

Subsidizing Inequality: Performance Pay and Risk Selection in Medicare

Michele Fioretti, Hongming Wang

▶ **To cite this version:**

Michele Fioretti, Hongming Wang. Subsidizing Inequality: Performance Pay and Risk Selection in Medicare. 2019. hal-03393070

HAL Id: hal-03393070

https://sciencespo.hal.science/hal-03393070

Preprint submitted on 21 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

SciencesPo
DEPARTMENT OF ECONOMICS

Discussion Paper

SUBSIDIZING INEQUALITY: PERFORMANCE PAY AND RISK SELECTION IN MEDICARE

Michele Fioretti and Hongming Wang

SCIENCES PO ECONOMICS DISCUSSION PAPER

No. 2019-15

Subsidizing Inequality: Performance Pay and Risk Selection in Medicare*

Michele Fioretti† **Hongming Wang**‡

November, 2019

Abstract

Pay-for-performance is commonly employed to improve the quality of social services contracted out to firms. We show that insurer responses to pay-for-performance can widen the inequality in accessing social services. Focusing on the U.S. Medicare Advantage market, we find that high-quality insurance contracts responded to quality-linked payments by selecting healthier enrollees with premium differences across counties. The selection is profitable because the quality rating fails to adjust for pre-existing health differences of enrollees. As a result, quality improved mostly due to selection, and the supply of high-quality insurance shifted to the healthiest counties. Revising the quality rating could prevent these unintended consequences.

JEL classifications: I13, I14, L15
Keywords: pay-for-performance, Quality Bonus Payment demonstration, Medicare Advantage, risk selection, supply-side selection, quality ratings, health inequality

*We would like to thank Ghazala Azmat, Aaron Baum, Zach Brown, Moshe Buchinsky, Alice Chen, Golvine De Rochembau, Emeric Henry, Aljoscha Janssen, Bora Kim, Marleen Marra, Daria Pelech, Geert Ridder, Alejandro Robinson-Cortés, Mark Shepard, Andre Veiga, Gianluca Violante and participants of the 2018 ASHECON, the 2019 IIOC and at seminars at the University of Southern California and Sciences Po 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

Economists have long praised the benefits of markets. Competition can lower market prices while improving quality. 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 hospitalization (Gupta, 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, the financial incentives can spur firms to invest in the 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 worsens the quality rating. The selection response can distort the quality rating linked to payments, and have large adverse effects on consumers. In health insurance markets, for example, selection 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). Yet, 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 shows how insurers respond to pay-for-performance by exploiting the introduction of the Quality Bonus Payment demonstration (henceforth QBP) in the U.S. Medicare program, which is transitioning to pay-for-performance (Burwell, 2015).[1] QBP tied insurers' subsidies to the quality of their services.[2] The same quality measures used to rank insurance contracts were already available to enrollees before QBP. Therefore, in this setting, we can credibly identify insurers' responses to pay-for-performance because QBP shifted the supply of insurance without directly affecting demand.

We apply this intuition empirically and find that insurers with high-quality ratings served less risky enrollees after QBP. These insurance contracts selected healthier enrollees by lowering premiums in healthier, low-risk counties and by raising it in riskier ones. The selection incentive originates from the design of the quality rating which fails to adjust for

[1] Medicare is the program providing health insurance to over 49 million elderly Americans. The program currently costs the U.S. government $705 bn annually, or 20% of total health expenditure (CMS, 2016).

[2] The Congressional Budget Office (CBO) estimates that QBP will cost the federal government $8.35 billion over ten years, more than all other Medicare demonstrations combined since 1995 (CBO, 2012).

pre-existing health differences related to chronic diseases like diabetes and hypertension. As a result, over 70% of the health improvements after QBP is due to risk selection rather than value-added. The selection response substantially inflated the subsidies to selecting insurers, and shifted the distribution of good quality insurance to the healthiest counties, showcasing the challenges of designing quality-based payment incentives to insurers.

Our empirical strategy exploits two features of the payment reform. First, QBP introduced variation in bonus payments based on contracts' quality ratings. In particular, we distinguish between contracts of high and low quality according to an institutional feature of the policy. Second, although contracts already received quality ratings before the QBP demonstration, quality was not a factor in the payment model. Therefore in a difference-in-differences framework, we estimate intent-to-treat effects on contracts differentiated by baseline quality ratings. To explore the mechanisms of selection, we extend the framework to examine premium differences across counties with different levels of risk for potential enrollees. We also quantify the impact of selection on the quality rating using instrumental variables derived from our findings.

We find a substantial improvement in the risk pool of high-quality contracts, as measured by risk scores. The decrease in risk scores is not associated with other differential effects on the average price levels or service area characteristics between high- and low-quality contracts. Risk scores decreased even more in high-quality contracts already enrolling low-risk individuals in the baseline – in these *"high-selection"* contracts, risk scores dropped by 4-8 percentage points. This suggests that QBP incentivized insurers to select enrollees based on risk. We next ask both *how* firms screened 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 significantly with county risk scores in high-quality contracts, but not in low-quality contracts. For every ten percentage point increase in the county risk score, drug premiums increased by $1.70 per month in high-quality contracts after the reform, or 7% above the mean. We show that premiums responded directly to the health of enrollee across counties, but not to local socio-economic factors, market concentration, or the cost and quality of healthcare, confirming that demand and supply factors external to risk are not driving the premium differences over risk scores. We therefore conclude that high-quality contracts selected healthier enrollees with premium differences across counties.

To understand why the payment demonstration incentivized the selection of healthier individuals, the second half of the paper delves into the design details of the quality rating. In a stylized model, we show that a negative correlation between risk and quality makes healthier individuals more profitable to insurers under quality-adjusted payments. Em-

pirically, we document a negative correlation between risk scores and the health outcome measures most weighted in the quality rating. These measures rank contracts based on health improvements over time, but fail to adjust for pre-existing health differences at the time of enrollment. As such, healthier enrollees are associated with better ratings, and contracts with greater improvements in the risk pool also experienced greater relative gains in the health outcome rating. These results suggest a role of risk selection in the gaming of quality ratings and payments.

We correct for selected risk scores in the measurement of quality using an instrumental variable strategy, and calculate the financial gains of selection to insurers. We allow the quality rating to depend on the risk composition of contracts, and decompose the rating gains after QBP into a selection component and a value-added component. Building on our analysis of the selection mechanism, we instrument the risk composition of contracts using premium differences across counties. We confirm that the health outcome ratings were subject to substantial gaming by insurers, with selection accounting for 77% of the rating improvements in high-selection contracts. Removing the selection gains would downgrade the overall rating of most high-selection contracts by 0.5 to 1 star in 2014 (out of 5 stars). Selection costs Medicare $51.7 million in 2014, or 13% of the bonus payments.

The selection response has large distributional impacts on enrollees. Since average premiums and enrollee benefits did not differ by quality, premium differences to select healthier enrollees led to subsidies from high-risk counties to low-risk counties. This allowed high-quality insurers to fully capture the selection gains as profits. As a result, market shares of high-quality contracts polarized across county risk scores. In the lowest-risk counties, high-selection contracts lowered monthly premiums by over $13 per ten percentage point decrease in the risk score. Moving from the bottom to the top 15% of county risk scores, high-quality contracts increased monthly premiums by $6.50, and decreased market shares by almost 13 percentage points. Therefore selection shifted the supply of high-quality insurance away from the riskiest counties, worsening the regional disparities in healthcare access in the U.S.

Our analysis identifies several aspects of the rating design which contributed to the selection response of insurers. To contain the selection incentive, we suggest risk-adjusting or down-weighting outcome measures in the quality rating. Since no risk-adjustment guideline exists for the outcome measures, we propose one approach that adjusts for selected risk scores using some knowledge of the selection mechanism by firms. Besides, we suggest designing quality ratings at the level of small geographic markets rather than contract service areas – the latter gives rise to the cross-space selection response in the current setting. For contracts covering many markets, a localized rating is also more

informative to enrollees.

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.

This paper is also related 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 profitability created by the adjustment formula (Brown *et al.* 2014, Geruso *et al.* 2019, Carey 2017, Lavetti and Simon 2018). This paper suggests that the residual variation is a small price to pay relative to the systematic variation in profitability favoring healthier enrollees in the absence of risk adjustments on quality. The scale of the selection response and the aggregate effects illustrates the critical role of risk adjustment for the functioning of health insurance markets.[3]

More broadly, 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 private insurers. This finding reveals that supply-side regulations play important roles in shaping the regional disparities of healthcare, with particular implications for social insurance programs contracted out to private entities.[4] Insurer gaming in these programs not only leads to the capture of subsidies (Cabral *et al.*, 2018, Duggan *et al.*, 2016, Curto *et al.*, 2014) and price distortions (Decarolis, 2015, Decarolis *et al.*, 2015), but has large unintended consequences on the distribution of benefits across consumer demographics and geography (Polyakova and Ryan, 2019).

Taken together, our results cast doubts on the effectiveness of pay-for-performance in

[3]Risk adjustments can also trigger the strategic coding of risk scores above the actuarial risks of enrollees (Geruso and Layton, 2018). We consider potential biases from coding intensity when analyzing the selection responses of insurers across space.

[4]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). Exploiting a supply shock, we identify payment incentives as one particular supply-side channel that drives disparities, with direct implications for policy.

health insurance markets. Intended to promote high-value services while cutting costs, quality bonus payments in Medicare led to only modest improvements in health outcomes. Insurers exploited the design of the quality rating by selecting healthier enrollees and by subsidizing premiums in low-risk counties with higher prices in riskier counties. The selection resulted in large overpayments to insurers, and shifted the distribution of high-quality insurance to the healthiest counties. In the rest of the paper we examine the efficiency and equity implications of the subsidy design, and suggest ways to improve the quality rating based on our analyses.

2 Quality Ratings and Payments in Medicare Advantage

When looking for health insurance, all Americans 65 years of age or older can choose between receiving the traditional, or public, Medicare coverage or purchase additional coverage privately through Medicare Advantage (MA). Since the inception of the MA program in the 1980s, MA enrollment has grown to 33% of the Medicare population in 2017, with over 2,000 MA plans providing Medicare insurance to 18.9 million beneficiaries. Also referred to as the "Part C" coverage, MA contracts offer at least the same benefits as the traditional Medicare,[5] and typically offer more generous benefits and supplemental coverages. For example, the vast majority of MA contracts offer prescription drug coverage (called "Part D"),[6] and have lower enrollee cost sharing than the traditional Medicare.

Private insurers contract with the Centers for Medicare and Medicaid Services (CMS) to offer MA insurance plans. To assist beneficiaries in making informed plan choices, CMS provides quality ratings of MA contracts over a wide range of performance measures. In the early 2000s, ratings of selected measures were included in the "Medicare and You" handbook mailed to Medicare eligibles. In 2007, CMS adopted a 5-star rating metric of quality. In 2009, CMS started publishing overall star ratings in addition to ratings of individual measures. The overall ratings were computed from measure ratings for Part C and Part D coverage separately. In 2011, CMS instituted an overall star rating combining the Part C and D measures, which became the basis of bonus payments from 2012 onward.

Ratings of individual measures are revised annually according to the percentile ranks of contracts. Although MA contracts have multiple plan offerings, all plans share the same quality rating at the level of the contract. For most measures, the performance bar of

[5] Medicare Part A covers hospital costs. Part B covers outpatient and other medical costs not covered in Part A. In Medicare Advantage plans, Part A and Part B benefits are offered under the Part C coverage.

[6] Prescription drug coverage is available to FFS enrollees through a stand-alone health plan in the Part D program. Almost all MA contracts offer at least one MA-PD plan with integrated Part D coverage.

a given star rating increases over years.[7] The underlying information of quality comes from a variety of administrative and survey records, collected from different samples over multiple periods. For instance, health outcome measures in the year-t rating are based on clinical records of enrollees in year $t-2$, whereas patient complaints measures are based on complaint resolution in the first half of year $t-1$. Appendix Table A39 and A40 list the data source and the sample period of all measures in the 2013 star rating.

Individual ratings are aggregated to the overall star rating through a weighting procedure. All individual measures received equal weights in 2009-11. Starting in 2012, health outcome measures receive 3.0 weights in the overall rating, access and patient satisfaction measures receive 1.5 weights, and process measures concerning screening and primary care visits receive 1.0 weights.[8] A small bonus is added to contracts with low variances across the measure ratings. The weighted average plus the bonus is rounded to the nearest star rating shown to enrollees on a scale from 2.0 to 5.0 stars at half-star increments.

Although quality disclosure in principle improves choice quality and market functioning,[9] previous research in the Medicare context has found modest demand side responses to quality ratings (Dafny and Dranove 2008, Reid *et al.* 2013, Li and Doshi 2016). Using the discontinuity in the overall rating near rounding cut-offs, Darden and McCarthy (2015) finds reduced enrollments in low-rated contracts below 3.0 stars in 2009, but no effect in 2010. The null effect in the second year of the rating disclosure is potentially due to supply side pricing responses (McCarthy and Darden, 2017), or due to consumer inattention to quality information after the policy phase-in.[10]

2.1 The Quality Bonus Payment Demonstration

Payments to MA plans are determined by comparing the plan's asking price (bid) with the benchmark set by CMS. The bid (denoted b) reflects the projected cost of an average enrollee in the plan plus an administrative load. The bid is evaluated against a plan-

[7] The measure 'Access to Primary Care Doctor", for example, is dropped in 2013 because nearly all MA plans met high-quality standards (above 85% for 4.0 star and 95% for 5.0 star), and the measure "Plan All-Cause (30-day) Re-Admission" revised the threshold for 5.0 rating from below 5% in 2012 to below 3% in 2013; only a handful of local coordinated care plans (CCP) with very small enrollments ever obtained 5.0 rating on this measure.

[8] CMS also publishes domain-level ratings where individual measures are grouped into 5 Part C domains and 4 Part D domains. For our analysis, we mainly exploit the weight change which differentially varied the financial return to improving measure ratings across the three weighting categories. The weights of measures in the 2013 rating are listed in Appendix Table A39 and A40.

[9] In particular, Glazer and McGuire (2006) show theoretically that a functioning quality reporting system is as important as risk adjustment in correcting market inefficiencies.

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

specific benchmark (B) constructed from county benchmarks weighted by enrollment. For plans bidding below the benchmark, payment equals the plan's bid and a rebate. The rebate is passed on to enrollees as premium discounts or additional benefits. For plans bidding above the benchmark, payment from CMS is capped at the benchmark, and the difference is charged to enrollees as extra premiums. There is no rebate or additional benefits to enrollees.

In an effort to link payments to the quality of care, the Affordable Care Act (ACA) – signed into law in March, 2010 – introduced quality bonus payments to contracts rated 4.0 stars and above. The ACA payment model was scheduled to take effect in 2012. On November 10th, 2010, CMS proposed the Quality Bonus Payment (QBP) demonstration which increased bonus payments above the ACA levels and extended bonus payments to lower-rated contracts below 4.0 stars. The demonstration covers the period 2012-2014. Payments fully transitioned to the ACA model in 2015.[11]

The QBP rebate to consumers is determined through the following formula

$$rebate = \begin{cases} 0 & \text{if } b \geq \alpha^{star} \cdot B \\ \tau^{star} \cdot (\alpha^{star} \cdot B - b) & \text{if } b < \alpha^{star} \cdot B, \end{cases} \tag{1}$$

where the star rating affects rebate through bonuses on the benchmark α^{star} and the rebate percent τ^{star}. Prior to QBP, all contracts are subject to the same benchmark in a county, and 75% of the bid below the benchmark is rebated to enrollees. QBP varied both parameters with the star rating, awarding bonus benchmarks and higher rebate percentages to high-quality contracts. Table 1 summarizes the policy change.

To a large extent, the generous bonus payments in QBP offset the payment cuts implemented by the ACA. Starting in 2012, ACA revised the county benchmarks to more closely match the spending levels in the fee-for-service Medicare.[12] The bonus benchmarks more than offset the base rate cut for median- to high-rated contracts.[13] Similarly, the rebate percent is reduced for all contracts, but adjusted higher for higher-rated contracts.

Other than the standard bonus payments awarded at the contract level, high-quality contracts in "double bonus counties" are eligible for additional bonuses on the bench-

[11] The bonus payments to lower-rated contracts (below 4.0 stars) are discontinued after the transition.

[12] The new benchmarks ranged from 95% of fee-for-service spending in counties in the top quartile to 115% in those in the lowest quartile.

[13] Between 2010 and 2012, the mean (median) base benchmark across counties remained stable at $775 ($750). Bonus-adjusted benchmark for 3-star contracts increased to $800 ($780) in 2012, and increased to $805 ($785) for 4-star contracts. At the contract level, although the raw benchmark trended similarly by quality, bonus-adjusted benchmarks increased more in high-quality contracts.

mark.[14] Layton and Ryan (2015) finds that the double benchmark bonuses are associated with increased plan offerings but did not improve the average star ratings in these counties. We control for location-specific benchmarks and benchmark bonuses in the cross-county analyses of insurer responses.

3 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 premiums and drug deductibles across service areas (counties) covered by each plan. We drop Regional Preferred Provider Organization (PPO) plans and contracts with missing star ratings for payment purposes since different payment rules apply in these cases. We further restrict the sample to plans covering both medical and prescription drug expenditures, or the MA-PD plans. Details of the sample construction are available in Appendix C.

Table 2 summarizes the estimation sample. Panel A looks at contract-year observations, while Panel B expands the contract-year observations by the locations (counties) in the contract's service area. An average MA-PD contract offers 3.4 plans covering over 25 counties in the service area, consists of about 334.8 thousands enrollees yearly. Most MA-PD contracts place bids below their benchmark by $110.7, earning its enrollees about $ 78.4 in rebates per month. Finally, a large number of contracts offer zero premiums and zero drug deductibles.

In the remaining sections of the paper we will look at how contracts of different quality responded to QBP. Our main focus is on risk selection. We first show that contracts with greater quality improved their risk pool more than other contracts. Second, we identify the mechanism used by contracts to risk-select enrollees: varying premiums across counties to attract more enrollments from healthy individuals. We refer back to Table 2 when we analyze the selection response by insurers. Third, we turn to explaining why QBP incentivized contracts to select healthier enrollees. We show that health outcome measures in the quality rating fail to adjust for pre-existing health conditions at the time of enrollment, making healthier enrollees more profitable to insurers with quality-linked payments. We quantify the selection gains on the quality ratings and the bonus payments using an instrumental variable strategy. We discuss the distributional implications of

[14] Benchmark bonuses to 5-star contracts in these counties exceed the standard 5%. The additional bonus is applied to all contracts in the county. Double bonus counties tend to be urban counties with low fee-for-service spending and high Medicare Advantage penetration rates.

selection on health inequality and healthcare access, and suggest ways to combat the adverse effects through improved design of the quality rating.

4 QBP and Risk Selection

The Quality Bonus Payment demonstration introduced more generous payments to insurers with higher quality ratings. In this section we exploit the payment differential by quality to empirically examine how insurers responded to the policy with respect to changes to their risk pools and pricing strategies.

We formalize our analysis using a standard difference-in-difference framework. We distinguish two groups of contracts, high- and low-quality, and two periods, before and after 2011. The difference over the time horizon captures the base rate cut which affects all contracts regardless of quality. The second difference captures the greater bonus payments to high-quality contracts in the post-reform period. Despite QBP became effective only from 2012, the payment incentives can already affect insurer pricing in 2011. Because the payment model was signed into law as part of the ACA in March, 2010, insurers had until June, 2010 to finalize the bid and benefit design for enrollment year 2011. We hence allow for anticipatory effects in the first post-reform year (2011) in our analysis.

We define high- and low-quality contracts using star ratings in the 2009-2010 baseline. Since bonus payments were initially restricted to contracts rated 4.0 stars and above in the ACA model, extended to contracts below 4.0 stars by the QBP demonstration, we use the 4.0 star cut-off to define contract quality. We refer to contracts rated 4.0 stars and above in 2009 and 2010 as high-quality contracts, and those rated below 4.0 stars but no less than 3.0 stars in 2009-2010 as low-quality contracts. Lower-rated contracts below 3.0 stars are subject to suspension if the star rating does not improve in two years. We do not include these contracts in the low-quality group.

Table 2 summarizes the estimation sample. We aggregate plan-level characteristics to the contract level using plan enrollment weights, and show separate statistics for high- and low-quality contracts in our sample. Panel A looks at contract-year observations used in the difference-in-differences analysis. High-quality contracts tend to serve lower-risk counties, bid closer to the benchmark, and provide smaller rebates to enrollees. They also charge higher premiums and are less likely to offer zero-premium plans. Differences in the drug deductibles are not statistically significant.

We estimate the following contract-level difference-in-differences model,

$$y_{ct} = \beta \cdot high_c \cdot post_t + \alpha_c + \tau_t + \epsilon_{ct}, \tag{2}$$

where the main outcome of interest y_{ct} is the risk score of contract c in year t, and $high$ and $post$ are indicator variables for high-quality contracts (rated 4.0 stars or above in 2009-2010) and the post-QBP years (2011 and after), respectively. In our preferred specification, we include contract fixed effects α_c to control for unobserved heterogeneity across contracts. To the extent that time-varying quality rating can affect the contract's selection and pricing strategy, we also control for contemporaneous star rating in the robustness analysis. This has little effects on the point estimates.

Although we specify baseline high-quality contracts as the treated group, the differential effect could be driven by either the low- or high-quality contracts. The specification itself does not imply that high-quality contracts are the drivers behind the estimated effects. To empirically investigate which contracts responded to the reform and how, in subsequent analyses we extend the basic specification to examine selection mechanism across geographic locations. The results make clear that the effect on risk score, our main outcome variable, is not driven by the low-quality controls. Equation 2 therefore serves as the natural starting point of our analysis.

4.1 Risk Scores Improved More for High-Quality Contracts

We first estimate equation 2 where the dependent variable is a contract's risk score. Table 3 shows the results. As an insurance contract is composed of multiple plans, we obtain the contract risk scores by aggregating the plan risk scores to the contract level using enrollment weights in column 1-2, and using equal weights in column 3-4. We also use the plan risk scores as the outcome variable in column 5-6. The odd-numbered columns control for contract or plan fixed effects. Across specifications, risk score decreased significantly in high-quality contracts. The preferred specification in column 1 shows a 2.6 percentage point decrease in the high-quality risk score, with larger effects above 4 percentage points in the cross section.[15]

Figure 1 shows the raw trends and the event study. Before the payment demonstration became law in 2010, risk scores in both quality stayed on parallel trends. In 2011, high-quality risk scores decreased below the low-quality counterparts, and remained so throughout the demonstration period. In the event study, risk scores decreased more in high-quality contracts after 2011, and the difference grew larger in 2012-2014.

