deductibles in panel (e)-(f). County differences in risk scores are measured as the deviation to the mean 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 standard errors clustered two-way at the level of counties and contracts.

Figure B9: Outcome ratings by baseline enrollee risk scores, event study

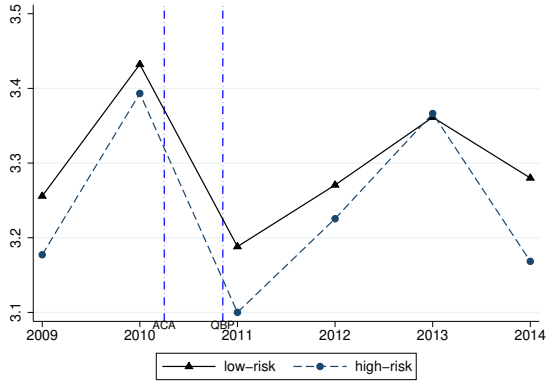
(a) Average outcome rating, raw trend



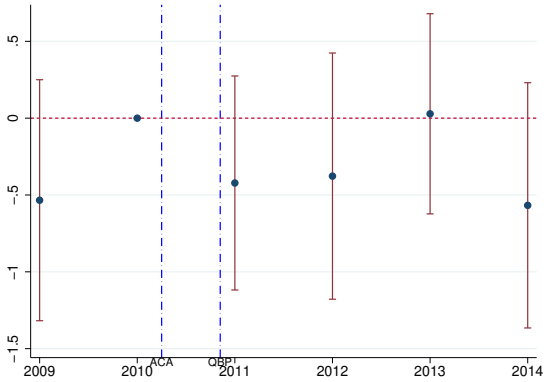
(b) Average outcome rating, event study



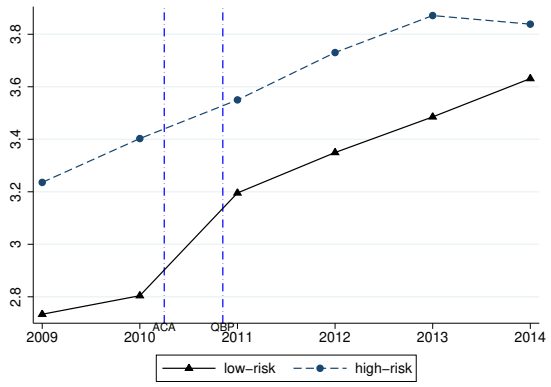
(c) Health improved, raw trend



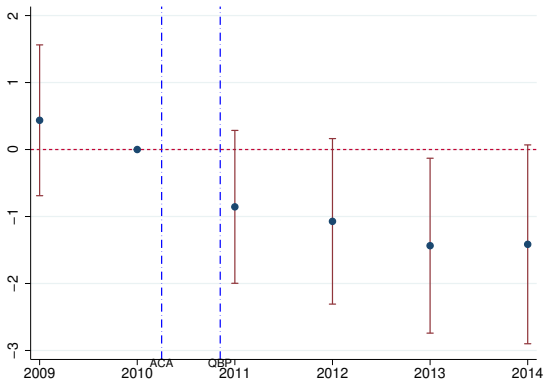
(d) Health improved, event study



(e) Diabetes & blood pressure, raw trend