[15] In fixed-effect models, contemporaneous star ratings have no effect on risk scores (Appendix Table A1). This is because high-quality contracts remain higher rated after the QBP (column 1-2), and the change in risk score is not driven by temporal shocks to quality. The result is also not driven by a handful of 5-star contracts accepting enrollees year round since 2012 (Decarolis and Guglielmo, 2017). Dropping 5-star contract-years has little effects on the estimates (Appendix Table A2).

The effect on risk scores is driven by high-quality plans. To see this, Figure 2 plots the kernel density of risk scores in high- and low-quality contracts, before (2009-2010) and after the QBP (2011-2014).[16] Risk scores decreased in high-quality contracts particularly in the lower deciles of the distribution. The density shift in low-quality contracts is negligible. Figure 3 confirms this result by plotting the distributional effects on deciles of risk scores, showing quantile difference-in-differences estimates in panel (a) and change-in-changes estimates (Athey and Imbens, 2006) in panel (b). Both sets of estimates suggest that the decrease in high-quality risk scores is concentrated in the 20%-40% deciles, where risk scores decreased by 4-8 percentage points, or 4%-9% below the baseline levels (Appendix Table A3).

Heterogeneous Effects. We also examine the heterogeneous effects by contracts with different exposure to risk types in the service area. Since Medicare beneficiaries who did not purchase private plans remain in the fee-for-service (FFS) program, MA contracts in counties with lower FFS risk scores are exposed to lower-risk enrollees. To explore the implications for the risk composition change in high-quality contracts, we construct service area risk at the contract level by averaging baseline FFS risk scores across the contract's service area.

Table 4 estimates the effects on the risk scores of high-quality contracts divided by service area risks. We find larger effects on high-quality contracts below the median service area risk (0.975), where risk scores decreased by 3.7 percentage points (column 1). The effect is small and insignificant above the median risk (column 2). In column 3-4, risk scores fell by 4.3 percentage points for high-quality contracts in the lower 25% of service area risk, but only slightly in the upper 25%. Figure 4 illustrates the results. Risk scores trended similarly between low-quality contracts and high-quality contracts serving riskier counties, but fell sharply in high-quality contracts serving lower-risk counties.[17]

The overall decrease in high-quality risk scores is therefore concentrated in what we call the *high-selection* contracts – high-quality contracts with below-median service area risks in the baseline. This heterogeneous effect suggests that exposure to low-risk enrollees in the service area potentially contributes to the risk composition gain at the contract level. The next subsection explores differential entry to low-risk counties as a potential mechanism of selection.

[16] We show yearly density plot in Appendix Figure B1, and yearly plan risk score distribution in Appendix Figure B2.

[17] The heterogeneous response by service area risk also applies to a smaller set of high-quality contracts rated 4.5 stars and above in the baseline (Appendix Table A4).

4.2 Market Characteristics do not Explain Risk Selection

One mechanism of the effect on risk scores is adjustments in the service areas of high-quality contracts. It is possible that high-quality contracts selectively entered counties with favorable characteristics, and the makeup of service areas affected the risk types enrolled in the contract. In this case, we expect contract-level market characteristics to partially explain the risk score differences between high- and low-quality contracts.

To detect adjustments in the service area, we replace yearly county characteristics with values in 2012, and average across the service area to generate contract-year observations of market characteristics. The resulting variable captures changes in the composition of a service area along measured characteristics, not the temporal variation in these characteristics. For instance, a decrease in the service area FFS risk score would indicate increased entry in low-risk counties or exits from high-risk counties, depending on the change in the market size.

Table 5 shows little change in the market characteristics. Basic compositions such as the number of counties and plans in the service area remained stable between high- and low-quality contracts. Important for the effect on risk scores, service area risk scores did not decrease in high-quality contracts after the payment reform (column 2). Although risk scores decreased more in contracts serving low-risk counties, differential entry into low-risk counties does not explain the risk composition gain of high-quality contracts.

Since the QBP demonstration also varied county benchmarks and awarded additional benchmark bonuses in the double bonus counties, we also check if contracts differentially selected into high-benchmark counties (column 3) or double-bonus counties (column 4).[18] We see no evidence of differential selection by high-quality contracts along these margins.

4.3 Pricing After QBP: Bids, Rebates, Premiums and Deductibles

We next investigate whether price differences between high- and low-quality contracts can explain the decrease in high-quality risk scores. On one hand, bonuses on rebate percentages tend to put downward pressure on the premiums of high-quality contracts. On the other hand, the benchmark bonuses allow high-quality contracts to place higher bids without affecting the rebate to enrollees (see equation 1). The ultimate pass-through of bonus payments to prices can affect the risk pool of high-quality contracts.

We analyze these aspects of the pricing response in Table 6, examining benchmarks,

[18] We define double bonus counties as having a 60% top-off above the standard bonus rate on benchmarks: 5-star contracts receive at least a 8% bonus in these counties, or 60% above the standard 5% bonus.

bids or rebates as the dependent variable in equation 2.[19] In column 1, benchmarks increased mechanically due to the benchmark bonuses to high-quality contracts.[20] Column 2 shows that high-quality contracts increased the bids by more than the benchmark bonuses, resulting in a narrowing of the benchmark-bid gap by $9.17 in column 3. Adjusting the gap by the bonus on rebate percentages, rebate dollars to enrollees increased by $0.40 in high-quality contracts, an effect indistinguishable from zero.

Appendix Figure B3 shows the raw trends. Bids and benchmarks followed parallel pre-trends but increased more in high-quality contracts after the payment reform. The bid-benchmark gap remained roughly constant between high and low quality throughout the sample period, and rebate increased only slightly in high-quality contracts. These results suggest that the bonus payments to high-quality contracts are largely retained as insurer revenues, with near-zero pass-through to enrollees in the form of rebates.

Table 7 estimates the effects of QBP on the premiums and drug deductibles charged by high-quality contracts using equation 2. We do not find meaningful premium differences between high- and low-quality contracts in either Part C or Part D (column 1-2). Appendix Figure B4 shows the raw trends. Evidence for significant differences in drug deductibles is also not strong. Column 3 estimates a marginally significant decrease in the drug deductibles of high-quality contracts. Relevant for 84% of MA-PD plans (cf Table 2), the share of zero-deductible plans did not differ by quality (column 4). Moreover, overall coverage generosity as measured by the rebate remained similar between high- and low-quality contracts.

The null effect on the price differences at the contract level is at odds with the significant effect on the risk scores of high-quality contracts. Since plan generosity and prices did not vary differentially by quality, risk scores tend to follow parallel trends. Yet, risk scores decreased in high-quality contracts. Within the same contract, however, premiums can differ across counties with different exposure to risk types, and the margin of response is not captured by the analysis so far. We examine this margin of response next.

5 Varying Prices Across Counties to Risk Select Enrollees

The previous section suggests that the effect on risk score is not explained by the contract-level differences in prices or service area characteristics. Within contracts, however, there is large variation across service areas overlooked in the contract-level analysis. Panel B

[19] To avoid the confounding effect due to changes in the risk score, the payment variables are risk-neutralized to reflect the cost of insuring an average Medicare enrollee of risk score 1.

[20] The effect is mechanic since we do not observe a similar increase in the raw benchmarks of high-quality contracts (point estimate $0.52, p-value 0.85), or an increase in the service area benchmarks in Table 5.

of Table 2 summarizes the variation. High-quality contracts offer a greater number of plans in a given county, charge higher premiums for both the Part C and Part D coverage, and have slightly larger market shares among rated contracts. Difference in the drug deductibles remain statistically insignificant.

This section investigates the risk composition change from within the contract's service area, comparing prices charged by the same contract across counties with different FFS risk scores. Specifically, we estimate the following tripe-difference design

$$y_{clt} = \beta_0 \cdot risk_{cl} \cdot high_c \cdot post_t + \beta_1 \cdot risk_{cl} \cdot post_t + \beta_2 \cdot high_c \cdot post_t + \beta \cdot X_{lt} + \alpha_{cl} + \tau_t + \epsilon_{clt}. \quad (3)$$

The variables *high* and *post* identify the high-quality group and the post QBP period as in Section 4. The outcome variables are prices varying at the level of contract c, year t, and location l. Locations refer to the counties in a contract's service area. 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. Variable $risk_{cl}$ measures the risk score differences across counties in a contract's service area. Specifically, we calculate county l's distance to the median county risk score in the service area of contract c, and use the distance-to-median measure in $risk_{cl}$ as the key independent variable in the analysis.[21] By construction, $risk_{cl}$ varies across locations within contracts, and varies across contracts given location.

We include contract-county fixed effects α_{cl} to absorb the baseline differences in prices across contracts and counties, as well as contract-specific unobserved heterogeneity across counties. We control for year fixed effects in τ_t. Assuming that price differences over county risk scores would have remained similar absent the reform, β_1 gives the effect of bonus payments on the price variation in low-quality contracts. Further assuming that the price variation over risk scores is parallel between high and low quality absent the reform, β_0 gives the effect of larger payments on the price variation in high-quality contracts.

In X_{lt}, we control for time-varying, location-specific payment incentives that may affect prices in these locations. Specifically, we include yearly raw benchmarks, bonus rates, and bonus-adjusted benchmarks in X_{lt}.[22] 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-quality contracts differentially entered high-bonus counties or

[21] 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 distance-to-median measure for each county in the service area. Section 5.5 shows robustness of key results to alternative measures of risk differences within contracts.

[22] We use the maximum bonus applied to 5-star contracts to measure the benchmark generosity in a county. The maximum bonuses are higher in the double bonus counties.

exited high-risk counties, then selected service area characteristics would result in biased estimates of the price variation. However, Section 4.2 finds little evidence of selection over service area risk or benchmark, mitigating this endogeneity concern.

To illustrate the identifying variation, we show difference-in-differences estimates separately for high- and low-quality contracts, before showing the triple-difference estimates on high quality. 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 requires that the within-contract price variation over county risk scores remains similar in pre-reform years. In the triple-difference setting, it further requires that the differential variation by quality remains similar in pre-reform years.

5.1 Varying Premiums to Risk Select Enrollees

We estimate the premium response of insurers by first investigating Part C premiums, and then Part D premiums. Table 8 shows the differences in Part C premiums in low- and high-quality contracts (column 1-2), and the differential effect on high-quality contracts (column 3).[23] We do not find significant differences over county risk scores in either quality. Column 4-6 restrict the within-contract locations to the lower and upper 15% of county risks in the contract's service area. Although selection in these counties has greater impacts on the overall risk score, variation in the Part C premium is similar in the risk tails and remains statistically insignificant.[24]

Premiums varied significantly over county risk scores in Part D. Across the 15% risk tails, Table 9 estimates that Part D premiums were $1.70 higher in high-quality contracts for every ten percentage point increase in the county risk score (column 6), or 7% above the mean. The differential effect is driven primarily by the high-quality contracts (column 5), whereas the variation in low-quality contracts is small and insignificant (column 4). We find similar patterns across the full service area in column 1-3, where the variation is weaker in high-quality contracts and remains insignificant in low-quality contracts.[25]

Figure 5 displays the effects on Part D premiums. The event study shows similar premium variation over county risk scores between 2009 and 2010 in each quality, and similar differential by quality in 2009-2010. After the reform, Part D premiums increased

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

[24] Appendix Figure B5 plots the event study and the raw trends. Although premiums stayed on parallel trends in 2009-2010, there is no significant change in premiums across county risks after the payment reform.

[25] The effect on high-quality contracts is concentrated in those rated 4.0-4.5 stars in the baseline (Appendix Table A5). Only a small number of high-quality contracts ever obtained 5-star ratings, and the premium variation within these contracts is somewhat weaker.

by as much as $2.40 per ten percentage point increase in the county risk score ($2.48 across 15% tails) in high-quality contracts. The variation in low-quality contracts remained insignificant throughout the sample period. On the raw trend, instead of showcasing the premium variation over risk score itself, we split the service area of each contract into high- and low-risk counties – grouping either by the median or across 15% tails – and plot the price variation over binary groups of county risks for an average high- and low-quality contract.[26] We find larger increases in premiums in the upper 15% of county risk scores – especially in 2011 and 2012 – relative to the parallel price levels in low-quality contracts over the same period. Premium differences were similar in pre-reform years both within and between quality.

Since a fair amount of Medicare Advantage contracts are priced at zero premiums (cf Table 2), we also examine the offer of zero-premium plans across service areas as one particular margin of response by insurers. Consistent with the variation in the price levels, high-quality contracts significantly increased the offer of zero-premium plans in low-risk counties (Appendix Table A6), and an increasing share of enrollment in zero-premium plans came from low-risk counties (Appendix Table A7). In both cases the effects are concentrated in the Part D premiums. Appendix Figure B6 shows the raw trends and the event study.

Combining Part C and Part D premiums, total premiums increased by $3.56 more in high-quality contracts per ten percentage point risk score in the risk tails (Appendix Table A8). Effect sizes are comparable over the QBP demonstration period (Appendix Figure B7). The premium differences favoring enrollees in low-risk counties potentially contribute to the risk composition gain in high-quality contracts. We examine this mechanism in detail in Section 5.3. We next look at drug deductibles.

5.2 Varying Drug Deductibles to Risk Select Enrollees

The pricing of drug deductibles did not vary substantially across service areas or contract quality. Appendix Table A9 shows the results of estimating equation 3 using drug deductible as dependent variable. The estimates show that both quality 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-quality contracts (Appendix Figure B8). Since 2010, event study estimates indicate rising drug deductibles with county risk scores in high-quality contracts. Due to the noise in the data, we do not pursue the variation in drug deductibles as a potential mechanism of the risk composition

[26] Because of this, the variation picked up in the event study graph appears more muted on the raw trend.

gain in high quality, and mostly focus on the variation in premiums. This is justified also by the fact that, as we discuss briefly in Section 5.4, differences in premiums account for around 70%-80% of the differences in the overall generosity by quality.

5.3 Mechanism

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

Socio-Economic Factors. Appendix Table A10 estimates the premium variation by county differences in per capita income and transfer income. We do not find significant premium variation 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-quality contracts. Appendix Table A11 finds similar null effects by county demographics such as racial composition and college education.

Market Competition. Appendix Table A12 explores market competition as an alternative driver of the premium variation. Premiums increased with market concentration in high-quality contracts (column 2).[28] However, risk scores tend to be lower in more concentrated markets, implying lower premiums in riskier counties due to market competition. Moreover, controlling for the effect of competition increased the premium variation over risk scores (column 5). The evidence indicates that the premium variation is not driven by the competitive effects on prices, but is constrained by such effects in more competitive markets.

Provider Quality. We next consider differences in provider costs and quality as alternative drivers of the premium variation 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 variation over risk scores. To investigate the quality channel, we use hospital readmission rates and preventable

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

[28] Specifically, greater concentration at the county level increases premiums in high-quality contracts, but concentration within quality markets has no significant impacts on premiums (column 7-9).

hospital stays as measures of inpatient and outpatient quality. We do not detect meaningful variation over either quality in Appendix Table A13.

Provider Cost. We investigate the cost channel exploiting adjustments on fee-for-service (FFS) costs in Appendix Table A14. Premiums did not vary over costs in either quality. 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 column 5-8.[29] Adjusting price-standardized costs by risk scores in column 9-12 cuts the effect size on high-selection contracts by half and renders the variation insignificant.[30] 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 A16 adjusts risk scores with the diagnosis intensity factors developed in Finkelstein *et al.* (2017).[31] Upon adjustment, we find stronger variation of Part D premiums over risk scores relative to the main results in Table 9. 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 variation 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. We further analyze the economic incentives driving the premium variation over health in Section 6, where we examine the design details of the quality rating to understand the pricing response by insurers.

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

[30] Appendix Table A15 shows similar effects across binary groups of high- and low-cost counties. In these specifications the scale of the cost variable has no direct impact on the effect sizes. Premium differences over costs are greatly reduced once we take out the risk component in costs.

[31] 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 heath conditions.

5.4 Insurance Generosity

Although we mainly focus on premiums, other price and non-price designs of the insurance contract may also vary in favor of healthier individuals. Instead of analyzing every detail of the insurance contract in relation to the risk composition change in high quality, we use rebate as a measure of overall insurance generosity and seek to understand the extent of generosity differences that can be explained by differences in premiums.[32]

We estimate equation 3 using rebates as the dependent variable in column 4 of Appendix Table A17. The estimate suggests that rebates increased by \$4.44 less in high-quality contracts for every ten percentage point increase in the county risk score. Of the \$4.44 loss of rebates, \$3.56 was added onto premiums in high-quality contracts (Appendix Table A8). Put together, premium differences account for 80% of the differences in the overall generosity by quality.

In high-selection contracts, the loss of rebates relative to low-quality contracts was \$7.17 for every ten percentage point risk score (column 5, Appendix Table A17), whereas premiums increased by \$5.07 above low-quality contracts (column 6, Appendix Table A18). Based on these estimates, about 70% of the rebate differences between low-quality and high-selection contracts are attributable to premiums. We therefore do not expect generosity differences to be a major channel of selection once we control for the differences in premiums.

5.5 Sensitivity Analysis

Alternative Weights. In the main analysis, we weight plan-level premiums and deductibles by enrollment to generate contract-level price variables. Plans and prices with larger enrollments are weighted more at the contract level. Alternatively, taking simple averages across plans, Appendix Table A19 to A21 find similar variation of premiums and drug deductibles in the service area. Appendix Figure B9 to B11 show the raw trends and event study estimates. Moreover, to limit the effect of abnormally large prices on the estimated variation, we use the median plan price as the contract price, and find similar variation as in the main analysis (Appendix Table A22). Appendix Figure B12 shows the raw trends and event study of median prices.

Alternative Risk Measures. The results are robust to alternative measures of risk score

[32] Unlike premiums, rebates vary at the level of plans but not locations. Plans then distribute the rebates to lower prices across locations in the service area. Assuming that the quality differential in the distribution of rebates across counties is similar to the quality differential in the premium variation across counties, we compare the triple-difference estimates to inform the extent of rebate differences explained by premiums.

differences within contracts. The main analysis uses the distance-to-median measure. Using the distance-to-mean measure, Appendix Table A23 to A25 find similar variation in premiums and drug deductibles. Appendix Figure B13 plots the event study estimates for this set of estimates. In the risk tails, instead of looking at the lower and upper percentiles of county risk scores, Appendix Table A26 looks at counties more than one standard deviation away from the mean county risk score. The effects are comparable to those in the main analysis across the 15% tails.

6 Why Does QBP Induce Risk Selection?

So far we have shown that high-quality contracts have significantly improved their risk pool since the QBP demonstration. The composition change is not driven by differences in the average price levels between quality but rather by premium differences within contracts across service areas. However, we have been silent on why risk selection is incentivized in the first place: what are the specific aspects of the payment model which made low-risk individuals more profitable to high-quality insurers?

This section suggests that the selection incentive may lie with the design of the quality rating itself, activated by the financial returns to quality introduced in the payment demonstration. Specifically, when differences in patient health conditions are not fully adjusted for in the quality measures, healthier individuals are associated with higher quality ratings and bonus payments to insurers. We illustrate the selection incentive in a stylized model in Section 6.1, characterize it empirically using between- and within-contract variation in Section 6.2 and 6.3, and examine the pricing responses in Section 6.4.

6.1 A Model of Quality Rating, Risk and Insurer Profit

To illustrate how the risk-quality linkage can affect premiums across locations, we build a simple two-county model where the insurer can use premium differences to select healthier enrollees and obtain higher quality ratings and payments. The insurer's revenue depends on the premiums in the two counties (p_1 and p_2), and on the benchmark B, which increases with the quality rating q.

The demand for the contract in county l is given by $s_l = s_l(p)$, and the average risk score of contract enrollees is given by $r(p_1, p_2)$. Due to the measurement bias, rating q can depend on the risk score r. This allows the insurer to affect the quality rating by changing its risk pool across markets.[33] We assume perfect risk adjustments on benchmarks and

[33] We follow Curto *et al.* (2014), and model p_l as the extra revenue covering the difference between the

bids, so that the insurer faces a constant marginal cost (c). In this world, risk selection would have no bearing on the insurer profits absent the linkage with quality and quality 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. From the first order conditions, the optimal premium in county l solves

$$p_l^* = c - B + \left(1 + \frac{dB}{dq} \cdot \frac{dq}{dr} \cdot \frac{\partial r}{\partial p_l} \cdot \frac{s_1 + s_2}{s_1}\right) \cdot |\varepsilon_l|^{-1},$$

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

Absent the risk-quality linkage, optimal premium equals marginal cost plus a mark-up which is inverse to demand semi-elasticity. When the average risk score affects the quality rating ($\frac{dq}{dr} \neq 0$), it distorts prices through $\frac{\partial r}{\partial p_l}$. Therefore, the insurer will set different premiums across the two counties if p_1 and p_2 affect the average risk score differently. Specifically, if higher risk score is associated with lower quality rating ($\frac{dq}{dr} < 0$), premium will be lower in county one relative to county two if, other things equal, county one is more conducive to risk selection (i.e., $\frac{\partial r}{\partial p_1} > \frac{\partial r}{\partial p_2}$). To see this, because $\frac{dB}{dq} \cdot \frac{dq}{dr} < 0$ when $\frac{dq}{dr} < 0$, premiums respond negatively to $\frac{\partial r}{\partial p_l}$, or decrease in counties more conducive to risk selection.

Hence, the observed premium variation is consistent with the selection of healthier individuals if lower enrollee risk scores contribute to higher quality ratings. Under this configuration, we expect premiums to decrease in counties with healthier enrollees, relative to the baseline period where payment did not depend on the quality rating. This effect should be stronger for contracts with greater gains from selection on the quality rating ($\frac{dB}{dq} \cdot \frac{dq}{dr}$). We next examine the risk-quality mechanism in detail, signing $\frac{dq}{dr}$ empirically.

6.2 Selection on Health Outcome Measures

Under the lenses of the model, the Quality Bonus Payment demonstration could incentivize selection if some measures in the quality rating are sensitive to the risk scores of enrollees. In this case, selection would be financially incentivized by the greater returns to quality after QBP. In this section, we first indicate that ratings on the health outcome measures tend to favor contracts with lower enrollee risk scores. The relationship is primarily driven by the outcomes measures for chronic conditions such as diabetes, which do not account for the health status of enrollees at the time of enrollment. We then confirm that this link allows high-quality contracts with low-risk enrollees to achieve greater improvements in

insurer's bid and the benchmark.

their outcome ratings.

Risk Scores and the Outcome Rating. As mentioned in Section 2, starting in 2012, the overall rating is a weighted average of measure ratings applying 3.0 weights to the outcome measures, 1.5 weights to the access measures, and 1.0 weights to the process measures. All measures received 1.0 weights prior to 2012. As a result, the contribution of outcome measures to the overall rating increased significantly after QBP (Appendix Table A27). To detect the effect of risk scores on the outcome rating, we exploit the introduction of QBP and the cross-contract variation over baseline risk scores (denoted by *Risk*) in a difference-in-differences analysis analogous to equation 2.

We report the estimates in Table 10. The table considers only outcome measures that were consistently included in the quality rating from 2009 to 2014 ("health improved" and "diabetes and blood pressure").[34] On average, a 10 percentage point increase in the baseline risk score is associated with a 12.2 percentage point decrease in subsequent outcome ratings (column 1). Grouped by the source of outcome measures, self-reported health improvement measures from the Health Outcome Survey (HOS) are minimally correlated with risk scores (column 2). By contrast, measures of diabetes and blood pressure management, drawn from the Healthcare Effectiveness Data and Information Set (HEDIS), are significantly and negatively correlated with risk scores (column 3). The overall correlation between risk scores and outcome ratings is therefore completely driven by the HEDIS measures.[35]

A closer inspection of the HEDIS measures suggests that a lack of risk adjustment on these measures potentially contributes to the negative correlation with risk scores. All three measures – *Diabetes Care: Blood Sugar Controlled*, *Diabetes Care: Cholesterol Controlled*, and *Controlling Blood Pressure* – consider chronic conditions controlled if results from medical testings are below certain thresholds, and rank contracts based on the percent of patients showing successful control of these conditions. *Blood sugar controlled*, for instance, is measured by the percent of diabetes patients with a recent blood sugar (A1c) testing below 9%.[36] However, the patient pool in the denominator is a simple sum of diabetic and

[34]Outcome measures later introduced to the quality rating include hospital re-admission measures (first appearing in 2012), drug adherence measures (2012), and quality improvement measures (2014). The difference-in-differences approach is more suitable for measures that were present in the quality rating throughout the sample period.

[35]Appendix Figure B15 plots the raw trends and event study estimates.

[36]Similarly, *Cholesterol Controlled* is measured by the percent of diabetic enrollees with a recent LDL-C testing below 100 mg/dL. *Controlling Blood Pressure* is measured by the percent of hypertension enrollees who lowered blood pressure below 140/90 during enrollment. Details of the outcome measures are available in the yearly Technical Note. The 2013 Note is accessible at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/

hypertensive enrollees, and the medical thresholds in the numerator are uniformly applied to all patients without adjusting for the severity of conditions. The lack of risk adjustment implies that outcome ratings would be negatively correlated with risk scores, if lower-risk enrollees are more likely to pass the medical thresholds and boost the outcome ratings of insurers.[37] In this case, contracts that selected healthier enrollees after the payment reform should also experience greater improvements in the outcome rating. We examine this relationship below.

Linking Selection and Contract Quality. We connect the change in the outcome rating to the risk-selection behavior of high-quality contracts that emerged in Section 4 and 5. In particular, we test whether high-quality contracts with riskier enrollees experienced smaller increase in the outcome ratings, and whether high-selection contracts with greater improvements in the risk pool also improved more in the outcome ratings.

Appendix Table A28 reports difference-in-differences estimates from the specification in equation 2, where we compare the rating dynamics across high-quality contracts with different baseline enrollee risk scores. Contracts that started out with riskier enrollees had lower outcome ratings in subsequent years. The loss of outcome ratings was larger in higher-rated contracts (4.5 stars) and higher percentiles (top 25%) of risk scores, driven completely by lower ratings on the chronic condition measures from HEDIS. Appendix Figure B16 plots the raw trends. Ratings on access and process measures did not differ across baseline enrollee risk scores (Appendix Table A29).

Using a similar approach, we show that lower service area risk scores mitigate the loss of outcome ratings in higher-rated contracts (4.5 stars). Contracts serving lower-risk areas saw smaller decrease in the outcome rating (Appendix Table A30). In the lower 25% of service area risks, the loss of HEDIS rating is only marginally larger than in low-quality contracts, implying significant relative gains over high-quality contracts in riskier counties. Figure 6 plots the raw trends. Contracts in the lowest-risk counties also decreased risk scores more than other high-quality contracts (Appendix Table A4), and risk selection was not associated with significant gains in the access and process ratings of these contracts

Downloads/2013-Part-C-and-D-Preview-2-Technical-Notes-v090612-.pdf

[37] The health literature has raised similar concerns over the lack of risk adjustments on the HEDIS quality measures (e.g., Zhang *et al.*, 2000, Safford *et al.*, 2009). 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.

(Appendix Table A31).

Next, we provide additional evidence on the link between risk scores and outcome ratings in two ways. First, we confirm that the outcome ratings responded to risk scores by exploiting the fact that outcome ratings released in year t depend on measures obtained two years prior. Second, we draw on additional data on chronic conditions to show that contracts increased premiums in areas with higher prevalence of chronic conditions using an approach similar to that in Section 5. Ultimately, these findings support the construction of the instruments that we will use to quantify selection in Section 6.5.

6.3 Quality and Risk Scores Over Time

This section examines the within-contract correlation between outcome ratings and risk scores over time. We exploit the 2-year delay between the measurement of chronic conditions in HEDIS and their entry in the quality rating to construct contemporaneous risk-outcome correlations, and compare them with correlations across other lag or lead periods *for the same contract*. If riskier individuals have worse measured outcomes, then a negative correlation should occur between outcome ratings and risk scores 2 years prior.

Figure 7 plots the correlation coefficients between HEDIS outcome ratings in year t and risk scores in year $t-3$, $t-2$, $t-1$ and t. We restrict the correlation to within-contract variation in ratings and risk scores over time, and group the coefficients by baseline quality. The contemporaneous pattern in panel (b) shows an average correlation coefficient of -0.5 (-0.6 median) for contracts with at least a 4.5 star in the baseline, where 75% have a negative coefficient over the sample period. The median correlation coefficient in high-quality contracts is -0.36, with more skewed distribution to the negative in high-selection contracts. We do not find similar magnitude of correlation or differentials by quality in other lag or lead periods.[38]

Appendix Table A32 shows the regression estimates of outcome ratings on risk scores across periods. Estimated effects of risk scores are large and significant in high-selection contracts, and insignificant in the lag and lead periods. However, these estimates may be biased due to selected risk scores in the outcome measures. We correct for the selection and estimate the causal impact of risk scores on outcomes using an instrumental variable approach in Section 6.5.

[38] Similar correlation patterns exist between risk scores and the overall star ratings in Appendix Figure B17. Lower risk scores are associated with higher star ratings two years later, and contracts with greater decreases in risk scores exhibit more negative within-contract correlations.

6.4 Varying Premiums to Deter Enrollees with Chronic Conditions

This section inspects the selection response of insurers to the outcome ratings. Adopting the framework of Section 5.1, we compare premiums across counties with different exposure to chronic conditions. We report estimates based on a triple difference regression analogous to equation 3, but instead of county risk scores in the variable *risk*, we focus on chronic condition prevalence rates interacted with coding-adjusted county risk scores to account for the health differences across counties. In this way, the raw prevalence rate is adjusted downward in counties where patients have fewer and less severe conditions.[39]

We focus on diabetes in Appendix Table A33, which estimates the premium variation over diabetes prevalence rates. Part D premiums increased by \$8.85-\$11.25 per ten percentage point increase in the prevalence rate in high-selection contracts (column 6- 7). Figure 8 plots the premium differences across high- and low-prevalence counties. Although the raw trends suggest larger premium increases in 2011-2012 in high-prevalence counties, effect sizes are comparable over years in the event study. We find similar effects on higher-rated contracts (4.5 stars) in Appendix Table A34 (raw trends in Appendix Figure B18), where larger selection effects are concentrated in high-selection contracts serving lower-risk counties. We also find similar results for hypertension.[40]

To summarize, premiums varied significantly over chronic conditions in high-selection and higher-rated contracts. Both types of contracts saw large improvements in the risk pool and outcome ratings after the payment reform. The selection response is incentivized by the design of outcome ratings favoring low-risk enrollees, and the correlations with risk scores are stronger within high-selection and higher-rated contracts. Building on these results, we develop an instrumental variable approach in the next section to causally assess the selection incentive due to QBP.

6.5 Quantifying Risk Selection in the Health Oucome Measure

Our previous analyses suggest that the HEDIS outcome measures are subject to substantial gaming by insurers. To causally estimate the impact of risk scores on the HEDIS outcomes, this section develops an instrumental variable (IV) strategy that relies on the finding that contracts varied premiums across counties to attract healthier individuals and improve

[39] We obtain diabetes prevalence rates by county from the Center of Disease Control (CDC), and hypertension prevalence rates from Olives *et al.* (2013). Appendix C provides more details on the underlying data and the construction of prevalence rates.

[40] Appendix Table A35 and Figure B19 show the premium variation over hypertension prevalence rates in higher-rated contracts. The variation is generally weaker but remains concentrated in high-selection contracts below the median service area risk score (0.975).

the composition of the risk pool.

We formulate the production of health outcomes as determined by a contract-specific quality component and a component due to the risk scores of enrollees. Specifically, we estimate the following equation

$$y_{ct} = \alpha_c + \alpha_c \cdot post + \beta \cdot risk_{ct-2} + \tau_t + \epsilon_{it}, \tag{4}$$

where y_{ct} is the health outcome (as measured by HEDIS) of contract c in year t. Outcomes are measured from enrollees two years prior, and $risk_{ct-2}$ denotes the concurrent risk score of these enrollees at the contract level.[41] 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 control the chronic conditions of a unit-risk enrollee: $\alpha_c \cdot post$ measures the value-added in quality after the payment reform. 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 instrumental variable (IV) strategy to separate selection from the value-added in quality. Since not all contracts improved outcomes by lowering risk scores, we estimate β across sub-groups of contracts and in particular focus on high-selection contracts with larger decreases in the risk score.

We exploit the premium variation over county risk scores to construct instruments for $risk_{ct-2}$. Our empirical analyses show that premiums responded *directly* to the health of enrollees. The cross-space selection of healthy individuals affected the risk composition of contracts, which we exploit in the first-stage prediction of contract-level risk scores. Moreover, premiums did not respond to the local supply or quality of care, or to other demand factors potentially correlated with health (Section 5.3). For the IV research design, it implies that the premium variation is unlikely to affect outcomes through the composition effect on other, non-health determinants of outcomes, lending support to the exclusivity condition. We hence construct the main instrument using correlations between premiums and coding-adjusted county risk scores within contract-years (Finkelstein *et al.*, 2017), and construct similar correlations with diabetes and hypertension prevalence rates as additional instruments for over-identification.[42]

Table 11 shows the OLS and two-stage-least-square (TSLS) estimates of the effects of

[41] The 2-year lag in the subscript distinguishes between the release and the measurement of outcomes: outcomes released in year t are obtained from enrollees in year $t-2$ (Figure 7).

[42] The instruments are $Corr(p_{ct-2}, R_{c0})$, where p_{ct-2} is the vector of Part D premiums in the service area of contract c and year $t-2$, and R_{c0} is the vector of baseline coding-adjusted county risk scores, health-adjusted diabetes prevalence rates and hypertension prevalence rates in the service area of contract c. We construct the correlation coefficients within contract-years to instrument for contract-level risk scores in $risk_{ct-2}$.

risk scores on the HEDIS outcomes. Recall that the HEDIS outcomes calculate the share of diabetes or hypertension patients who have controlled their conditions below specific medical thresholds. We look at the percentage of such enrollees as the dependent variable in the table. Premium variation strongly predicts risk scores in high-quality contracts serving low-risk counties (column 4-5), but not in low-quality contracts (column 1) or high-quality contracts serving riskier counties (column 3). Correcting for selection, the causal effect of risk scores on the outcome measures is significantly larger than the OLS estimates. The implied gains from selection is substantial. In high-selection contracts (column 4), lowering risk scores by 10 percentage points increased outcome measures by 10.28 percentage points. Evaluated against the risk composition gain over low-quality contracts from 2010 to 2012 (0.025), selection increased outcome measures in high-selection contracts by 2.57 percentage points,[43] whereas the value-added per unit of risk is a modest 0.48 percentage point. Put differently, around 85% of the outcome gains in high-selection contracts is attributable to selection.

Appendix Table A37 estimates the effects of risk scores separately for the three HEDIS outcomes in high-selection contracts. Selection has larger effects on controlling blood sugar and cholesterol levels, where 57% to 85% of the outcome change is attributable to selection. For blood pressure, the share of patients with blood pressure below 140/90 increased by 5.21 percentage points after the payment reform, and selection on risks explains little of the improvement. Overall, the quality value-added is smaller in outcomes based on lab tests, where other types of gaming, such as repeated measurements and selection of the most favorable result, are less likely to affect outcomes.

In conclusion, a significant portion of the outcome gains in high-selection contracts is due to the selection of healthier individuals. Given the effect of risk scores on outcomes, the lack of risk adjustments on outcomes allowed high-selection contracts to receive unearned bonus payments from Medicare. In the next section, we calculate the cost of selection to Medicare by estimating the effect of selection on the overall star rating linked to payments. We further discuss the broader impacts of selection on the access to high-quality insurance across counties, and suggest ways to improve the design of the quality rating.

[43] Specifically, $\Delta\text{Risk} \cdot \widehat{\beta_{TSLS}} = -0.025 \cdot (-102.81) = 2.57$. ΔRisk is the event study coefficient for year 2012 in the contract-level analysis of risk scores (panel b of Figure 4).

7 Discussion

7.1 The Cost of Selection

Our findings show that the health outcome measures in the quality rating are sensitive to the risk composition of contracts, and in response, insurers enrolled healthier individuals using premium variation across counties. Risk selection is profitable if it raises the overall star rating and hence the bonus payments to insurers. We regard bonus payments rewarding selected risk types rather than value-added as overpayments, and use it as a measure of the cost of selection to Medicare. To estimate this cost, we first estimate the effect of selection on the overall star rating. We then calculate the savings in bonus payments from removing the selection gains in the overall rating.

Selection and the Quality Rating. We apply the IV strategy developed in Section 6.5 to estimate the effect of risk scores on the quality rating. Specifically, 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 4. We show results for high-selection contracts in Appendix Table A38. Risk scores have large, negative impacts on the outcome rating (column 1-2), where 71% to 77% of the star change after the payment reform is due to selection.[44] The risk composition gain over low-quality contracts from 2010 to 2012 increased outcome ratings by 0.46 star in high-selection contracts. Adjusted by the weights of outcome measures (50% across all measures), overall rating increased by 0.23 star in high-selection contracts on average. Ratings on the lower-weighted access and process measures are minimally affected by risk scores.

To construct counterfactual ratings absent the change in risk scores, we first recover the underlying continuous star rating of each contract in 2014.[45] From the continuous rating, we subtract the selection component due to the risk score change from 2010-2012. We estimate the risk score change for each high-selection contract using low-quality contracts as controls.[46] The adjusted rating holds the risk composition fixed at 2010, and removes the effect of selected risk scores since 2011 on the quality rating.

[44] The selection effect is comparable to but different from the 85% calculated in Section 6.5 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 6.5 we focused only on the three HEDIS measures.

[45] The underlying 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.

[46] We obtain similar risk score changes when we construct a weighed average of low-quality contracts as the synthetic control for each high-quality contract (Abadie *et al.*, 2010).

We compare the selection-adjusted rating and the original star rating in panel (a) of Figure 9. The horizontal axis groups high-selection contracts by the original star rating in 2014. The vertical axis shows the percent of enrollees experiencing a change in the star rating, conditional on the original rating. Adjusting for selection, around 50% of enrollees in 4.0-star contracts and all enrollees in 3.5-star contracts would see a reduction of quality rating to 3.0 stars. The affected contracts were marginal high-quality contracts rated 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-quality 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 on the star rating in the 4.5-5.0 range, where 96% of enrollees were in contracts rated above 4.5 stars on the continuous scale. Adjusting for selection, 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 alters the generosity of bonus payments to insurers (Table 1). To map the change in the star rating to the change in firm bidding and hence program cost, we make the behavioral assumption that insurers adjust bids to the new benchmark but keep the rebate to enrollees unchanged. This assumption is supported by our empirical finding of a near-zero pass-through of bonus payments to enrollees (Table 6). We therefore infer counterfactual bids by inverting equation 1 holding rebates at the pre-adjustment level.[47]

We plot the difference in program costs after risk adjustment in panel (b) of Figure 9, where costs are calculated on a monthly basis given the 2014 star rating. Selection has larger impacts on the program cost in the 3.5-4.0 star range, where overpayment amounted to $24 per enrollee. Both the change in the star rating and the drop in benchmark bonuses below 4.0 stars added to the cost of selection in this group. Put differently, marginal high-quality contracts would have lost sizeable bonus payments, if they did not improve the quality rating above 4.0 stars and qualify for higher benchmark bonuses. In the 4.5-5.0 range, since all contracts remained above 4.0 stars where there is little bonus variation, the cost of selection averaged $1 per enrollee in this group.[48]

To evaluate the size of the overpayment, we compare it with the benchmark bonus to high-selection contracts. Relative to low-quality contracts, high-selection contracts received $70 more in benchmarks in 2014.[49] Selection raised the bonus payments to

[47] For insurers bidding above the benchmark, the program cost is the benchmark adjusted by the new star rating.

[48] Appendix Figure B20 plots the results when the risk score change is derived from the synthetic controls. We find similar costs of selection to most contracts and larger costs to 3.5-star contracts. The average cost of selection in the 3.5-4.0 range remains similar ($25).

[49] The estimate is the event study coefficient for 2014 in the contract-level analysis of bonus-adjusted

high-selection contracts by $9 per enrollee, or 13% of the benchmark bonus in 2014. In marginal high-quality contracts (3.5-4.0 stars), selection increased program costs by 34% of the benchmark bonus. Put differently, for every dollar of payment that improved quality, another $0.52 was paid to selected risk scores. Scaled by the enrollment-months in high-selection contracts, cost of selection amounted to $51.7 million annually in 2014, with nearly $45 million concentrated in marginal high-quality contracts. Insurers retained these payments as profits.

In conclusion, we show that risk scores have significant effects on the quality rating linked to payments. Without risk adjustment, healthier individuals are more profitable to insurers, and the selection response led to sizeable costs to the program. Furthermore, selection can have large impacts on program enrollees through the supply of high-quality insurance across markets. We discuss the effect of selection on the spatial distribution of quality and the welfare implications next.

7.2 Distribution of Quality Across Counties

The premium variation implies heterogeneous pass-through of bonus payments favoring low-risk service areas. More generally, the selection response can lower the market share of high-quality contracts in riskier counties, and shift the spatial distribution of quality across county risk scores. We estimate the effect of the QBP demonstration on county-level market shares of high- and low-quality contracts using the following specification

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

where $risk_l$ is the baseline fee-for-service risk score in county l. y_{clt} is the market share of contract c in county l and year t.

The left panels of Figure 10 plot the market share changes in the lower and upper 15% tails of county risk scores. We see clear diverging patterns by quality. At the contract level (panel a), high-quality contracts increased market shares in the lowest-risk counties (gray lines) and decreased market shares in the highest-risk counties (purple lines). Panel c aggregates contract market shares by high and low quality, and shows similar differences at the quality level. There is substantial divergence in high-quality market shares across the risk tails (solid lines), and modest divergence in low-quality market shares (dotted lines).

benchmarks (Table 6).

We present detailed evidence on the market share changes and premiums across the 15% risk tails in Appendix D. Premiums of high-quality contracts increased by $6.51 over low-quality contracts in the top 15% riskiest counties, where high-quality market share decreased by 12.96 percentage points. In the lower risk tail, high-quality market share grew to over 17 percentage points higher than the low-quality market share in 2014, whereas the difference was only 2 percentage points in 2009-2010.[50]

The shifts in the spatial distribution of quality are the result of a supply shock on insurer revenues. Consumer knowledge of quality ratings and preferences for quality are not directly affected by the payment reform. Therefore, the supply-side response to subsidies plays important roles in the regional disparities of healthcare access in the Medicare population. To the extent that a larger share of bonus payments rewarded health rather than quality, the supply-side response directed health spending to the healthiest counties, and increased the transfer from taxpayers to insurers. Insurers captured the payments by redistributing insurance benefits from risky to healthy enrollees, directly hurting the welfare of the riskier enrollees. These distributional impacts lowered the insurance value of the program to the riskiest population, and substantially widened the gap in the access to high-quality insurance across regions.

7.3 Improving the Design of the Quality Rating

The selection response to quality bonus payments suggests that some form of risk adjustment is necessary to inform the cross-contract comparison of outcomes. Since no official risk-adjustment guideline exists for the HEDIS measures, we presented one approach where an outcome production function is estimated to correct for selection using some knowledge of the selection mechanism by firms. Alternatively, a pre-specified adjustment formula can be applied to the outcome data directly. The ex-ante approach cannot in itself address the coding and selection incentives created by the adjustment formula. To limit these unintended effects on insurers, a complementary policy would down-weight the outcome measures and re-orient the quality rating towards access and process measures less susceptible to the selection response by insurers.

A second issue with the quality rating is that performance measures are pooled from the contract's service area. An average contract covers 25 counties in the service area. However, since the design of the insurance contract and the delivery of healthcare services both

[50] Appendix Table A36 estimates the market share changes over the full range of county risk scores. For every 10 percentage point increase in the risk score, market share of high-quality contracts decreased by 11.8 percentage points (column 2), or by 9 percentage points differentially over low-quality contracts (column 3). The event studies are in the right panels of Figure 10.

occur at the county level,[51] issuing quality ratings at the contract level is uninformative of the actual quality of care experienced by enrollees, and generates cross-space selection incentives favoring high-quality counties and counties with greater selection gains on quality. The within-contract premium variation is driven by both the effect of risk scores on quality and the measurement of quality at the level of contracts rather than counties.