(f) Diabetes & blood pressure, event study

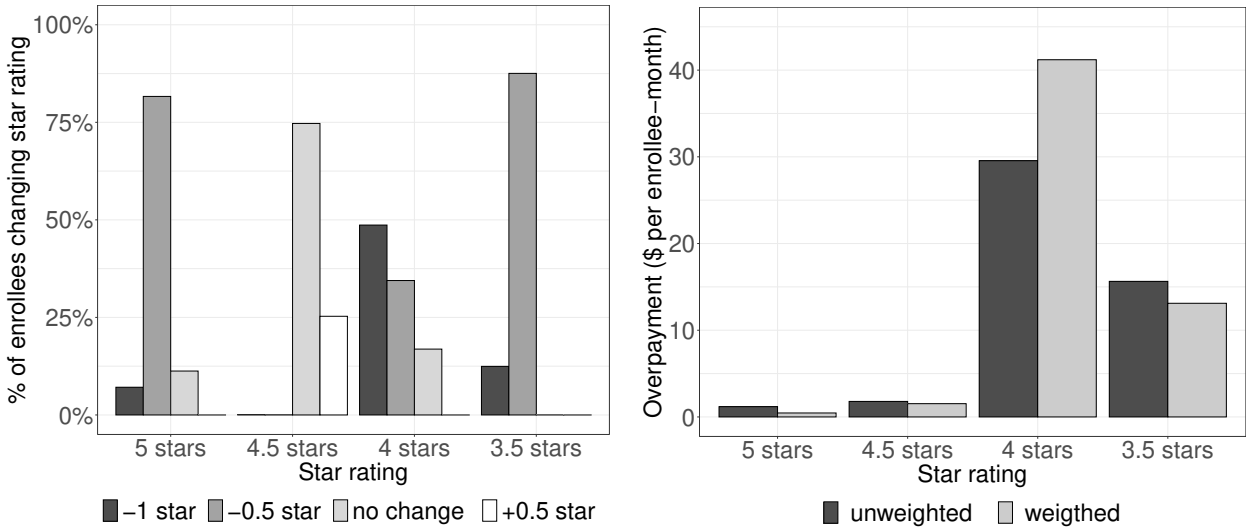


Notes: The figure shows the dynamics of outcome ratings by baseline enrollee risk scores. The raw trends in the left panels plot separate trends for binary groups of contracts above and below the median enrollee risk score (0.97) in the baseline. The right panels show event study estimates from difference-in-differences 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 level of contracts.

Figure B10: Effects of selection on the quality rating and overpayments, synthetic control

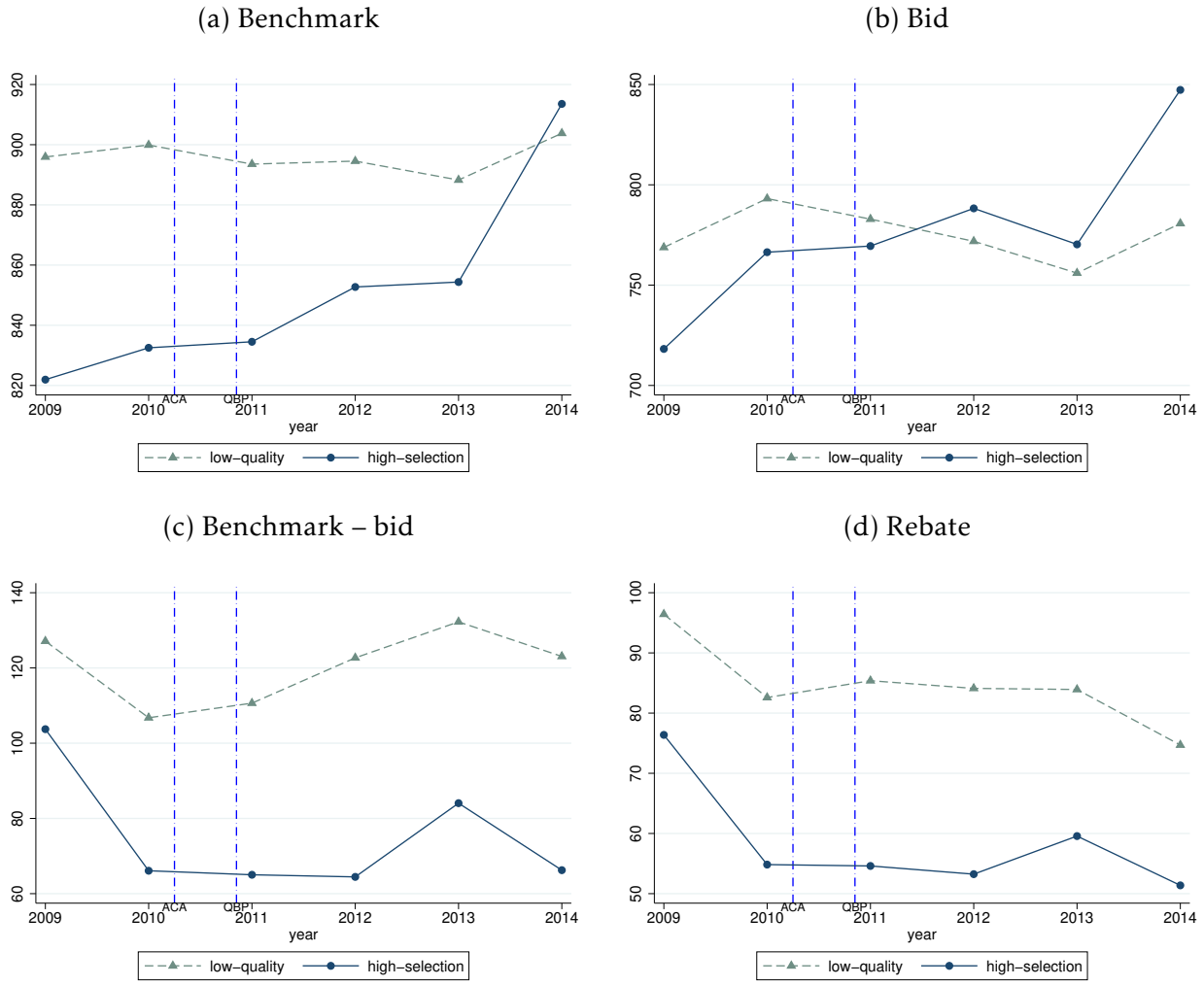
(a) Share of enrollees with star rating change absent selection

(b) Overpayments due to selection



Notes: The figure shows the effect of 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). Panel (a) plots for each overall star rating level in 2014 (horizontal axis) 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. 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). 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. 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. 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 B11: 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.

C Theoretical Appendix

In this section we illustrate in details the model in Section 3. The model shows that relative to the pre-reform prices, premiums decrease more in counties with lower fee-for-service risk score (Γ_l^{FFS}). From the insurer's first order condition (equation 2), we define the price change due to the selection incentive after the payment reform as

$$\Delta p_l = \frac{d\,B}{d\,q} \cdot \frac{\partial q}{\partial r} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_l + s_{-l}}{s_l} \cdot |\varepsilon_l|^{-1}. \tag{C1}$$

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 the two counties. Specifically,

$$\begin{aligned} r &= \frac{s_l \cdot \Gamma_l^{MA} + s_{-l} \cdot \Gamma_{-l}^{MA}}{s_l + s_{-l}} \\ &= \frac{\bar{\Gamma}_l - (1 - s_l) \cdot \Gamma_l^{FFS} + \bar{\Gamma}_{-l} - (1 - s_{-l}) \cdot \Gamma_{-l}^{FFS}}{s_l + s_{-l}}, \end{aligned} \tag{C2}$$

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 C2 therefore expresses contract risk score r in terms of enrollment share s_l and the 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 C2 w.r.t. premium in county l yields

$$\frac{\partial r}{\partial p_l} = \frac{s_l'}{s} \cdot \left(\frac{1 + s_{-l}}{s} \cdot \Gamma_l^{FFS} + \frac{1 - s_{-l}}{s} \cdot \Gamma_{-l}^{FFS} - \frac{\bar{\Gamma}_l + \bar{\Gamma}_{-l}}{s} \right) - \frac{1 - s_l}{s} \cdot \frac{\partial \Gamma_l^{FFS}}{\partial p_l}, \tag{C3}$$

where $s = s_l + s_{-l}$.

The first bracket in equation C3 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 $\frac{\partial \Gamma_l^{FFS}}{\partial p_l}$. We examine premium responses to the level differences in Γ_l^{FFS}, but do not directly exploit the marginal term in the empirical analysis.

Equation C3 states that for a similar enrollment response, county l is more effective at lowering 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 C1). Substituting equation C3 into equation C1 nets out the semi-elasticity

$\varepsilon_l = s_l'/s_l$. The resulting price change Δp_l relative to the pre-reform level is given by

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

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

$$\Delta p_l - \Delta p_{-l} \propto -\frac{dB}{dq}\frac{\partial q}{\partial r}\cdot\left(\Gamma_l^{FFS} - \Gamma_{-l}^{FFS}\right). \tag{C4}$$

Equation C4 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

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

Compared to equation C4, 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 C4 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.

D Data Appendix

D.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 D1 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.[67] 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.[68] Data on

[67] 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.

[68] Less than 1% of the rated contracts in year t have missing star ratings in $t+1$ in the estimation sample.

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.

Table D1: 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.

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 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

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.