Finally, changes in the quality ratings are generally uninformative of the absolute quality gains or losses experienced by contracts. This is because the quality rating is a relative ranking of contracts within the Medicare Advantage market. Apart from unadjusted differences in risk scores and service areas, the lack of an external benchmark further weakens the link between quality ratings and the underlying quality. To improve the quality rating, we echo MedPAC (2010) in proposing a localized ranking of contracts at the county level, where stars are assigned based on quality gains over similar case-mixes in the fee-for-service program. The within-county ranking nets out location-specific factors and limits cross-space selection in the current setting. Moreover, the ranking is directly useful to enrollees choosing between the fee-for-service program and different quality tiers of the private option. The new rating can be supported by more consistent keeping and sharing of data between the fee-for-service and the private Medicare.

8 Conclusions

The Medicare Advantage (MA) market is an important setting to understand the strategic response of private insurers to government payments. Created alongside the traditional, fee-for-service Medicare, whether the private option delivers better care quality at lower costs is central to the success of the MA program. Moreover, as the pay-for-performance model proliferates in the healthcare sector, the design of market-wide quality ratings as the basis of payments becomes increasingly important. Lessons from the MA rating provide valuable guidance on this effort. More generally, the incentive effects inform the design of other public-private partnerships where social insurance is contracted out to private entities. Promoting quality while limiting the regulatory capture by private firms is the central concern in these contexts.

Our results point to significant challenges of designing pay-for-performance contracts in selection markets. Even small design details can have large-scale distributional impacts when coupled with high-powered financial incentives. In the MA market, failing to risk-adjust the quality rating for enrollee health differences resulted in systematic selection of

[51] Specifically, the network design of contracts must meet adequacy criteria at the county level, including the minimum number of providers and the maximum travel time to access them for enrollees in a county.

healthier individuals by high-quality insurers. The selection imposed significant costs on the program, shifted the distribution of high-quality insurance towards healthier counties, and lowered the benefit pass-through to riskier enrollees. Adjusting for selection, the value-added in quality is modest. The unintended effects on insurers and the consequent increase in health inequality illustrate the critical role of risk-adjustment mechanisms in the design of quality-based payment models in insurance markets.

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.

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.

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

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

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

CBO (2012). *Medicare Advantage: Quality Bonus Payment Demonstration Undermined by High Estimated Costs and Design Shortcomings:[letter to Orrin G. Hatch, Ranking Member, Committee on Finance, United States Senate*. US Government Accountability Office.

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.

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

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. (2014). Can health insurance competition work? Evidence from Medicare Advantage, National Bureau of Economic Research, No. w20818.

Dafny, L. and Dranove, D. (2008). Do report cards tell consumers anything they don't already know? The case of Medicare HMOs. *The Rand Journal of Economics*, **39** (3), 790–821.

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.

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*.

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

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.

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

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.

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

Li, P. and Doshi, J. A. (2016). Impact of Medicare Advantage prescription drug plan star ratings on enrollment before and after implementation of quality-related bonus payments in 2012. *PloS One*, **11** (5), e0154357.

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

MedPAC (2010). *Report to the Congress, Medicare Payment Policy*. Medicare Payment Advisory Commission.

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.

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

Polyakova, M. and Ryan, S. P. (2019). Subsidy targeting with market power, National Bureau of Economic Research, No. w26367.

Reid, R. O., Deb, P., Howell, B. L. and Shrank, W. H. (2013). Association between Medicare Advantage plan star ratings and enrollment. *Jama*, **309** (3), 267–274.

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: Quality adjustments on benchmarks and rebates

Year	Star Rating					
	≤2.5	3.0	3.5	4.0	4.5	5.0
Benchmark Bonus $\alpha^{star} = 1 + \%$						
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%
Rebate Percentage τ^{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%

Source: Centers for Medicare and Medicaid Services, Advance Notice of Methodological Changes, Calendar Year 2009-2014.

Table 2: Summary statistics

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
	Full Sample		Low Quality		High Quality		(V)-(III)
	mean	s.e.	mean	s.e.	mean	s.e.	p-value
			Panel A: Contract-Year Observations				
Risk Score	0.97	0.0075	0.97	0.0093	0.96	0.012	0.55
Number of Counties	25.09	5.40	28.19	7.74	18.18	2.21	0.22
Number of Plans	3.40	0.23	3.53	0.31	3.12	0.28	0.33
Service Area Risk	0.99	0.006	1.00	0.008	0.96	0.008	0.00
Enrollment (k)	334.75	34.95	328.35	39.19	349.06	71.56	0.80
Benchmark	874.10	5.72	883.08	6.52	854.06	10.87	0.023
Bid	763.38	6.28	763.65	7.58	762.80	11.25	0.95
Benchmark-Bid	110.72	5.55	119.43	6.90	91.27	8.68	0.012
Rebate	78.37	3.73	83.55	4.68	66.80	5.74	0.025
Part C Premium	29.64	2.44	20.25	2.51	50.63	4.51	0.00
Zero Part C Premium (%)	48.74	2.81	59.27	3.29	25.23	3.90	0.00
Part D Premium	19.42	1.19	15.00	1.37	29.30	1.73	0.00
Zero Part D Premium (%)	44.23	2.87	54.98	3.42	20.23	3.68	0.00
Drug Deductible	32.62	4.42	32.85	5.72	32.11	6.40	0.93
Zero Drug Deduc (%)	84.21	1.89	84.70	2.36	83.11	3.07	0.68
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	0.35
Number of Plans	1.76	0.073	1.59	0.088	2.22	0.093	0.00
Part C Premium	31.96	2.65	25.17	2.74	49.94	5.47	0.00
Zero Part C Premium (%)	37.36	3.25	43.06	4.03	22.25	4.83	0.00
Part D Premium	20.74	1.43	17.77	1.73	28.60	2.17	0.00
Zero Part D Premium (%)	35.04	3.27	41.49	4.06	17.97	4.29	0.00
Drug Deductible	28.65	5.88	30.05	7.69	24.92	6.23	0.60
Zero Drug Deduc (%)	84.25	2.95	83.40	3.87	86.55	2.91	0.511
Market Share (%)	33.51	1.72	31.77	1.97	38.12	3.20	0.089
N	20,472		14,861		5,611		

Notes: Table summarizes the estimation sample. Benchmark and bid variables are in dollars per risk-neutral enrollee per month, while the other pricing variables are in dollars per enrollee per month. Enrollment is measured as enrollee-month counts in a given year. Contracts rated at least 3.0 stars in the 2009-2010 baseline are included in our sample, summarized in column 1-2. Baseline high- and low-quality contracts are summarized in column 3-4 and 5-6, respectively. We define high-quality contracts as those rated 4.0 stars and above in 2009-2010, and low-quality contracts as those rated below 4.0 stars but no less than 3.0 stars in 2009-2010. Column 7 tests the null of zero differences between high- and low-quality contracts. Plan characteristics are aggregated to the contract-year level in Panel A, and to the contract-county-year level in Panel B, both weighted by enrollment. Details of the sample construction are in Appendix C.

Table 3: Effect of QBP on risk scores

	(I)	(II)	(III)	(IV)	(V)	(VI)
High · Post	-0.026***	-0.041***	-0.035***	-0.042***	-0.020***	-0.045***
	(0.008)	(0.014)	(0.012)	(0.015)	(0.007)	(0.015)
Weights	plan enrollment		equal weights		unweighted	
Fixed Effects	contract		contract		plan	
y mean	0.97	0.97	0.97	0.97	0.96	0.96
R^2	0.86	0.007	0.76	0.012	0.79	0.009
N	1,122	1,127	1,122	1,122	4,549	4,549

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

Notes: Table shows difference-in-differences estimates on risk scores. Column 1-2 show the effects on contract risk scores aggregated from plan risk scores weighted by enrollment. Column 3-4 show the effects on contract risk scores as simple averages across plans. Column 5-6 show the effects on plan risk scores. We control for contract or plan fixed effects in the odd-numbered columns, and show the cross-sectional differences in the even-numbered columns. Standard errors clustered at the level of linked contracts (column 1-4) or plans (column 5-6) in the parenthesis.

Table 4: Effect on risk scores, by service area risks

	(I)	(II)	(III)	(IV)
Treat · Post	-0.037***	-0.016	-0.043***	-0.015
	(0.010)	(0.010)	(0.012)	(0.016)
Treat	high-quality + risk		high-quality + risk	
	< median	> median	< 25%	> 75%
Control	low-quality		low-quality	
y mean	0.97	0.97	0.97	0.97
R^2	0.86	0.86	0.86	0.86
N	920	977	851	894

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

Notes: Table shows difference-in-differences estimates on risk score, dividing the sample of high-quality contracts by the service area risk in the baseline. Column 1 and 2 estimate the effect on high-quality contracts below and above the median service area risk (0.975), relative to low-quality contracts. Column 3 estimates the effect on high-quality contracts in the lower 25% of service area risk (<0.902), and column 4 estimates the effect on high-quality contracts in the upper 25% of service area risk (>1.007). All specifications control for contract fixed effects. Standard errors clustered at the level of contracts in the parenthesis.

Table 5: Effect of QBP on market characteristics

	(I) # Counties	(II) Risk	(III) Benchmark	(IV) Double-Bonus County	(V) # Plans
High · Post	8.70	0.002	1.80	-0.020	-0.17
	(8.39)	(0.002)	(2.94)	(0.021)	(0.23)
y mean	25.09	0.99	799.15	0.72	3.40
R^2	0.73	0.98	0.96	0.90	0.87
N	1,122	1,122	1,122	1,122	1,122

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

Notes: 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. 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 variation in these characteristics. All regressions include contract fixed effects. Standard errors clustered at the level of contracts in the parenthesis.

Table 6: Effect of QBP on benchmarks, bids, and rebates

	(I) Benchmark	(II) Bid	(III) Benchmark-Bid	(IV) Rebate
High · Post	27.84***	37.01***	-9.17	0.39
	(7.10)	(7.50)	(6.07)	(3.68)
y mean	874.10	763.38	110.72	78.37
R^2	0.83	0.84	0.83	0.87
N	1,122	1,122	1,122	1,122

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

Notes: Table shows difference-in-differences estimates on benchmarks, bids and rebates. 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 7: Effect of QBP on premiums and drug deductibles

	(I) Part C Premium	(II) Part D Premium	(III) Drug Deductible	(IV) Zero Deductible
High · Post	0.81	2.33	-16.98*	0.051
	(2.87)	(1.58)	(8.98)	(0.045)
y mean	29.64	19.42	32.62	0.84
R^2	0.87	0.81	0.69	0.63
N	1,122	1,122	1,122	1,122

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

Notes: Table shows difference-in-difference estimates on premiums and drug deductibles. We aggregate plan level variables 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 8: Effect of QBP on Part C premiums, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			18.08*			18.63
			(10.45)			(13.50)
Risk · Post	-10.43	7.44	-11.33*	-8.74	6.88	-10.34
	(6.42)	(10.25)	(6.29)	(6.75)	(12.80)	(6.48)
High · Post			-3.68			-4.66
			(3.79)			(3.99)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	25.17	49.94	31.96	24.97	75.47	31.20
R^2	0.77	0.85	0.82	0.77	0.85	0.82
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part C premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 9: Effect of QBP on Part D premiums, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			12.43*			16.98**
			(7.49)			(7.98)
Risk · Post	-4.56	13.56**	-3.42	-4.75	14.76**	-4.32
	(4.83)	(6.31)	(4.71)	(5.17)	(6.99)	(4.97)
High · Post			2.44			3.50*
			(2.31)			(1.90)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	17.77	28.60	20.74	17.46	27.60	20.22
R^2	0.76	0.67	0.75	0.75	0.70	0.75
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part D premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 10: 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: 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 11: Effect of selection on the HEDIS outcome

	(I)	(II)	(III)	(IV)	(V)
			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)
$\alpha_c \cdot$ Post	1.39	-0.35	-1.10	2.49	2.96
			Panel B: TSLS		
Risk Score		-87.03*		-102.81***	-161.89**
		(49.25)		(36.52)	(65.25)
First-stage F-stat	2.00	10.01	4.26	11.11	22.90
Over-id p-value	–	0.28	–	0.38	0.13
$\alpha_c \cdot$ Post		-0.00		0.48	-3.89
ΔRisk $\cdot\widehat{\beta_{TSLS}}$		1.57		2.57	5.83
Contract	low	high	high	high	high
Service area risk			>50%	≤50%	≤25%
y mean	65.86	71.04	71.77	70.37	64.51
N	1,280	669	396	228	116

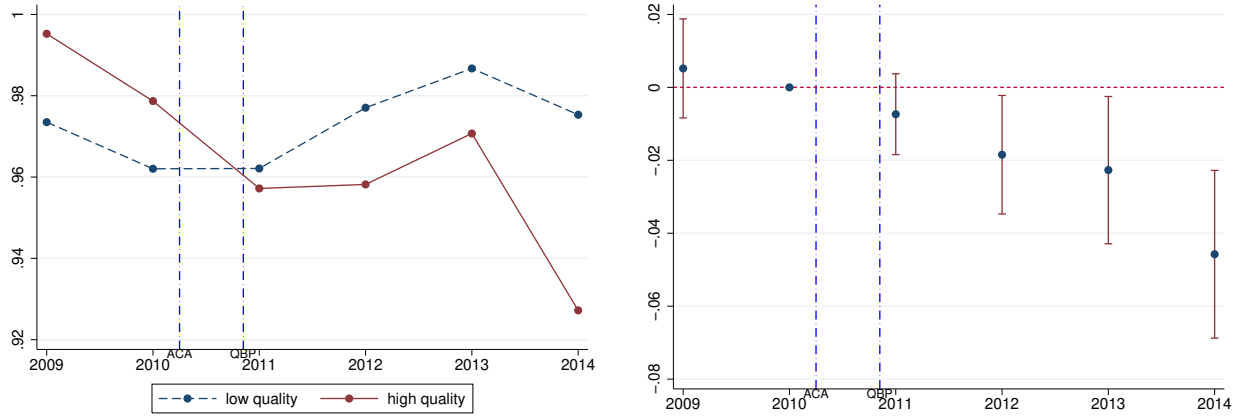
*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Notes: Table shows the effect of risk scores on the HEDIS outcomes, measured in percentages with chronic conditions controlled. Panel A shows the OLS estimates. Panel B shows two-stage-least-squares (TSLS) estimates where contract-level risk scores are instrumented by the correlation between Part D premiums and county risk scores in the service area. Specifically, instruments are constructed as $Corr < p_{ct-2}, R_{c0} >$, where p_{ct-2} is the Part D premiums in contract c and year $t-2$, and R_{c0} is the baseline coding-adjusted county risk scores in the service area of contract c. We construct additional instruments using correlations with health-adjusted diabetes and hypertension prevalence rates in R_{c0}. The instruments strongly predicted contract-level risk scores in high-quality contracts (column 2), with effects concentrated in high-selection contracts serving lower-risk counties (column 4-5). For these contracts, we calculate quality value-added in $\alpha_c \cdot$Post, and gains from selecting lower risk types in ΔRisk $\cdot\widehat{\beta_{TSLS}}$. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. We obtain stronger first-stage prediction and similar IV estimates when the endogenous variable is plan rather than contract risk score and when we control for plan-level fixed effects and contract-level value-added. Robust standard errors clustered at the level of contracts in the parenthesis.

Figure 1: Effect on risk scores, event study

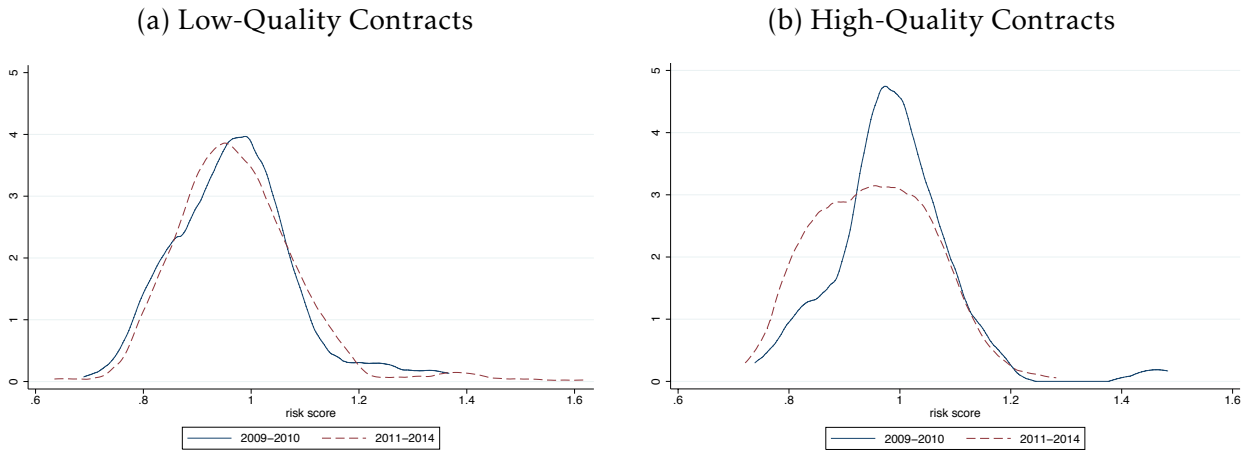
(a) Raw Trend

(b) Event Study



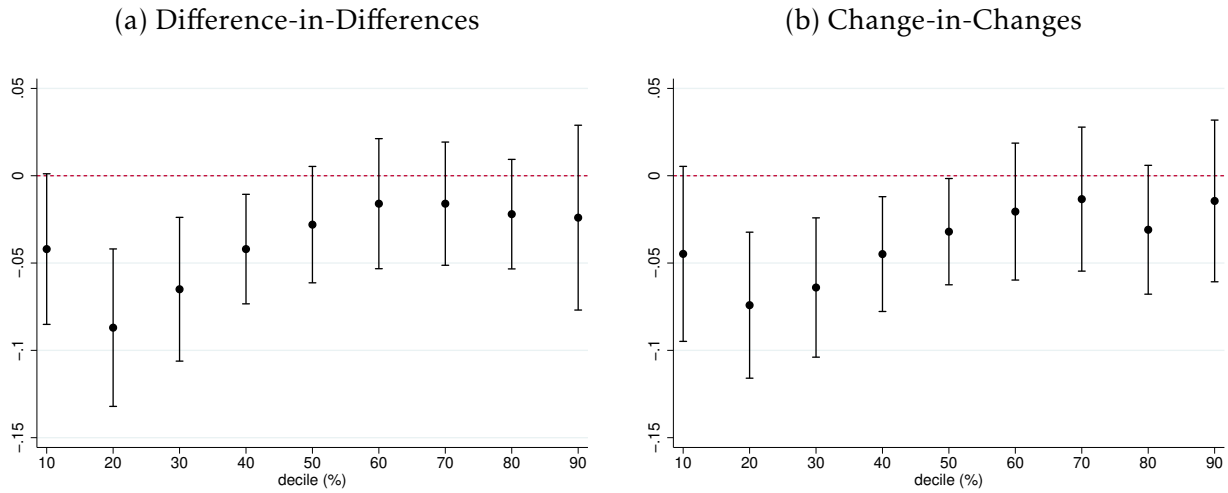
Notes: Panel (a) shows the raw trend of risk scores for baseline high- and low-quality contracts. Risk scores are aggregated from plan risk scores weighted by enrollment. Panel (b) plots the event study estimates on the differential change in high-quality risk scores controlling for contract and year fixed effects. 95% confidence intervals are plotted based on robust standard errors clustered at the level of contracts.

Figure 2: Effect on risk scores, kernel density



Notes: Figure plots the kernel density of risk scores for high-quality contracts in panel (a), and for low-quality contracts in panel (b). Separate densities are drawn for the pre- (2009-2010) and post-QBP (2011-2014) period. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

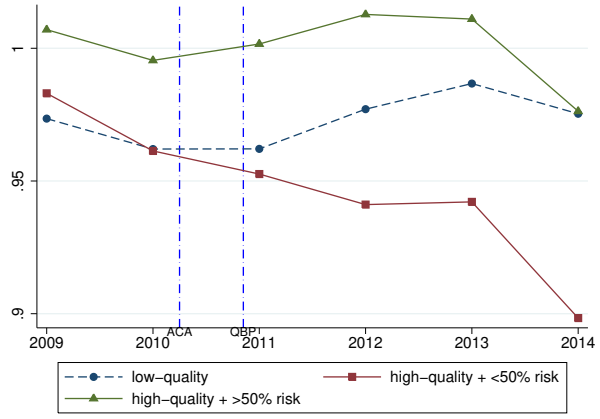
Figure 3: Effect on risk scores, by deciles



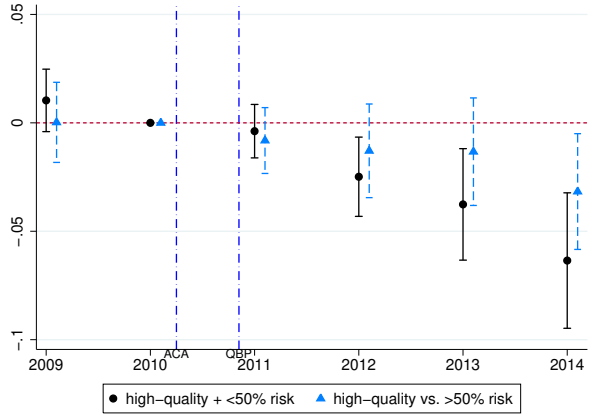
Notes: Figure plots the quantile difference-in-differences estimates on deciles of risk scores in panel (a), and the change-in-changes estimates in panel (b). 95% confidence intervals based on bootstrapped standard errors are plotted. A tabulated version of the results listing point estimates, standard errors, and the baseline deciles of high-quality risk scores is available in Appendix Table A3.

Figure 4: Effect on risk scores, by service area risks, event study

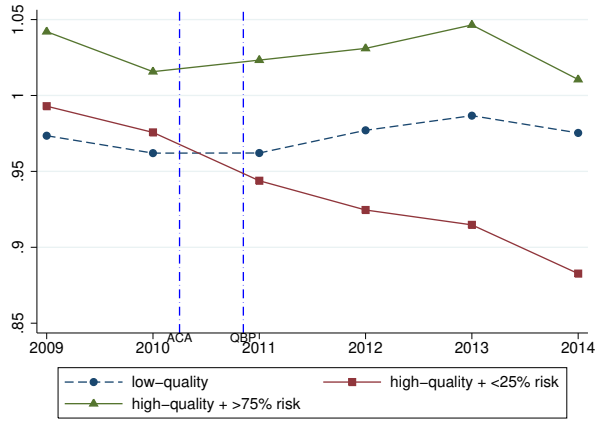
(a) Raw Trend



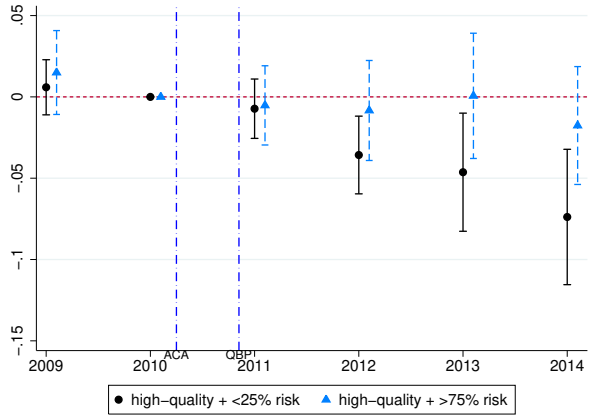
(b) Event Study



(c) Raw Trend, 25% Tails



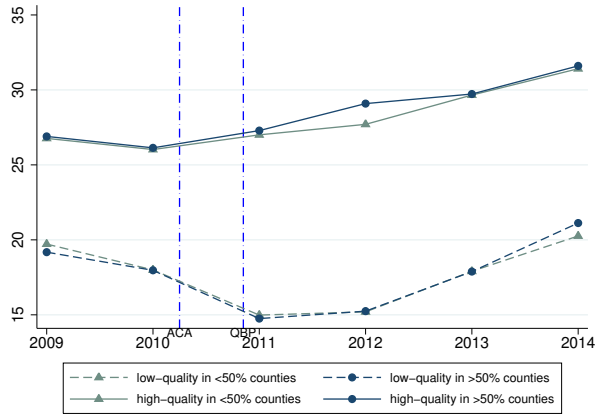
(d) Event Study, 25% Tails



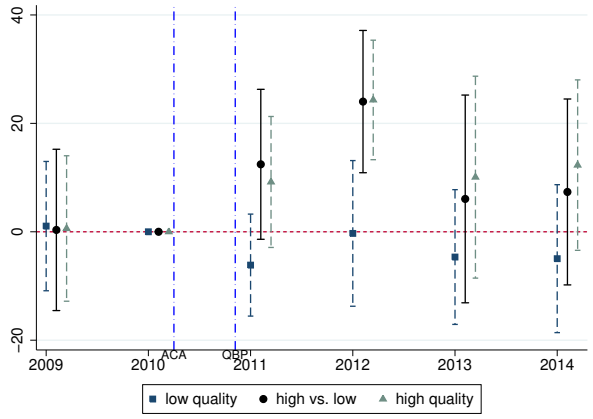
Notes. Panel (a) shows the raw trends of risk scores in high-quality contracts above and below the median service area risk (0.975) and in low-quality contracts. Panel (b) shows the event study estimates separately for high-quality contracts above and below the median service area risk. Plotted 95% confidence intervals are based on robust standard errors clustered at the level of contracts. Panel (c) plots the raw trends of risk scores in high-quality contracts across the 25% tails of the service area risk (below 0.902 or above 1.007) and in low-quality contracts. Corresponding event study estimates are shown in panel (d).

Figure 5: Effect on Part D premiums, within-contract variation, event study

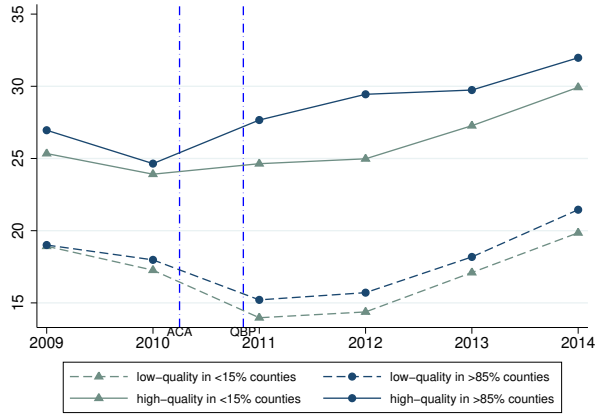
(a) Raw Trend



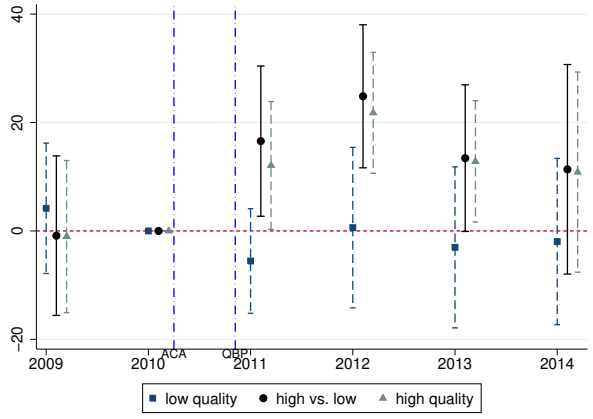
(b) Event Study



(c) Raw Trend, 15% Tails



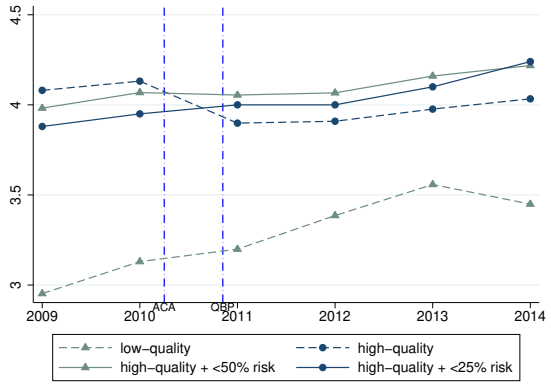
(d) Event Study, 15% Tails



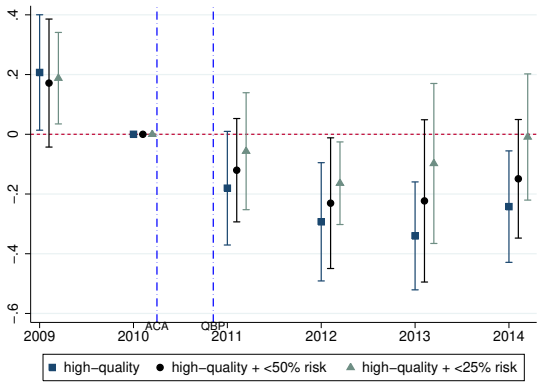
Notes. Figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality 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-quality contract (dotted lines) and an average high-quality contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 6: Outcome ratings by quality and service area risks, event study

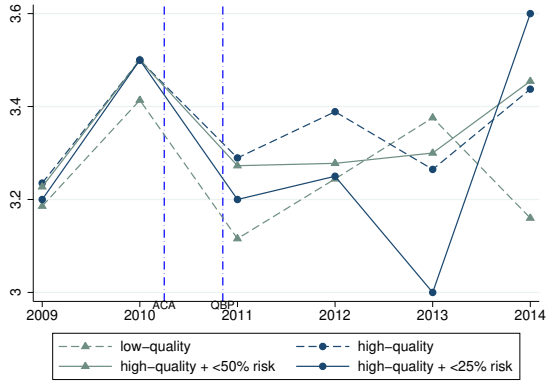
(a) Average Outcome Rating, Raw Trend



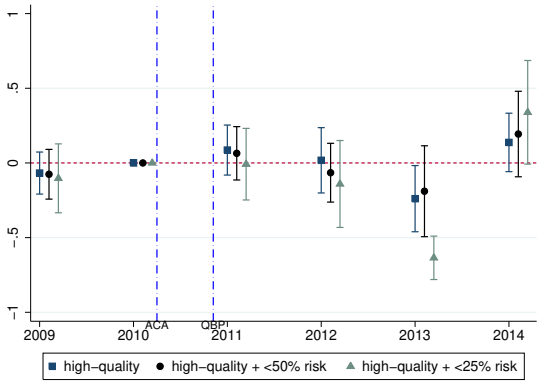
(b) Average Outcome Rating, Event Study



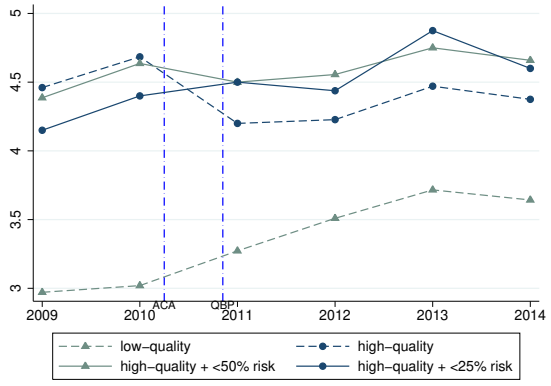
(c) Health Improved, Raw Trend



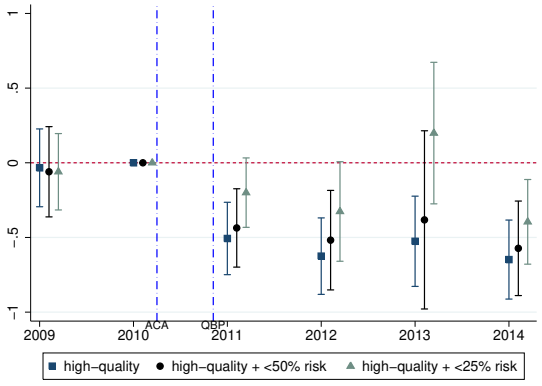
(d) Health Improved, Event Study



(e) Diabetes & Blood Pressure, Raw Trend



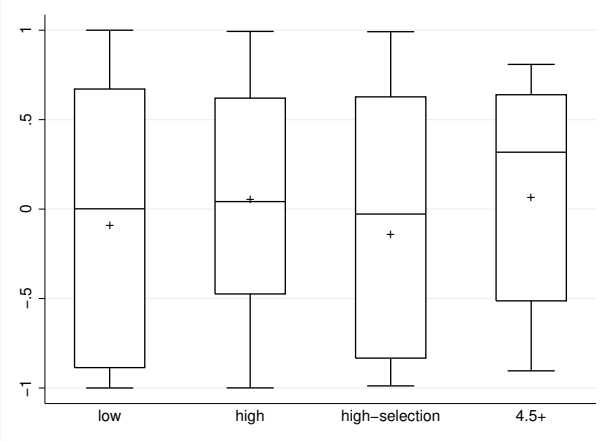
(f) Diabetes & Blood Pressure, Event Study



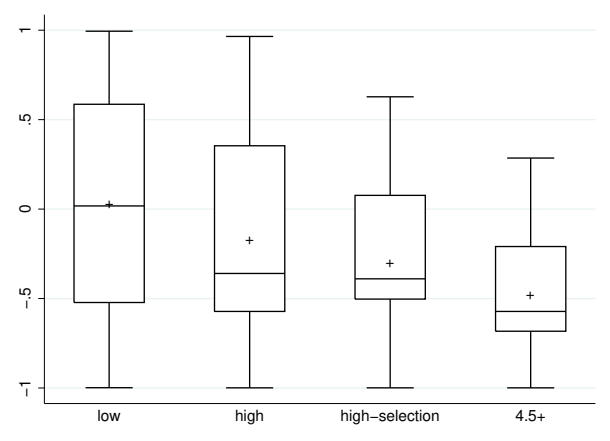
Notes. Figure shows the dynamics of outcome ratings by baseline quality and service area risk. High-quality contracts are restricted to those rated 4.5 stars and above in the baseline, or the higher-rated contracts. The left panels plot separate trends for sub-groups of higher-rated contracts with service area risk scores below the median (0.975) and in the lower 25% (less than 0.902) of high-quality contracts, and for low-quality contracts. The right panels show the event study estimates from difference-in-differences specifications comparing ratings in (sub-groups of) higher-rated contracts with the low-quality controls. Panel (a) and (b) look at the average ratings 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 7: Within-contract correlations between risk scores and HEDIS outcome ratings

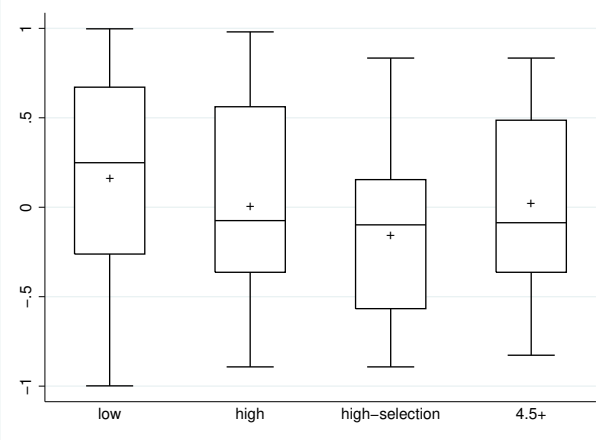
(a) $Corr(Outcome_t, Risk_{t-3})$



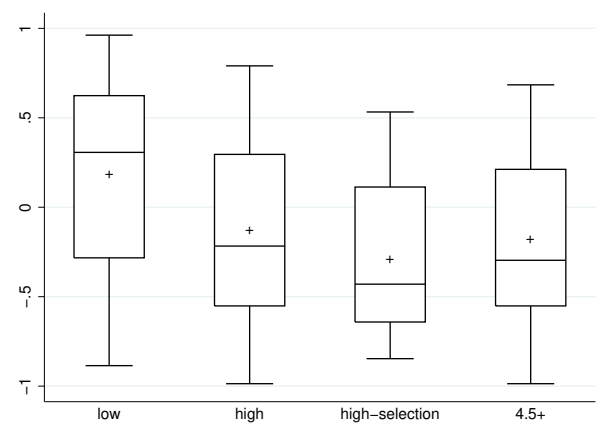
(b) $Corr(Outcome_t, Risk_{t-2})$



(c) $Corr(Outcome_t, Risk_{t-1})$



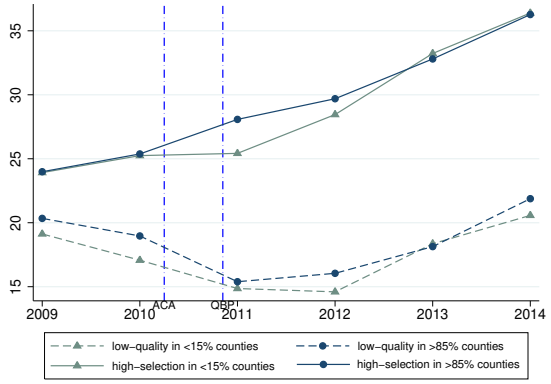
(d) $Corr(Outcome_t, Risk_t)$



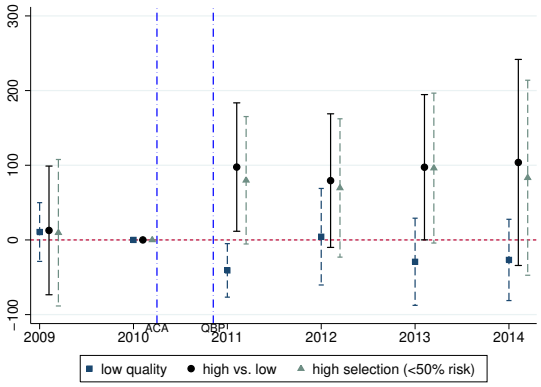
Notes. Figure plots the distribution of correlation coefficients between HEDIS outcome ratings in year t and risk scores in year $t-3$ (panel a), $t-2$ (panel b), $t-1$ (panel c), and year t (panel d). The coefficients are calculated within contracts over time, grouped by the baseline contract quality. The middle line represents the median coefficient. The lower and upper edges of the box bound the 25th-75th percentile. The antennas reach the nearest observations outside. The plus (+) marks the mean coefficient.

Figure 8: Effect on Part D premium, within-contract variation over health-adjusted diabetes prevalence rates, high-selection contracts, event study

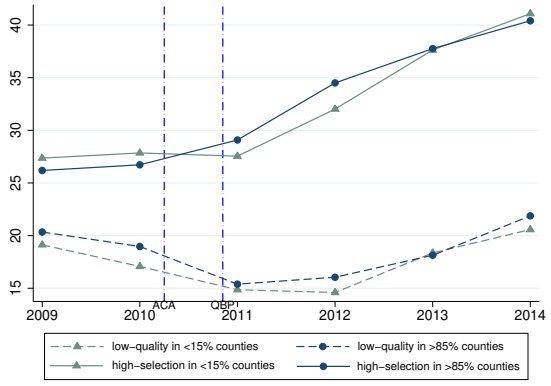
(a) High-Selection (<50% Risk), Raw Trend



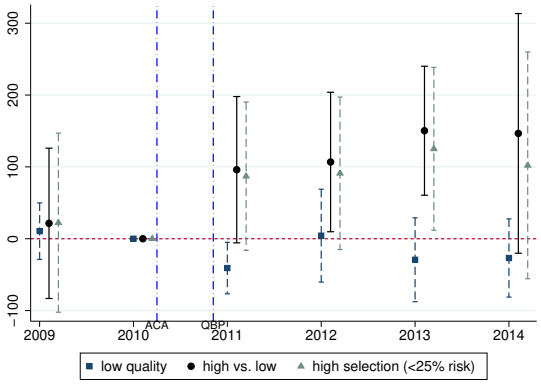
(b) High-Selection (<50% Risk), Event Study



(c) High-Selection (<25% Risk), Raw Trend



(d) High-Selection (<25% Risk), Event Study



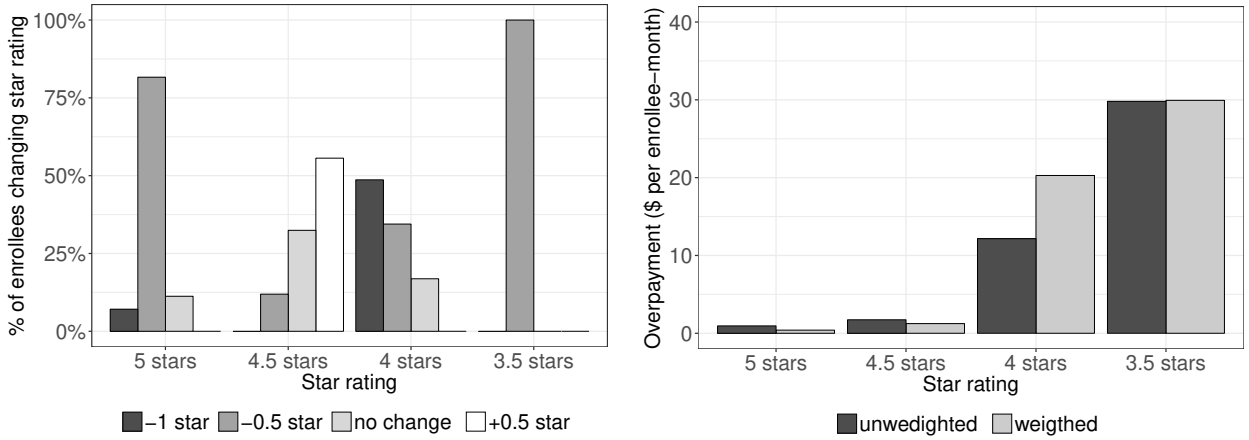
Notes. Figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract variation over county differences in health-adjusted diabetes prevalence rates in the right panels. The health-adjusted prevalence rate multiplies the raw prevalence rate by the coding-adjusted 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-quality 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 variation 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 9: Effects of selection on the quality rating and overpayments

(a) Share of Enrollees with Star Rating Change

Absent Selection

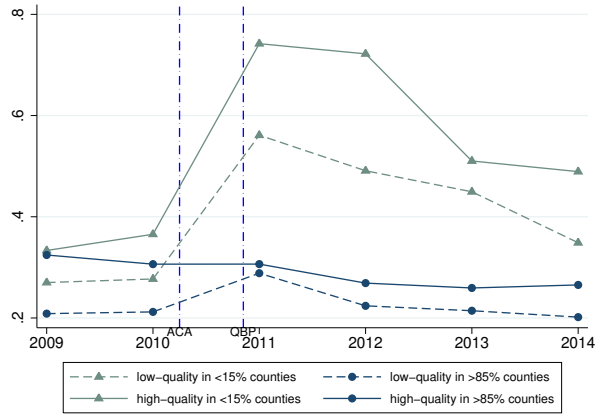
(b) Overpayments due to Selection



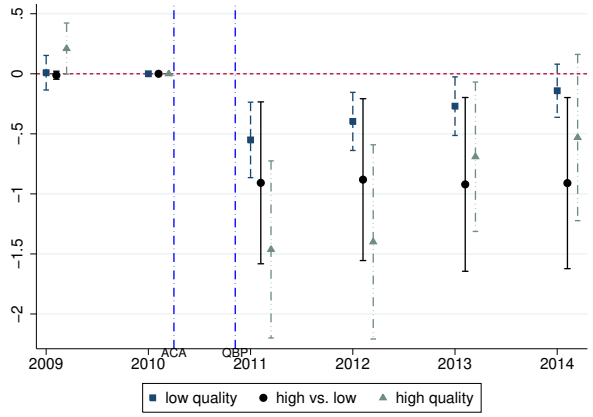
Notes. Figure shows the effect of adjusting risk selection on the overall star ratings of high-selection contracts in panel (a) and on the payments to these contracts in panel (b). Panel (a) plots for each overall star rating level in 2014 (horizontal axis) the percentage of enrollees receiving lower (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 the payments across 2014 star ratings. 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 behavioral assumption is supported by our empirical analysis of bidding and pricing strategies at the contract level. 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 10: Effect on market shares, cross-county variation, event study

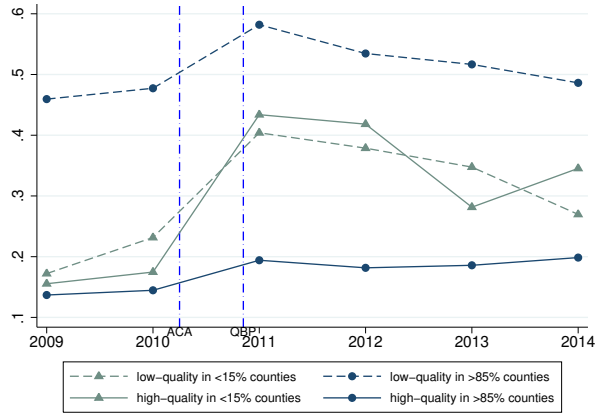
(a) Contract-County-Year, Raw Trend, 15% Tails



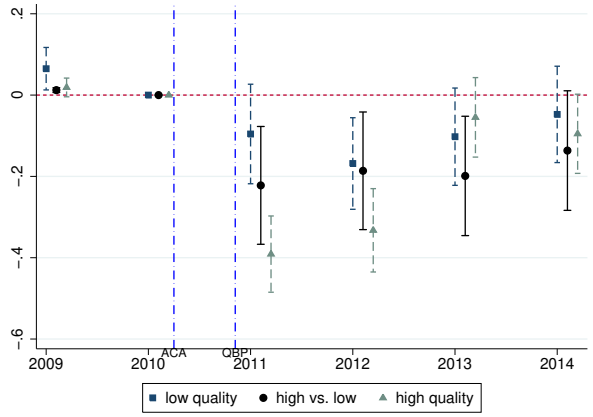
(b) Contract-County-Year, Event Study, All Counties



(c) Quality-County-Year, Raw Trend, 15% Tails



(d) Quality-County-Year, Event Study, All Counties



Notes. Figures shows the cross-county variation in market shares at the level of contracts in panel (a) and (b), and by high and low quality in panel (c) and (d). Quality-level market shares are constructed for a balanced panel of quality-county-years where counties with masked enrollment data in some but not all years receive zero market shares in years with missing enrollments. Event study estimates in panel (b) and (d) show the variation in market shares over continuous county risk scores. 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). Raw trends in panel (a) and (c) show differences in the levels of market shares across the 15% tails of county risk scores.