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

D.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.[69] 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 D2 summarizes the FFS risk scores and costs by county.

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

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

Table D2: 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.

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.

E Distributional Impacts: Additional Evidence

Following the discussion in Section 8.2, we provide additional evidence on the premium and market share changes in the upper and lower 15% tails of county risk scores. We focus on the risk tails because the within-contract differences implies decreasing (increasing) premiums in the healthiest (riskiest) counties. In the intermediate range, premiums can either increase or decrease depending on the ranking of county risk scores within contracts and the distribution of contracts across counties.

Table E1 estimates the effect on premiums separately for the two risk tails using the specification in equation 13. We expect the effects to be asymmetric since high-rated contracts in the lower risk tail are predominantly high-selection contracts, whereas in the upper risk tail, high-selection contracts account for only 10% of the high-rated sample. Consistent with stronger within-contract differences in 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). In low-rated insurance, premiums did not vary with risk scores in either risk tail.

Table E2 estimates the premium differences over pooled county risk scores across 15% tails in column 1-3, and the effects on market shares in column 4-6. We focus on high-selection contracts in the table. Premiums in high-selection contracts increased significantly with county risk scores, especially in the lower risk tail. Premiums in low-rated insurance did not similarly increase with county risk scores. Market shares in general decreased with county risk score, but decreased much more in high-selection contracts. Moving from the lower to the upper 15% of county risk scores, premiums of high-rated contracts increased by an additional $6.51 after the payment reform, and the market share of high-rated contracts decreased by an additional 12.96 percentage points. Figure E1 shows the event study.

E.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.

E.1.1 Estimation Sample

Figure E2 shows the market share changes when high- and low- rated contracts are defined using contemporaneous ratings. In the lower 15% risk tail (gray lines), high-rated market shares narrowed with and overtook low-rated market shares during the sample period. At the same time, market shares in general decreased in low-rated insurance (dotted

lines), but decreased less in the riskier counties. In high-rated insurance (solid lines), market shares were comparable across risk tails prior to the payment reform but increased differentially in the lower risk tail after the payment reform. Table E3 shows the estimated effects.

E.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. Contracts with at least a 4.0-star rating in the given year are classified as high-rated insurance. We construct market shares of high-rated insurance for county-years with at least one Medicare Advantage contract listed in the Landscape Files.

Table E4 shows that high-rated market share decreased significantly with county risk across the 15% risk tails: a ten percentage point increase in the county risk score lowered high-rated market share by 1.6 percentage points (column 3). We find similar effects across county risk scores in the estimation sample (column 5 of Table E3). The market share differences are concentrated in the lower risk tail (column 1), where a ten percentage point increase in the county risk score decreased high-rated market share by 8.4 percentage points. Moving from the bottom to the top 15% risk tail, high-rated market share decreased by 3.7 percentage points (column 4).

Figure E3 compares the growth of high-rated market share in the lower (gray line) and upper (blue line) risk tail in panel (a). The widening gap across risk tails is driven by accelerating growth rate of high-rated insurance in the healthiest 15% counties. By 2014, high-rated market share in the bottom 15% counties surpassed that in the riskiest counties by as much as 17 percentage points (panel b). Prior to the payment reform, market shares in both risk tails stayed on close and parallel trends. The temporary drop in 2011 was caused by a change in the computation of the quality rating combining the Part C and Part D ratings. 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 is not visible in the estimation sample (Figure E2) and does not affect the main analysis of continuing contracts differentiated by baseline quality ratings.

Table E1: Premium differences over county risk scores, lower and upper 15% risk tails

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			133.51**			37.73
			(55.65)			(32.80)
Risk · Post	63.81	121.40**	-15.87	-25.18	32.15	-17.57
	(119.09)	(54.36)	(85.71)	(15.85)	(58.03)	(9.34)
High · Post			-109.90**			-54.80
			(51.75)			(43.18)
Risk · High			-66.25			-57.18
			(40.44)			(70.02)
Counties		<15% risk			>85% risk	
Contracts	low	high	all	low	high	all
y mean	41.57	98.14	72.39	41.50	73.33	48.12
R^2	0.82	0.82	0.81	0.81	0.76	0.74
N	1,303	1,044	2,322	3,924	1,022	4,935

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

Notes: The table estimates the premium differences over county risk scores in the lower 15% risk tail (<0.87046) in column 1-3, and in the upper 15% risk tail (>1.028) in column 4-6. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table E2: 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 +	all	low	high +	all
Service area risk		<50%			<50%	
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. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table E3: 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 level of counties in the parenthesis.