Online Appendix

A Additional Tables

Table A1: Quality differences before and after QBP, and the effect of current-period ratings on risk scores

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
	≥4.0 stars		risk score		risk score		risk score	
High · Post	0.081	0.051	-0.026***	-0.041***	-0.035***	-0.042***	-0.020***	-0.050***
	(0.076)	(0.073)	(0.008)	(0.014)	(0.012)	(0.015)	(0.007)	(0.016)
Star Rating			0.003	-0.016	-0.003	-0.022	0	-0.036***
			(0.007)	(0.014)	(0.008)	(0.014)	(0.005)	(0.011)
Weights	unweighted		plan enrollment		equal weights		unweighted	
Fixed Effects	contract		contract		contract		plan	
y mean	0.29	0.29	0.97	0.97	0.97	0.97	0.96	0.96
R^2	0.59	0.46	0.86	0.007	0.76	0.012	0.79	0.016
N	1,122	1,127	1,122	1,127	1,122	1,122	4,549	4,549

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

Notes: Table shows difference-in-differences estimates on current period star rating ≥4.0 stars in column 1-2, and on risk scores in column 3-8. Different from the main results in Table 3, column 3-8 also control for current period star ratings in the difference-in-differences specification. We find that quality differences are stable before and after the QBP: baseline high-quality contracts continue to score above 4.0 stars relative to baseline low-quality contracts (column 1-2). Current period star ratings have little effect on risk scores, particularly in specifications with contract or plan fixed effects. The main difference-in-differences estimates barely change with controls of current period star ratings. Standard errors clustered at the level of contracts (column 1-8) or plans (column 7-8) in the parenthesis.

Table A2: Effect of QBP on risk scores, excluding 5-star contract-years

	(I)	(II)	(III)	(IV)	(V)	(VI)
High · Post	-0.024***	-0.040***	-0.035***	-0.041***	-0.017**	-0.043***
	(0.008)	(0.014)	(0.012)	(0.016)	(0.008)	(0.016)
Weights	plan enrollment		equal weights		unweighted	
Fixed Effects	contract		contract		plan	
y mean	0.97	0.97	0.97	0.97	0.96	0.96
R^2	0.86	0.005	0.76	0.011	0.78	0.008
N	1,112	1,117	1,112	1,117	4,332	4,502

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

Notes: Table shows difference-in-differences estimates on risk scores. Column 1-2 show the effects on contract risk scores aggregated from plan risk scores weighted by enrollment. Column 3-4 show the effects on contract risk scores as simple averages across plans. Column 5-6 show the effects on plan risk scores. 5-star contract-years are excluded from the sample. We control for contract or plan fixed effects in the odd-numbered columns, and show the cross-sectional differences in the even-numbered columns. Standard errors clustered at the level of linked contracts (column 1-4) or plans (column 5-6) in the parenthesis.

Table A3: Distributional effects of QBP on risk scores

	(I) Difference-in-Differences	(II) Change-in-Changes	(III) Baseline High
10%	-0.042*	-0.045**	0.848
	(0.022)	(0.026)	
20%	-0.087***	-0.074***	0.919
	(0.023)	(0.021)	
30%	-0.065***	-0.064***	0.951
	(0.021)	(0.020)	
40%	-0.042***	-0.045***	0.966
	(0.016)	(0.017)	
50%	-0.028*	-0.032**	0.980
	(0.017)	(0.016)	
60%	-0.016	-0.021	1.002
	(0.019)	(0.020)	
70%	-0.016	-0.013	1.026
	(0.019)	(0.021)	
80%	-0.022	-0.031	1.057
	(0.017)	(0.019)	
90%	-0.024	-0.014	1.096
	(0.033)	(0.024)	

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

Notes: Table shows the distributional effect of QBP on the deciles of risk score in high-quality contracts relative to the low-quality contracts. Estimates are from a quantile difference-in-differences specification in column 1, and from a change-in-changes specification in column 2. Bootstrapped standard errors from 100 replications in the parenthesis. Column 3 shows the baseline (2009-2010) deciles of high-quality risk scores.

Table A4: Effect on risk scores, high-quality (≥4.0 stars) and higher-rated contracts (≥4.5 stars), by service area risks

	(I)	(II)	(III)	(IV)	(V)	(VI)
Treat · Post	-0.026***	-0.040***	-0.046***	-0.032***	-0.050***	-0.032
	(0.008)	(0.009)	(0.012)	(0.010)	(0.019)	(0.021)
Treat	≥4.0	≥4.5	≥4.5 + risk		≥4.5 + risk	
Service area risk			< median	> median	< 25%	> 75%
Control	low-quality		low-quality		low-quality	
y mean	0.97	0.97	0.97	0.97	0.97	0.97
R^2	0.86	0.86	0.85	0.86	0.85	0.86
N	1,122	901	833	843	801	808

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

Notes: Table shows difference-in-differences estimates on the risk scores of high-quality contracts (rated 4.0 stars and above) in column 1, and on the risk scores of higher-rated contracts (4.5 stars and above in the baseline) in column 2-6. Column 3-4 divide higher-rated contracts by the median service area risk score (0.975) in high-quality contracts. Column 5-6 show results on higher-rated contracts in the lower (below 0.902) and upper 25% (above 1.007) of service area risk scores of high-quality contracts. All specifications control for contract fixed effects. Standard errors clustered at the level of contracts in the parenthesis.

Table A5: Effect of QBP on Part D premiums, within-contract variation, high-quality, higher-rated, and 5-star contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Risk · High · Post					16.98**	19.27	14.64*
					(7.98)	(13.07)	(7.83)
Risk · Post	-4.75	14.76**	20.21*	14.76*	-4.32	-3.84	-4.62
	(5.17)	(6.99)	(11.02)	(7.85)	(4.97)	(5.15)	(5.17)
High · Post					3.50*	5.34**	0.74
					(1.90)	(2.33)	(1.90)
Counties				15% tails			
Contracts	low	≥4.0	≥4.5	5.0	(2) vs. (1)	(3) vs. (1)	(4) vs. (1)
y mean	17.46	27.60	30.39	15.58	20.22	18.87	17.42
R^2	0.75	0.70	0.73	0.88	0.75	0.76	0.75
N	4,393	1,641	535	155	6,034	4,928	4,530

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

Notes: Table shows the within-contract variation in Part D premiums over county risk scores. We examine heterogeneous effects on high-quality contracts by the baseline ratings and 5-star status over the sample period. Column 1-3 show the difference-in-differences estimates on the premium variation in low-quality (baseline ratings below 4.0 stars, high-quality (4.0 stars and above), and higher-rated contracts (4.5 stars and above), respectively. Column 4 shows the difference-in-differences estimate on contracts ever achieving a 5-star rating after 2012 and hence subject to year-round enrollments in the contract. Column 5-7 show the triple-difference estimates on the differential variation in high-quality contracts defined in column 2-4, respectively. We restrict 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 A6: Effect on zero premiums and zero drug deductibles, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(XI)
	Zero Part C Premium			Zero Part D Premium			Zero Drug Deductible		
Risk · High · Post			-0.18			-0.44**			0.093
			(0.12)			(0.19)			(0.22)
Risk · Post	0.11	-0.083	0.10	0.20	-0.30*	0.19	-0.31***	-0.27	-0.29**
	(0.10)	(0.084)	(0.10)	(0.13)	(0.15)	(0.13)	(0.11)	(0.21)	(0.11)
High · Post			0.034			0.022			0.061
			(0.031)			(0.032)			(0.053)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	0.44	0.23	0.38	0.43	0.19	0.36	0.84	0.87	0.85
R^2	0.75	0.71	0.75	0.74	0.74	0.75	0.63	0.59	0.62
N	4,393	1,641	6,034	4,393	1,641	6,034	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in the offering of zero-premium and zero-deductible plans over county risk scores. Specifically, the outcome variable is the percent of plans with zero premium (or deductible) offered in the contract-county pair, unweighted 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 show the difference-in-difference estimates of the effect on zero Part C premiums in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate giving the differential effect on high-quality contracts. Column 4-6 repeat the analysis for zero part D premiums. Column 7-9 look at zero 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 on enrollment in plans with zero premiums or drug deductibles, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(XI)
	Zero Part C Premium			Zero Part D Premium			Zero Drug Deductible		
Risk· High · Post			-0.20			-0.52**			0.14
			(0.15)			(0.21)			(0.23)
Risk · Post	0.12	-0.13	0.098	0.23	-0.39**	0.21	-0.19**	-0.098	-0.18*
	(0.12)	(0.10)	(0.12)	(0.14)	(0.17)	(0.14)	(0.094)	(0.23)	(0.096)
High · Post			0.024			0.011			0.058
			(0.032)			(0.032)			(0.053)
Counties	15% tails			15% tails			15% tails		
Contracts	low	high	all	low	high	all	low	high	all
y mean	0.46	0.23	0.40	0.45	0.19	0.38	0.84	0.87	0.85
R^2	0.77	0.71	0.77	0.76	0.78	0.78	0.65	0.65	0.65
N	4,393	1,641	6,034	4,393	1,641	6,034	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in the share of zero-premium and zero-deductible enrollments over county risk scores. Specifically, the outcome variable is the share of a contract's zero-premium (or zero-deductible) enrollment contributed by the given contract-county pair. 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 enrollments in plans with zero Part C premiums, showing separate difference-in-differences estimates for low- and high-quality contracts. Column 3 shows the triple-difference estimate giving the differential effect on high-quality contracts. Column 4-6 (7-9) repeat the analysis focusing on enrollment in plans with zero Part D premium (drug deductible). for zero part D premiums. All regressions include contract-county fixed effects. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Table A8: Effect of QBP on the total premium (Part C and D), within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			30.51**			35.61**
			(12.54)			(14.26)
Risk · Post	-14.99*	20.99*	-14.75*	-13.49	21.64	-14.66*
	(8.73)	(11.20)	(8.66)	(8.88)	(13.37)	(8.60)
High · Post			-1.25			-1.16
			(4.86)			(4.83)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	42.93	78.55	52.69	42.44	75.47	51.42
R^2	0.85	0.85	0.87	0.85	0.86	0.87
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in total premiums over county risk scores. Column 1-2 show the difference-in-differences estimates on the premium variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 A9: Effect of QBP on drug deductibles, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-7.06			-22.48
			(46.09)			(52.21)
Risk · Post	36.54*	48.45	39.25**	33.62**	29.33	35.50**
	(18.78)	(46.03)	(19.46)	(16.87)	(51.01)	(17.26)
High · Post			-12.10			-13.42
			(10.02)			(9.52)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	30.05	24.92	28.65	28.62	24.91	27.61
R^2	0.70	0.61	0.67	0.68	0.67	0.68
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in drug deductibles over county risk scores. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 QBP on Part D premiums over income, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	per capita income (thousands)			p.c. transfer income (thousands)		
Treat · High · Post			-0.15			0.27
			(0.14)			(0.70)
Treat · Post	-0.011	-0.13	-0.0026	0.035	0.14	0.0059
	(0.033)	(0.13)	(0.034)	(0.25)	(0.66)	(0.24)
High · Post			2.50			3.33
			(2.38)			(2.18)
Counties		15% tails			15% tails	
Contracts	low	high	all	low	high	all
y mean	17.67	28.38	20.73	17.86	27.81	20.62
R^2	0.77	0.66	0.76	0.75	0.69	0.75
N	4,385	1,753	6,138	4,498	1,726	6,224

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

Notes: Table shows the within-contract variation 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-quality contracts, followed by the triple-difference estimate on high quality. We restrict locations to counties in the lower or upper 15% of the income distribution within 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 A11: Effect of QBP on Part D premiums over socio-economic status, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	white (%)			some college (%)		
Treat · High · Post			-0.15			0.075
			(0.15)			(0.066)
Treat · Post	0.021	-0.17	0.016	-0.058	0.029	-0.057
	(0.026)	(0.14)	(0.025)	(0.043)	(0.059)	(0.044)
High · Post			2.04			1.32
			(2.60)			(2.72)
Counties		15% tails			15% tails	
Contracts	low	high	all	low	high	all
y mean	17.43	28.83	20.57	17.89	28.36	20.81
R^2	0.76	0.66	0.76	0.75	0.65	0.75
N	4,479	1,703	6,182	4,729	1,833	6,562

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

Notes: Table shows the within-contract variation 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 both measures of SES. We show separate difference-in-differences estimates on low- and high-quality contracts, followed by the triple-difference estimate on high quality. We restrict locations to counties in the lower or upper 15% of the SES distribution within 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 A12: Effect of QBP on Part D premiums over market competitiveness, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
County variation in *Treat*:		HHI			risk score			quality-specific HHI	
Treat · High · Post			17.24*			21.26***			-4.28
			(10.28)			(7.67)			(7.71)
Treat · Post	1.88	18.57*	1.75	-3.09	20.16***	-2.77	3.89	-0.96	3.37
	(3.16)	(9.72)	(3.14)	(5.35)	(6.53)	(5.15)	(4.45)	(6.08)	(4.48)
High · Post			4.02*			3.63**			3.12
			(2.06)			(1.80)			(2.69)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	16.67	29.25	20.21	17.46	27.60	20.22	15.38	22.42	17.82
R^2	0.76	0.65	0.76	0.75	0.71	0.75	0.75	0.78	0.76
N	4,136	1,619	5,755	4,393	1,641	6,034	3,574	1,889	5,463

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

Notes: Table shows the within-contract variation in Part D premiums over county differences in market competitiveness (column 1-3 and column 7-9) and in risk scores (column 4-6). Competitiveness is (inversely) measured by the Herfindahl-Hirschman Index (HHI) at the county level in column 1-3, and by a quality-specific HHI restricted to contracts of the same (binary) quality in column 7-9. In each case, we show separate difference-in-differences estimates on low- and high-quality contracts, followed by the triple-difference estimate on high quality. We restrict locations to counties in the lower or upper 15% of the (quality-specific) HHI distribution within the contract's service area. Column 4-6 estimate the premium variation over county risk scores but directly control for the variation over HHIs on the right hand side: county differences in HHIs are also interacted with the Post indicator in the difference-in-differences specification, and interacted with High · Post in the triple difference. The within-contract locations in column 4-6 are from the lower and upper 15% of county risk scores as in the main analysis (Table 9). 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 QBP on Part D premiums over provider quality, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)
County variation in Treat:	hospital re-admission (%)			preventable hospital stay (%)		
Treat · High · Post			0.36*			0.39
			(0.20)			(0.28)
Treat · Post	-0.23	0.21	-0.20	-0.043	0.26	-0.053
	(0.13)	(0.15)	(0.13)	(0.18)	(0.19)	(0.18)
High · Post			2.89			2.09
			(2.00)			(2.25)
Counties	15% tails			15% tails		
Contracts	low	high	all	low	high	all
y mean	17.30	27.98	20.20	17.41	27.98	20.18
R^2	0.76	0.69	0.76	0.76	0.67	0.76
N	4,336	1,618	5,954	4,381	1,555	5,936

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

Notes: Table shows the within-contract variation 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-quality contracts, followed by the triple-difference estimate on high quality. We restrict locations to counties in the lower or upper 15% of the quality distribution within 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 A14: Effect of QBP on Part D premiums over fee-for-service (FFS) costs, within-contract variation

	(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				1.52**				1.63*				1.62**
				(0.66)				(0.84)				(0.64)
Treat · Post	-0.13	0.45	1.57**	-0.075	-0.15	0.40	1.60*	-0.13	-0.28	0.61	0.84	-0.35
	(0.29)	(0.55)	(0.65)	(0.28)	(0.36)	(0.67)	(0.81)	(0.34)	(0.51)	(0.60)	(0.52)	(0.50)
High · Post				6.77***				5.67**				5.32**
				(2.34)				(2.21)				(2.26)
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	17.60	27.48	28.27	19.22	17.46	27.72	28.27	19.05	17.52	28.30	28.93	19.23
R^2	0.75	0.76	0.74	0.75	0.75	0.69	0.75	0.76	0.76	0.66	0.75	0.77
N	4,418	1,671	787	5,205	4,459	1,661	771	5,230	4,366	1,662	766	5,132

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

Notes: Table shows the within-contract variation 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-quality 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-quality controls. We restrict locations to counties in the lower or upper 15% of the cost distribution within 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 A15: Effect of QBP on Part D premiums over binary fee-for-service (FFS) costs, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
Binary Variation in Treat:	Per Capita FFS Cost ≥85% Unadjusted (thousands)				Per Capita FFS Cost ≥85% Price-Standardized (thousands)				Per Capita FFS Cost ≥85% Price-Standardized Risk-Adjusted (thousands)			
Treat · High · Post				4.61***				4.42**				2.72*
				(1.76)				(1.86)				(1.45)
Treat · Post	-0.37	1.56	4.52***	-0.22	-0.45	0.75	4.12**	-0.38	-0.45	-0.085	1.60	-0.53
	(1.08)	(1.58)	(1.54)	(1.06)	(1.08)	(1.88)	(1.70)	(1.05)	(0.93)	(1.19)	(1.35)	(0.92)
High · Post				4.85**				3.90*				4.31*
				(2.40)				(2.31)				(2.28)
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	17.60	27.48	28.27	19.22	17.46	27.72	28.27	19.05	17.52	28.30	28.93	19.23
R^2	0.75	0.69	0.74	0.76	0.75	0.69	0.75	0.76	0.76	0.66	0.75	0.77
N	4,418	1,671	787	5,205	4,459	1,661	771	5,230	4,366	1,662	766	5,132

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

Notes: Table shows the within-contract differences in Part D premiums across binary groups of low- and high-cost counties. The cost variable is unadjusted per capita fee-for-service (FFS) 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-quality 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-quality controls. We restrict locations to counties in the lower or upper 15% of the cost distribution within 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 QBP on premiums and drug deductibles over coding-adjusted risk scores, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(XI)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			16.72			23.05***			-25.20
			(13.58)			(8.48)			(49.68)
Risk · Post	-8.60	5.22	-10.22	-8.50	17.44**	-7.87	37.71*	34.79	41.53**
	(8.53)	(12.75)	(8.23)	(5.68)	(7.30)	(5.51)	(19.59)	(48.91)	(19.11)
High · Post			-4.58			3.52*			-13.49
			(3.99)			(1.90)			(9.45)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	24.97	47.87	31.20	17.46	27.60	20.22	28.62	24.91	27.61
R^2	0.77	0.85	0.82	0.75	0.70	0.75	0.68	0.67	0.68
N	4,393	1,641	6,034	4,393	1,641	6,034	4,393	1,641	6,034

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

Notes: Table re-estimates the within-contract cross-county variation in Part C premiums (Table 8), Part D premiums (Table 9), and drug deductibles (Appendix Table A9), adjusting county risk scores with the diagnosis intensity factors developed in Finkelstein *et al.* (2017). For comparability, we restrict locations to counties in the lower and upper 15% tails of unadjusted risk scores within the contract's service area, or the original sample used in the main analysis. We show difference-in-differences estimates on low- and high-quality contracts, followed by the triple-difference estimate on high quality. 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 QBP on rebates, within-contract variation

	(I)	(II)	(III)	(IV)	(V)
Risk · High · Post				-44.41*	-71.57***
				(24.19)	(25.80)
Risk · Post	32.07*	-16.15	-45.87**	33.58*	33.10*
	(17.80)	(16.73)	(19.94)	(18.22)	(17.85)
High · Post				5.07	-0.54
				(3.76)	(4.39)
Counties			15% tails		
Contracts	low	high	high +	(2) vs. (1)	(3) vs. (1)
Service area risk			< 50%		
y mean	68.03	60.66	49.25	66.03	65.31
R^2	0.80	0.78	0.80	0.80	0.80
N	4,393	1,641	743	6,034	5,136