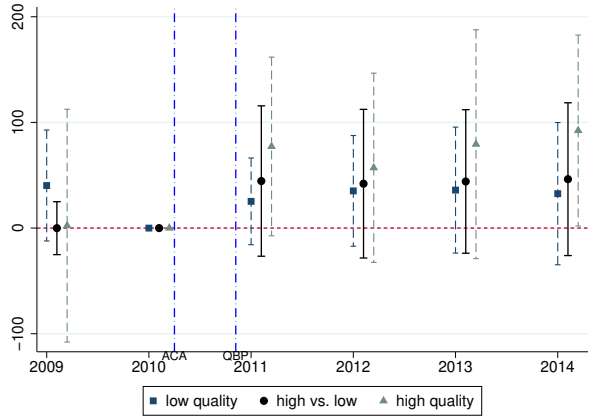
Table E4: Effects on high-rated market shares, contemporaneous quality rating

	(I)	(II)	(III)	(IV)
Risk · Post	-0.84***	-0.071	-0.15***	-0.034***
	(0.27)	(0.12)	(0.048)	(0.013)
Risk	continuous county risk scores			>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

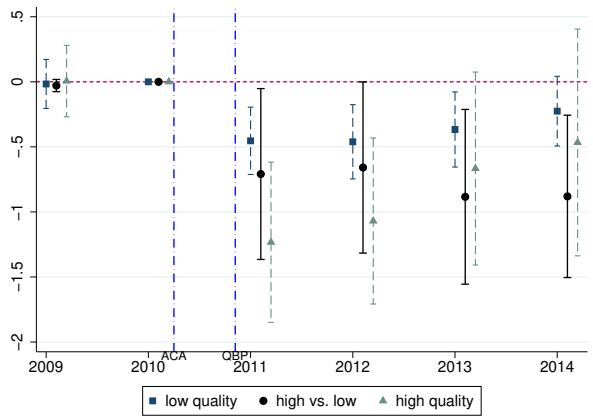
Notes: The table shows the effect of bonus payments on the market share of high-rated insurance across county risk scores. We distinguish across quality ratings using the contemporaneous rating, and classify contracts with a 4.0 star rating and above as high rated. From contract market shares, we construct the market share of high-rated insurance for county-years with at least one MA contract listed in the Landscape Files. We focus on counties in the bottom 15% of county risk scores in column 1, in the top 15% risk scores in column 2, and across the 15% risk tails in column 3. In column 4, the Risk variable is a binary indicator of the top 15% tail, so that the estimate gives the discrete change in high-rated market share when risk scores increase from the bottom to the top 15% risk tail. Robust standard errors clustered at the level of counties in the parenthesis.