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

Notes: Table shows the within-contract variation 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 on the variation in low- and high-quality contracts, respectively. Column 3 restricts high-quality 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-quality (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 A18: Effect of QBP on the total premium (Part C and D), within-contract variation, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Risk · High · Post					35.61**	50.73**	63.63***
					(14.26)	(21.00)	(21.25)
Risk · Post	-13.49	21.64	38.33*	55.94**	-14.66*	-13.51	-13.46
	(8.88)	(13.37)	(20.38)	(19.92)	(8.60)	(8.73)	(8.85)
High · Post					-1.16	2.40	-1.04
					(4.83)	(5.22)	(8.62)
Counties	15% tails						
Contracts	low	high (+ service area risk)			(2)-(1)	(3)-(1)	(4)-(1)
Service area risk			<50%	<25%			
y mean	42.44	75.47	92.11	101.46	51.42	49.62	47.60
R^2	0.85	0.86	0.85	0.81	0.87	0.87	0.86
N	4,393	1,641	743	421	6,034	5,136	4,814

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

Notes: Table shows the within-contract variation 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 of the variation in low- and high-quality contracts, respectively. Column 3 restricts high-quality 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-quality contracts relative to the low-quality 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 A19: Effect of QBP on Part C premiums, within-contract variation, unweighted by enrollment

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			15.97			19.19
			(12.04)			(13.71)
Risk · Post	-11.32*	7.72	-10.75	-10.10	8.86	-10.26
	(6.65)	(11.55)	(6.69)	(7.00)	(12.28)	(6.86)
High · Post			-3.10			-3.96
			(3.44)			(3.53)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	26.92	49.55	33.12	26.73	48.21	32.57
R^2	0.76	0.80	0.79	0.76	0.82	0.79
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part C premiums over county risk scores. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-differences estimate on the differential variation in high-quality 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 A20: Effect of QBP on Part D premiums, within-contract variation, unweighted by enrollment

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			16.45***			18.53**
			(7.54)			(8.34)
Risk · Post	-5.07	15.58**	-4.15	-5.05	14.92*	-4.65
	(4.86)	(6.69)	(4.82)	(4.77)	(7.82)	(4.67)
High · Post			2.25			2.98
			(2.14)			(1.80)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	18.78	29.37	21.68	18.50	28.59	21.24
R^2	0.74	0.66	0.74	0.73	0.69	0.74
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part D premiums over county risk scores. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-differences estimate on the differential variation in high-quality 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 A21: Effect of QBP on drug deductibles, within-contract variation, unweighted by enrollment

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-25.18			-22.04
			(43.50)			(51.92)
Risk · Post	64.36**	58.54	65.77***	66.97**	63.30	67.24***
	(24.89)	(41.96)	(24.97)	(25.68)	(46.92)	(25.58)
High · Post			-10.53			-12.88
			(11.29)			(10.66)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	31.14	27.83	29.08	28.09	26.29	28.32
R^2	0.66	0.52	0.63	0.65	0.55	0.62
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in drug deductibles over county risk scores. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-differences estimate on the differential variation in high-quality 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 A22: Effect of QBP on median premiums and drug deductibles, within-contract variation

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IV)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			23.31*			17.06**			-12.97
			(13.51)			(8.33)			(52.28)
Risk · Post	-14.19*	8.70	-14.32*	-4.70	13.17*	-4.48	44.59*	48.88	44.58*
	(7.94)	(12.14)	(7.66)	(5.15)	(7.21)	(5.02)	(23.48)	(49.29)	(23.33)
High · Post			-5.27			2.62			-16.33*
			(3.88)			(1.82)			(9.33)
Counties		15% tails			15% tails			15% tails	
Contracts	low	high	all	low	high	all	low	high	all
y mean	26.15	47.28	31.89	18.64	28.66	21.36	27.23	22.67	25.99
R^2	0.73	0.83	0.78	0.73	0.67	0.73	0.64	0.57	0.62
N	4,393	1,641	6,034	4,393	1,641	6,034	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation 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 on the variation of Part C premium in low- and high-quality contracts, respectively. Column 3 shows the triple-differences estimate on the differential variation in high-quality 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 QBP on Part C premiums, within-contract variation, distance to mean

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			11.75			15.26
			(14.18)			(14.84)
Risk · Post	-10.56	1.10	-11.34*	-8.83	3.39	-10.39
	(6.86)	(13.44)	(6.67)	(6.94)	(14.25)	(6.66)
High · Post			-3.61			-4.60
			(3.76)			(3.98)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	25.17	49.94	31.96	24.97	47.87	31.20
R^2	0.77	0.85	0.82	0.77	0.85	0.82
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part C premiums over county risk scores. County differences in risk scores are measured as the distance to the mean county risk in the service area, as opposed to the distance-to-median measure in the main analysis. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 A24: Effect of QBP on Part D premiums, within-contract variation, distance to mean

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			14.64*			17.74**
			(7.66)			(8.65)
Risk · Post	-2.87	17.16**	-1.80	-4.09	16.19**	-3.65
	(4.70)	(7.16)	(4.57)	(5.16)	(7.73)	(4.96)
High · Post			2.43			3.51
			(2.29)			(1.89)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	17.77	28.60	20.74	17.46	27.60	20.22
R^2	0.76	0.67	0.75	0.75	0.70	0.75
N	14,861	5,611	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in Part D premiums over county risk scores. County differences in risk scores are measured as the distance to the mean county risk in the service area, as opposed to the distance-to-median measure in the main analysis. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 A25: Effect of QBP on drug deductibles, within-contract variation, distance to mean

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-38.25			-33.03
			(46.29)			(50.93)
Risk · Post	22.08	42.13**	44.76**	36.31**	21.34	38.37**
	(44.70)	(19.14)	(19.47)	(17.23)	(49.49)	(17.50)
High · Post			-12.03			-13.44
			(9.98)			(9.50)
Counties		all			15% tails	
Contracts	low	high	all	low	high	all
y mean	30.05	24.92	28.65	28.62	24.91	27.61
R^2	0.61	0.70	0.67	0.68	0.67	0.68
N	5,611	14,861	20,472	4,393	1,641	6,034

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

Notes: Table shows the within-contract variation in drug deductibles over county risk scores. County differences in risk scores are measured as the distance to the mean county risk in the service area, as opposed to the distance-to-median measure in the main analysis. Column 1-2 show the difference-in-differences estimates on the variation in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 A26: Effect on premiums and drug deductibles, within-contract variation, standard deviation from mean county risk

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(XI)
	Part C Premium			Part D Premium			Drug Deductible		
Risk · High · Post			16.81			16.16**			-28.31
			(13.37)			(7.94)			(50.39)
Risk · Post	-6.98	9.28	-7.41	-5.32	12.46	-4.92	60.55**	44.94	60.72**
	(8.25)	(11.54)	(8.14)	(5.13)	(7.63)	(5.02)	(23.54)	(46.46)	(24.05)
High · Post			-2.52			3.26*			-16.53
			(3.27)			(1.93)			(10.90)
Counties	distance to mean > s.d.			distance to mean > s.d.			distance to mean > s.d.		
Contracts	low	high	all	low	high	all	low	high	all
y mean	26.42	47.72	32.12	18.37	29.39	21.32	27.99	26.58	27.61
R^2	0.77	0.81	0.80	0.73	0.68	0.74	0.65	0.50	0.61
N	4,317	1,578	5,895	4,317	1,578	5,895	4,317	1,578	5,895

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

Notes: Table shows the within-contract variation in premiums and drug deductibles over county risk scores. We restrict the within-contract locations to counties more than one standard deviation away from the mean county risk in the contract's service area. Column 1-2 show the difference-in-differences estimates on the variation of Part C premium in low- and high-quality contracts, respectively. Column 3 shows the triple-difference estimate on the differential variation in high-quality 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 A27: Weight increase and the composition of measures in the overall star rating

	(I)	(II)	(III)	(IV)	(V)	(VI)
Measures in Rating	Outcome		Access		Process	
Rating · Post	0.15***	0.23***	-0.12***	0.086	-0.042	-0.022
	(0.036)	(0.073)	(0.027)	(0.011)	(0.034)	(0.086)
Rating	0.45***	0.32***	0.78***	0.59***	0.97***	0.86***
	(0.037)	(0.082)	(0.025)	(0.067)	(0.026)	(0.068)
Baseline quality	all	≥4.0	all	≥4.0	all	≥4.0
y mean	3.41	4.10	3.40	4.09	3.40	4.09
R^2	0.49	0.41	0.68	0.39	0.71	0.57
N	1,640	331	1,692	338	1,692	338

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

Notes: 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. Overall star ratings are regressed on ratings of component measures using contract-year observations of quality ratings. Column 1-2 estimate the contribution of outcome measures to the overall rating, where weights increased 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 are unchanged. Even-numbered columns show separate estimates for high-quality contracts with at least a 4.0 star rating or above in the baseline. Robust standard errors clustered at the level of contracts in the parenthesis.

Table A28: Effect of enrollee risk scores on the outcome ratings of high-quality contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
	Outcome Mean				Health Improved				Diabetes & Blood Pressure			
Treat · Post	-0.26***	-0.36***	-0.35***	-0.41***	0.026	0.058	0.042	-0.016	-0.42***	-0.56***	-0.61***	-0.70***
	(0.074)	(0.076)	(0.075)	(0.081)	(0.044)	(0.054)	(0.065)	(0.066)	(0.098)	(0.11)	(0.13)	(0.15)
y mean	3.45	3.40	3.35	3.33	3.28	3.27	3.26	3.25	3.59	3.52	3.46	3.43
Treat	high	≥4.5	≥4.5	≥4.5	high	≥4.5	≥4.5	≥4.5	high	≥4.5	≥4.5	≥4.5
Enrollee risk			>50%	>75%			>50%	>75%			>50%	>75%
Control		low quality				low quality				low quality		
Effect differences												
p-value		0.14	0.25	0.071		0.35	0.76	0.46		0.11	0.11	0.050
R^2	0.63	0.61	0.59	0.57	0.23	0.25	0.27	0.28	0.70	0.69	0.66	0.65
N	1,089	875	802	784	952	759	701	683	1,083	869	796	778

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

Notes: Table shows the difference-in-differences estimates on the outcome ratings of high-quality contracts, comparing sub-groups of high-quality contracts with different baseline enrollee risk scores. Column 1 shows the effect on the outcome ratings of high-quality contracts (rated 4.0 stars or above in the baseline) relative to the low-quality controls. Column 2 restricts high-quality contracts to higher-rated contracts rated 4.5 stars and above in the baseline. Column 3 restricts higher-rated contracts to those above the median risk score (0.977) of high-quality contracts. Column 4 restricts higher-rated contracts to those in the top 25% of risk scores (greater than 1.041) of high-quality contracts. p-values in column 2-4 test for the equality of effect sizes between sub-groups of high-quality contracts and the average high-quality contract in column 1. Column 5-8 repeat the analysis focusing on measures of self-reported health improvements from the Health Outcome Survey (HOS). Column 9-12 focus on measures of managing diabetes and hypertension conditions 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 A29: Effect of enrollee risk scores on the ratings of high-quality contracts, access and process measures

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
	Access Measures				Process Measures			
Treat · Post	-0.040	-0.049	-0.10	-0.11	-0.077*	-0.033	0.005	-0.021
	(0.042)	(0.043)	(0.063)	(0.073)	(0.043)	(0.046)	(0.058)	(0.057)
y mean	3.68	3.58	3.51	3.49	3.35	3.29	3.23	3.21
Treat	high	≥4.5	≥4.5	≥4.5	high	≥4.5	≥4.5	≥4.5
Enrollee risk			>50%	>75%			>50%	>75%
Control		low quality				low quality		
Effect differences								
p-value		0.83	0.32	0.36		0.26	0.13	0.32
R^2	0.79	0.78	0.74	0.73	0.75	0.75	0.70	0.68
N	1,122	901	828	810	1,122	901	828	810

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

Notes: Table shows the difference-in-differences estimates on the access and process ratings of high-quality contracts, comparing sub-groups of high-quality contracts with different baseline enrollee risk scores. Column 1 shows the effect on the ratings of access measures in high-quality contracts (rated 4.0 stars or above in the baseline) relative to the low-quality controls. Column 2 restricts high-quality contracts to higher-rated contracts rated 4.5 stars and above in the baseline. Column 3 restricts higher-rated contracts to those above the median risk score (0.977) of high-quality contracts. Column 4 restricts higher-rated contracts to those in the top 25% of risk scores (greater than 1.041) of high-quality contracts. p-values in column 2-4 test for the equality of effect sizes between sub-groups of high-quality contracts and the average high-quality contract in column 1. Column 5-8 repeat the analysis focusing on the process measures of the quality rating. All regressions include contract and year fixed effects. Standard errors clustered at the level of contracts in the parenthesis.

Table A30: Effect of service area risk scores on the outcome ratings of high-quality contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)
	Outcome Mean			Health Improved			Diabetes & Blood Pressure		
Treat · Post	-0.36***	-0.25***	-0.17**	0.058	0.077	0.052	-0.56***	-0.46***	-0.22*
	(0.076)	(0.082)	(0.066)	(0.054)	(0.053)	(0.060)	(0.11)	(0.15)	(0.12)
y mean	3.40	3.35	3.32	3.27	3.26	3.25	3.52	3.46	3.41
Treat	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5
Service area risk		<50%	<25%		<50%	<25%		<50%	<25%
Control		low quality			low quality			low quality	
Effect differences									
p-value		0.058	0.004		0.61	0.92		0.21	0.003
R^2	0.61	0.59	0.57	0.25	0.28	0.28	0.69	0.66	0.64
N	875	807	775	759	706	674	869	801	769

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

Notes: Table shows the difference-in-differences estimates on the outcome ratings of higher-rated contracts (4.5 stars and above in the baseline), comparing sub-groups of higher-rated contracts with different service area risk scores. Column 1 shows the effect on the outcome ratings of higher-rated contracts relative to the low-quality controls. Column 2 restricts higher-rated contracts to those below the median service area risk score (0.975) of high-quality contracts. Column 3 restricts higher-rated contracts to those in the lower 25% of service area risk scores (less than 0.902) of high-quality contracts. p-values in column 2-3 test for the equality of effect sizes between sub-groups of higher-rated contracts and the average contract in column 1. Column 4-6 repeat the analysis focusing on measures of self-reported health improvements from the Health Outcome Survey (HOS). Column 7-9 focus on measures of managing diabetes and hypertension conditions 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 A31: Effect of service area risk scores on the ratings of high-quality contracts, access and process measures

	(I)	(II)	(III)	(IV)	(V)	(VI)
	Access Measures			Process Measures		
Treat · Post	-0.049	-0.002	0.029	-0.033	-0.017	0.016
	(0.043)	(0.056)	(0.076)	(0.046)	(0.048)	(0.071)
y mean	3.58	3.51	3.48	3.29	3.23	3.20
Treat	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5	≥4.5
Service area risk		<50%	<25%		<50%	<25%
Control		low quality			low quality	
Effect differences						
p-value		0.15	0.23		0.64	0.42
R^2	0.79	0.75	0.73	0.75	0.70	0.67
N	1,122	833	801	901	833	801

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

Notes: Table shows the difference-in-differences estimates on the access and process ratings of higher-rated contracts (4.5 stars and above in the baseline), comparing sub-groups of higher-rated contracts with different service area risk scores. Column 1 shows the effect on the ratings of access measures in higher-rated contracts relative to the low-quality controls. Column 2 restricts higher-rated contracts to those below the median service area risk score (0.975) of high-quality contracts. Column 3 restricts higher-rated contracts to those in the lower 25% of service area risk scores (less than 0.902) of high-quality contracts. p-values in column 2-3 test for the equality of effect sizes between sub-groups of higher-rated contracts and the average contract in column 1. Column 4-6 repeat the analysis focusing on the process measures in the quality rating. All regressions include contract and year fixed effects. Standard errors clustered at the level of contracts in the parenthesis.

Table A32: 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: Table shows the within-contract variation 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-quality (3.0-3.5 stars), high-quality (4.0 stars and above) and high-selection (service area risk score below the high-quality 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 A33: Effect of QBP on Part D premium, within-contract variation over health-adjusted diabetes prevalence rates, high-selection contracts

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Diabetes · High · Post					100.15	88.45***	112.50***
					(64.56)	(33.70)	(34.05)
Diabetes · Post	-27.80	87.52	76.31**	88.15**	-26.48	-26.81	-26.81
	(22.09)	(57.99)	(28.19)	(38.49)	(22.18)	(22.20)	(22.24)
High · Post					2.88	5.89***	7.40***
					(2.05)	(2.25)	(2.64)
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	18.03	28.19	29.21	32.58	20.80	19.71	19.37
R^2	0.74	0.69	0.74	0.69	0.74	0.75	0.75
N	4,421	1662	784	450	6,083	5,207	4,871

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

Notes: Table shows the within-contract variation 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 on the variation in low- and high-quality contracts, respectively. Column 3 restricts high-quality 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-quality contracts relative to the low-quality 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 A34: Effect of QBP on Part D premium, within-contract variation over health-adjusted diabetes prevalence rates, high-selection and higher-rate contracts (≥4.5 stars)

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Diabetes · High · Post					67.19	121.08***	121.72***
					(62.97)	(36.53)	(35.63)
Diabetes · Post	-27.80	41.77	82.50	83.42	-26.77	-27.13	-27.06
	(22.09)	(53.63)	(46.77)	(55.12)	(22.22)	(22.18)	(22.17)
High · Post					5.10**	7.82***	6.64***
					(2.31)	(2.15)	(2.28)
Counties	15% tails						
Contracts	low	≥ 4.5 (+ service area risk)			(2) vs. (1)	(3) vs. (1)	(4) vs. (1)
Service area risk			<50%	<25%			
y mean	18.03	30.80	27.17	24.92	19.45	18.64	18.36
R^2	0.74	0.70	0.75	0.72	0.75	0.74	0.74
N	4,421	553	316	221	4,974	4,737	4,642

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

Notes: Table shows the within-contract variation 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 high-quality contracts to higher-rated contracts with 4.5 stars and above in the baseline, and 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 on the variation in low-quality and higher-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 higher-rated contracts relative to the low-quality 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 A35: Effect of QBP on Part D premium, within-contract variation over health-adjusted hypertension prevalence rates, high-selection contracts and higher-rated contracts (≥4.5 stars)

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Hypertension · High · Post					20.49	48.72*	26.01***
					(20.72)	(25.02)	(9.53)
Hypertension · Post	-5.75	13.69	41.06*	20.42	-5.24	-5.43	-5.44
	(6.60)	(18.75)	(18.74)	(10.71)	(6.56)	(6.57)	(6.58)
High · Post					5.66**	8.67***	7.71***
					(2.40)	(2.11)	(2.32)
Counties	15% tails						
Contracts	low	≥ 4.5 (+ service area risk)			(2) vs. (1)	(3) vs. (1)	(4) vs. (1)
Service area risk			<50%	<25%			
y mean	17.61	30.65	27.28	24.75	19.06	18.25	17.95
R^2	0.75	0.70	0.75	0.74	0.76	0.75	0.75
N	4,441	553	316	218	4,994	4,757	4,659

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

Notes: Table shows the within-contract variation 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 high-quality contracts to higher-rated contracts with 4.5 stars and above in the baseline, and 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 on the variation in low-quality and higher-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 higher-rated contracts relative to the low-quality 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 A36: Effect of QBP on market shares, across county risks

	(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			quality-county-year (balanced panel)		
Quality	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: Table shows the effect on the market shares of Medicare Advantage contracts over county risk scores in column 1-3, and the effect on quality-level market shares in column 4-6. Quality-level market shares are constructed for a balanced panel of quality-county-years where counties with masked enrollment data in some but not all years receive zero market shares in years with missing enrollments. Column 1-3 control for contract, county, and year fixed effects, and cluster standard errors two-way at the level of contracts and counties in the parenthesis. Column 4-6 control for county, year, and quality fixed effects, and cluster standard errors at the level of contracts in the parenthesis. Average market share (y mean) is in lower in column 4-6 due to zero market shares in the balanced panel.

Table A37: Effect of selection on the HEDIS outcomes, high-selection contracts

	(I) blood sugar controlled	(II) cholesterol controlled	(III) blood pressure controlled
		Panel A: OLS	
Risk Score	-31.32	-37.97	-47.64
	(34.30)	(35.55)	(34.32)
$\alpha_c \cdot$ Post	-0.032	5.56	2.10
		Panel B: TSLS	
Risk Score	-144.89*	-161.76***	-8.01
	(87.67)	(32.28)	(37.38)
First-stage F-stat	11.23	11.23	11.11
Over-id p-value	0.40	0.28	0.35
$\alpha_c \cdot$ Post	-2.73	-0.65	5.01
ΔRisk $\cdot\widehat{\beta_{TSLS}}$	3.62	4.04	0.20
y mean	82.49	60.46	68.50
N	226	226	228

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

Notes: Table shows the effect of risk scores on the three HEDIS outcomes, measured in percentages with the chronic condition controlled. Panel A shows the OLS estimates. Panel B shows two-stage-least-squares (TSLS) estimates where contract-level risk scores are instrumented by the correlation between Part D premiums and county risk scores in the service area. Specifically, instruments are constructed as $Corr < p_{ct-2}, R_{c0} >$, where p_{ct-2} is the Part D premiums in contract c and year $t-2$, and R_{c0} is the baseline coding-adjusted county risk scores in the service area of contract c. We construct additional instruments using correlations with health-adjusted diabetes and hypertension prevalence rates in R_{i0}. We show results for high-selection contracts – high-quality contracts with with below-median service area risk scores (<0.975), and calculate quality value-added in α_c·Post and gains from selecting lower risk types in ΔRisk $\cdot\widehat{\beta_{TSLS}}$. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. We obtain stronger first-stage prediction and similar IV estimates when the endogenous variable is plan rather than contract risk score and when we control for plan-level fixed effects and contract-level value-added. Robust standard errors clustered at the level of contracts in the parenthesis.