Figure E1: Effect on premiums and market shares across the 15% risk tails

(a) Premium



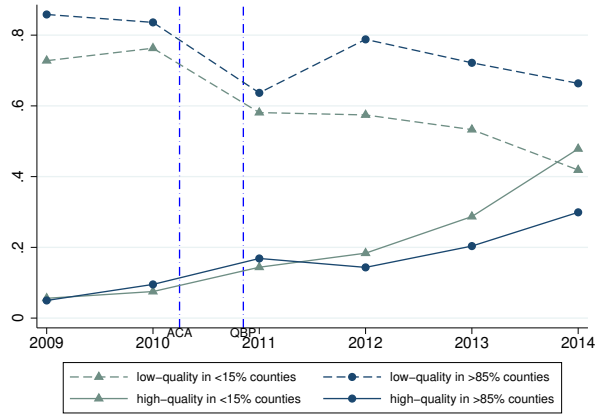
(b) Market Share



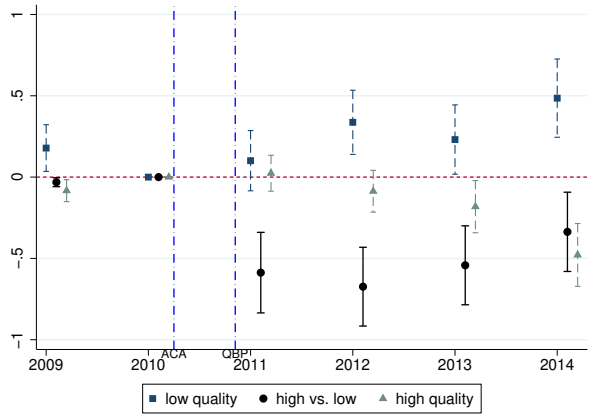
Notes: The figure plots the event study estimates on premiums (panel a) and market shares (panel b) across county risk scores in the 15% risk tails. We restrict high-rated contracts to high-selection contracts below the median service area risk (0.975) of high-rated insurance. 95% confidence intervals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

Figure E2: Effects on market shares, contemporaneous quality rating, 15% risk tails

(a) Raw trend

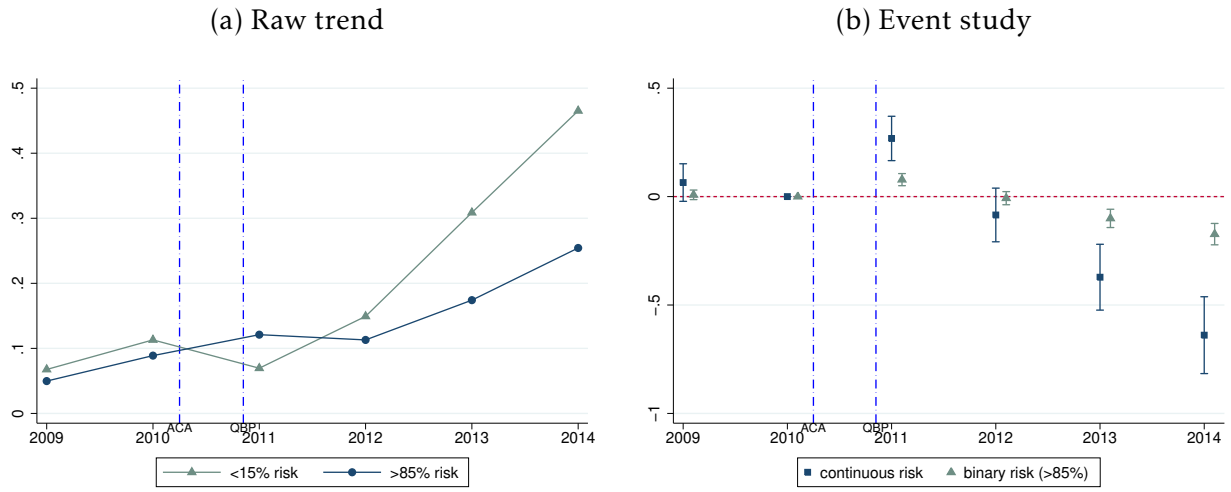


(b) Event study



Notes: The figure plots the raw trends (panel a) and the event study estimates (panel b) of market shares by high and low quality rating based on contemporaneous ratings. Contracts receiving a 4.0 star rating or above are classified as high rating. We construct the quality rating-level market shares in a balanced panel of county-rating-years where counties with masked enrollment data in some but not all years receive zero market shares for missing enrollments. We focus on 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 level of counties.

Figure E3: Effects on high-rated market shares, contemporaneous quality 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. We distinguish across quality ratings using the contemporaneous rating, and classify contracts with a 4.0 star rating and above as high rating. From contract market shares, 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 plot high-rated market shares in the two risk tails. The event study plots the yearly differences in market shares across continuous county risk scores on the left, and the discrete change in market shares when risk scores increase from the bottom to the top 15% risk tail on the right. 95% confidence intervals are plotted based on robust standard errors clustered at the level of counties.

References in the Online Appendix

Amy Finkelstein, Matthew Gentzkow, Peter Hull, and Heidi Williams. Adjusting risk adjustment–accounting for variation in diagnostic intensity. *The New England Journal of Medicine*, 376(7):608, 2017.

Casey Olives, Rebecca Myerson, Ali H Mokdad, Christopher JL Murray, and Stephen S Lim. Prevalence, awareness, treatment, and control of hypertension in United States counties, 2001–2009. *PloS One*, 8(4):e60308, 2013.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS USA: Version 9.0 [dataset], 2019.