Table A38: 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
	(1.50)	(2.54)	(2.51)	(4.32)	(0.83)	(1.52)
$\alpha_c \cdot$ Post	0.51	0.65	0.10	0.20	0.18	0.16
			Panel B: TSLS			
Risk Score	-18.48***	-14.66*	-1.86	0.046	0.035	3.82
	(6.73)	(7.74)	(2.37)	(5.80)	(4.28)	(3.48)
First-stage F-stat	7.42	10.85	7.42	10.85	7.42	10.85
Over-id p-value	0.98	0.23	0.42	0.52	0.33	0.53
$\alpha_c \cdot$ Post	0.19	0.16	0.02	0.12	0.09	0.29
ΔRisk $\cdot\widehat{\beta_{TSLS}}$	0.46	0.53	0.05	0.00	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: Table shows the effect of risk scores on the star ratings of outcome, access, and process measures, where the dependent variable is the average rating across measures in the weighting category. 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 the OLS estimates. Panel B shows two-stage-least-squares (TSLS) estimates where contract-level risk scores are instrumented by the correlation between Part D premiums and county risk scores in the service area. Specifically, instruments are constructed as $Corr < p_{ct-2}, R_{i0} >$, where p_{ct-2} is the Part D premiums in contract c and year $t-2$, and R_{c0} is the baseline coding-adjusted county risk scores in the service area of contract c. We construct additional instruments using correlations with health-adjusted diabetes and hypertension prevalence rates in R_{c0}. We show results for high-selection contracts – high-quality contracts with with below-median service area risk scores (<0.975) in odd-numbered columns, and those in the lower 25% (<0.902) in even-numbered columns. We calculate quality value-added in $\alpha_c \cdot$Post and gains from selecting lower risk types in ΔRisk $\cdot\widehat{\beta_{TSLS}}$. We show p-values from over-identification tests. To increase statistical power, we use plan-year observations in the table. We obtain stronger first-stage prediction and similar IV estimates when the endogenous variable is plan rather than contract risk score and when we control for plan-level fixed effects and contract-level value-added. Robust standard errors clustered at the level of contracts in the parenthesis.

Table A39: Part C measures in the quality rating, 2013

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

Notes: Table lists the Part C measures in the 2013 quality rating, name of the data source, the measurement period, and the weight received by the measure in the overall star rating linked to payments.

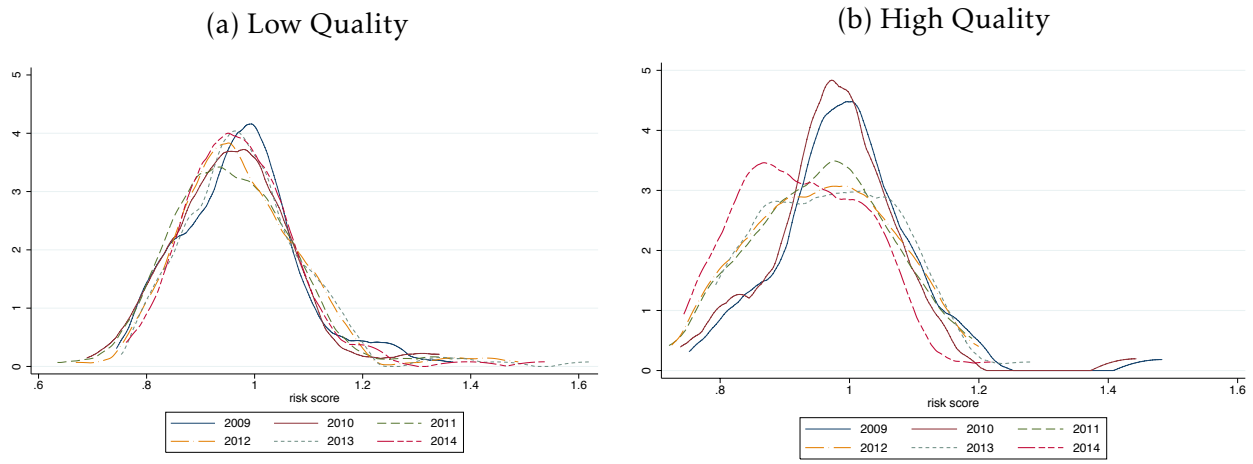
Table A40: Part D measures in the quality rating, 2013

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

Notes: Table lists the Part D measures in the 2013 quality rating, name of the data source, the measurement period, and the weight received by the measure in the overall star rating linked to payments. Measure D06-D09 in the part D domain "Member Complaints, Problems Getting Services, and Improvement in the Drug Plan's Performance" are identical to measures C30-C33 in the corresponding Part C domain; only C30-C33 are used in the computation of the final rating.

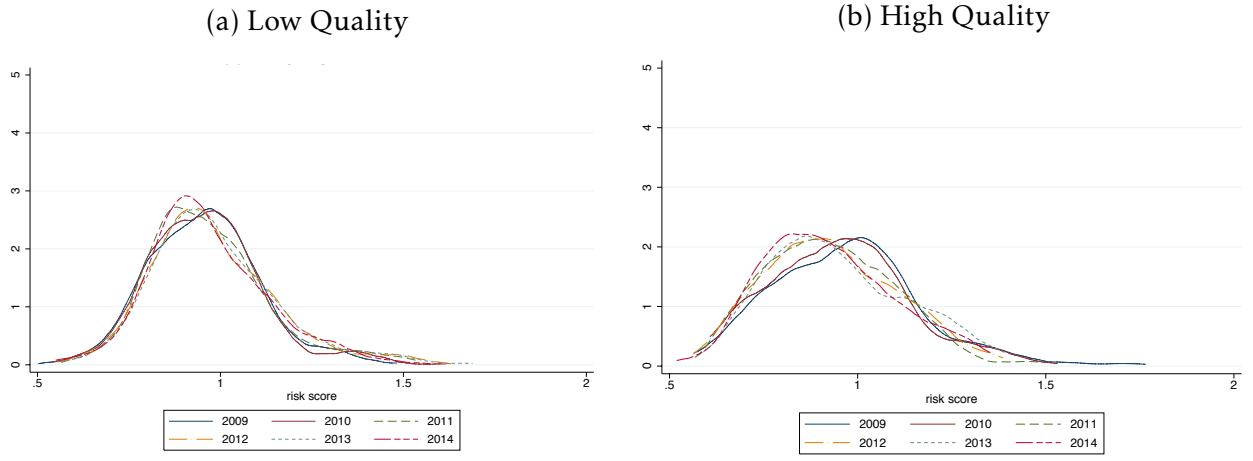
B Additional Figures

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



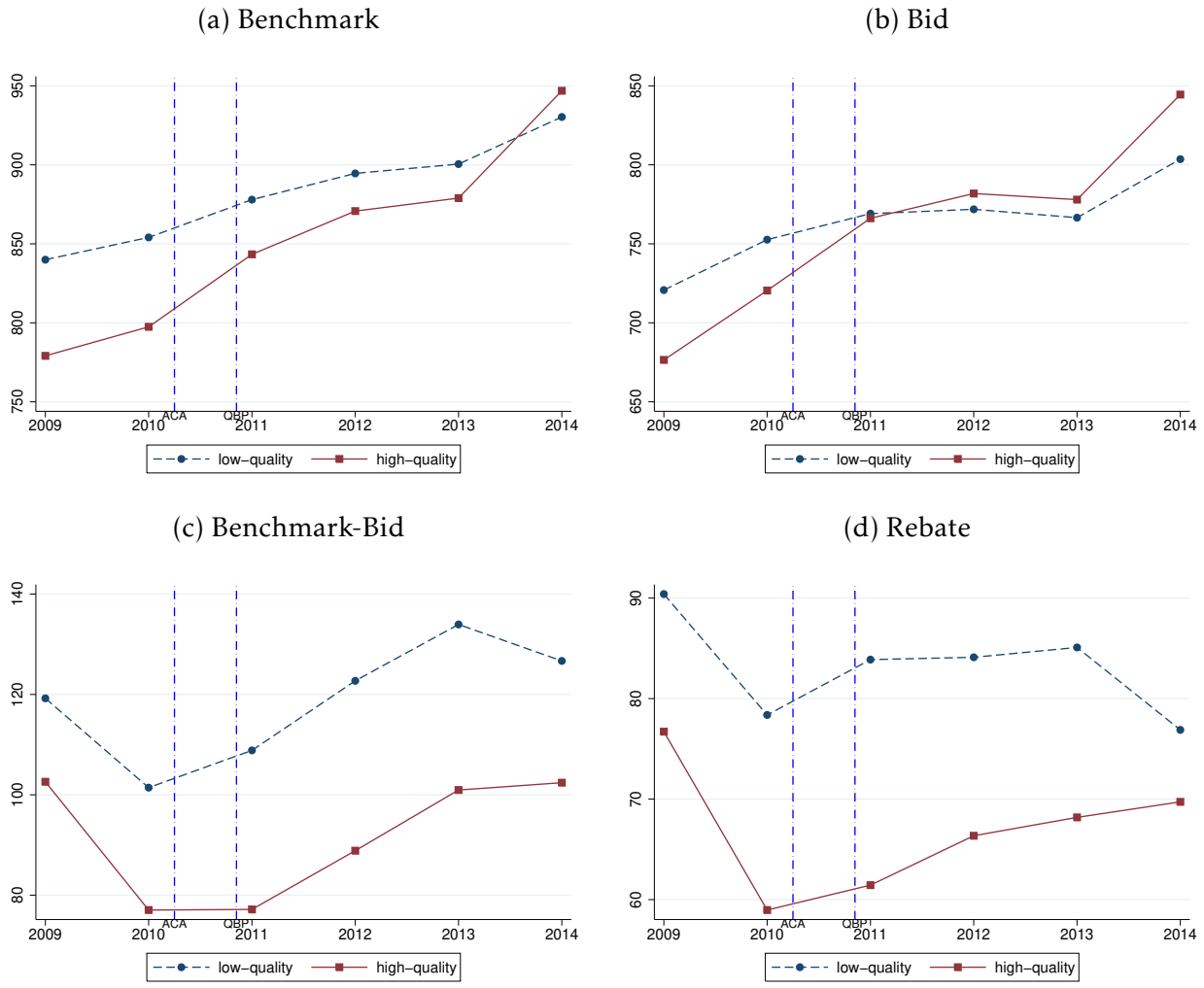
Notes. Figure plots the kernel density of high-quality risk scores in panel (a), and that of low-quality risk scores in panel (b). Separate density is drawn for each year. Risk scores are at the level of contracts aggregated from plan risk scores weighted by enrollment.

Figure B2: Plan risk scores, kernel density, by quality and year



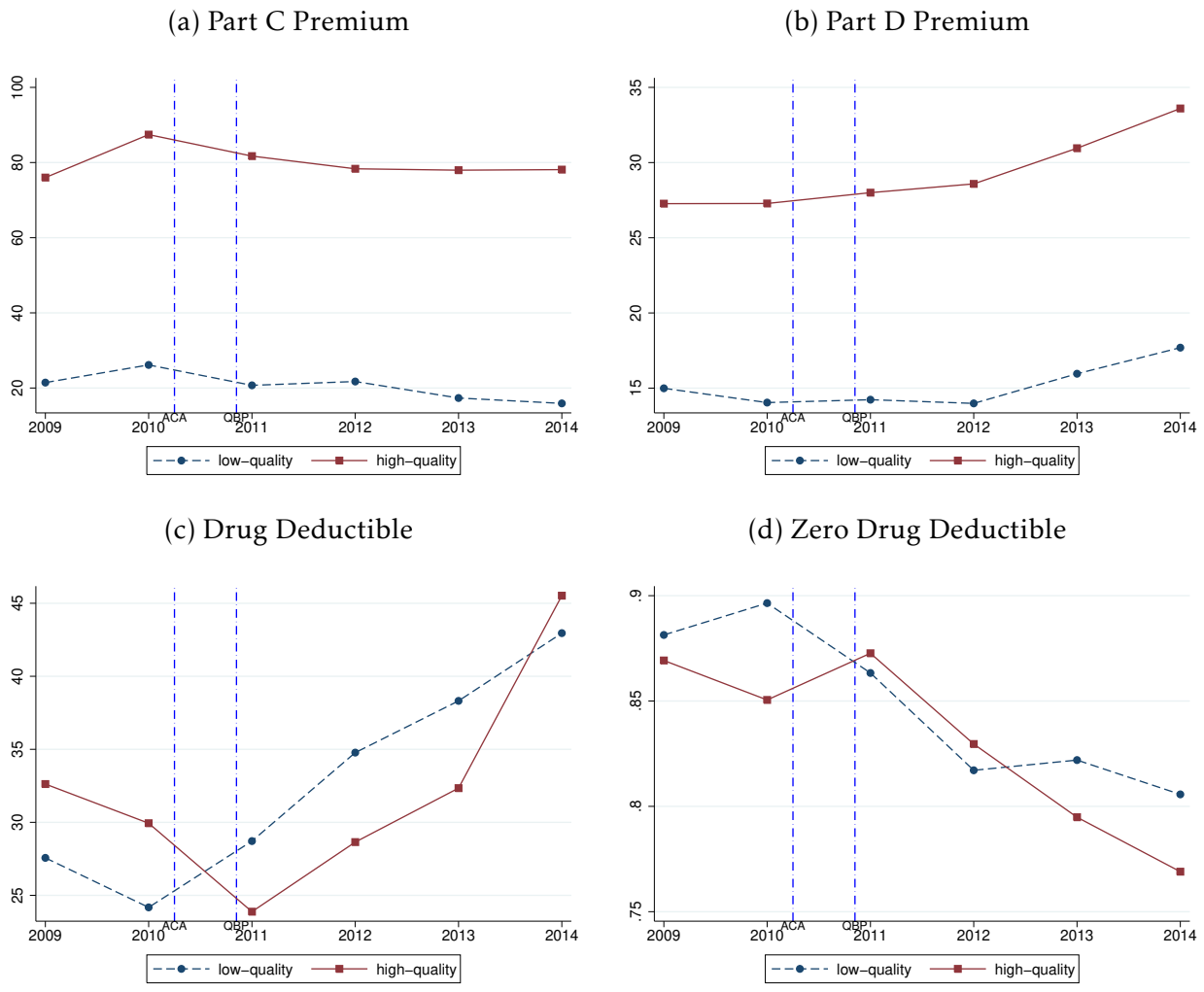
Notes. Figure plots the kernel density of plan risk scores in high-quality contracts in panel (a), and that in low-quality contracts in panel (b). Separate density is drawn for each year. Although quality is assigned at the level of contracts, this figure plots the plan risk score distributions.

Figure B3: Effect on benchmarks, bids, and rebates, raw trends



Notes. Figure shows the raw trends of benchmarks in panel (a), bids in panel (b), the difference between benchmarks and bids in panel (c), and rebates in panel (d). All variables are at the level of contract aggregated from plan variables weighted by enrollment. Quality bonus adjustments are included in benchmarks and rebates.

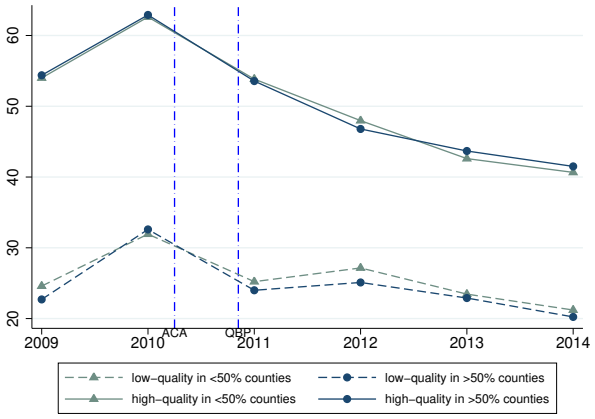
Figure B4: Effect on premiums and drug deductibles, raw trends



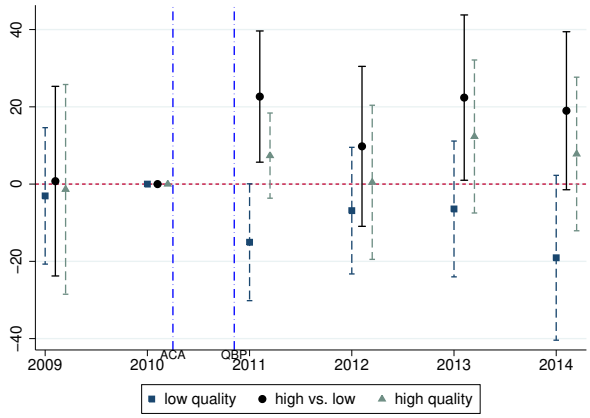
Notes. Figure shows the raw trends of Part C premium in panel (a), Part D premium in panel (b), drug deductible in panel (c), and the share of zero-deductible plans in panel (d). All variables are at the level of contract aggregated from plan variables weighted by enrollment.

Figure B5: Effect on Part C premiums, within-contract variation, event study

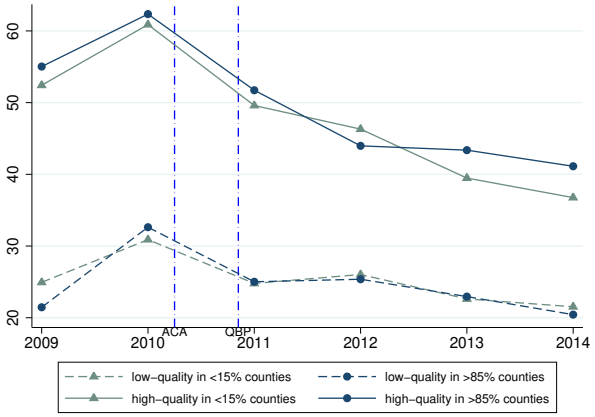
(a) Raw Trend



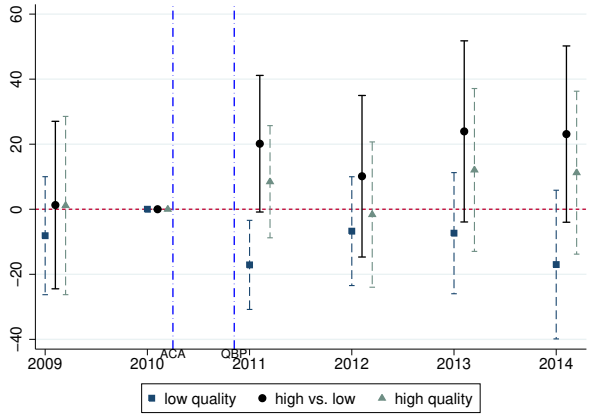
(b) Event Study



(c) Raw Trend, 15% Tails



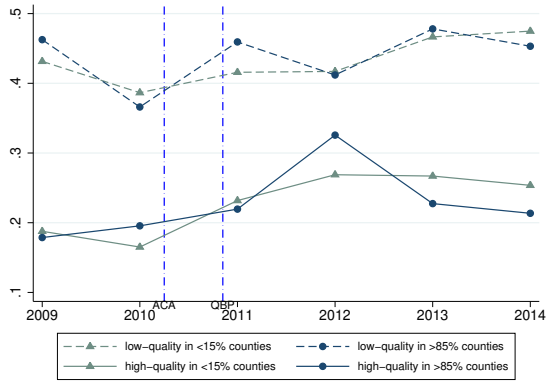
(d) Event Study, 15% Tails



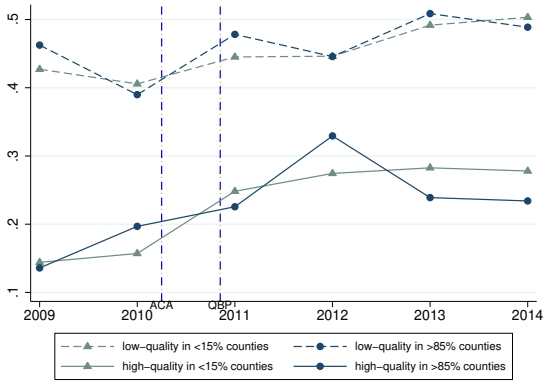
Notes. Figure plots the raw trends of Part C premiums in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality 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-quality contract (dotted lines) and an average high-quality contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 zero-premium and zero-deductible plans and enrollment, 15% tail counties, raw trend

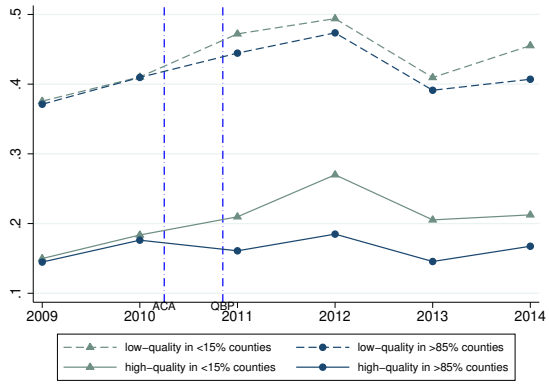
(a) Zero Part C Premium



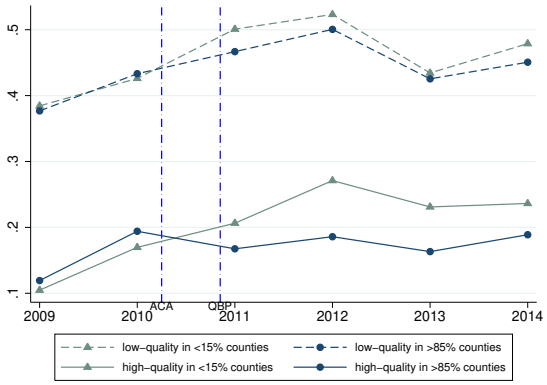
(b) Zero Part C Premium, Enrollment



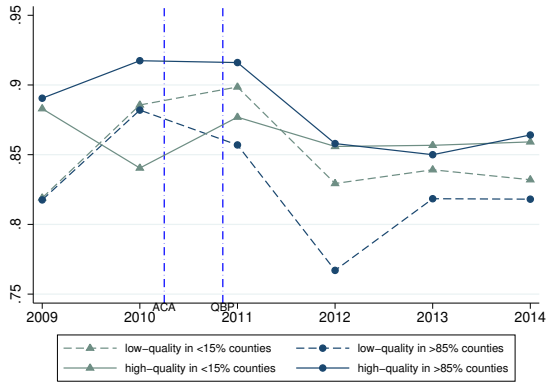
(c) Zero Part D Premium



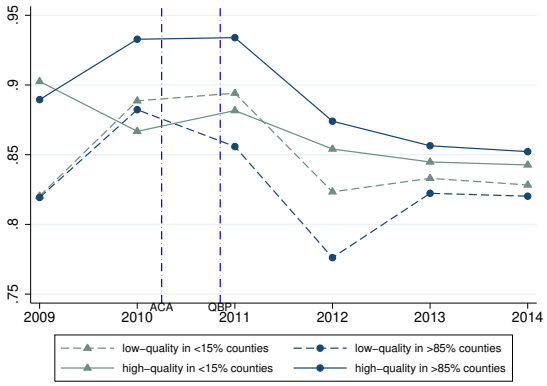
(d) Zero Part D Premium, Enrollment



(e) Zero Drug Deductible



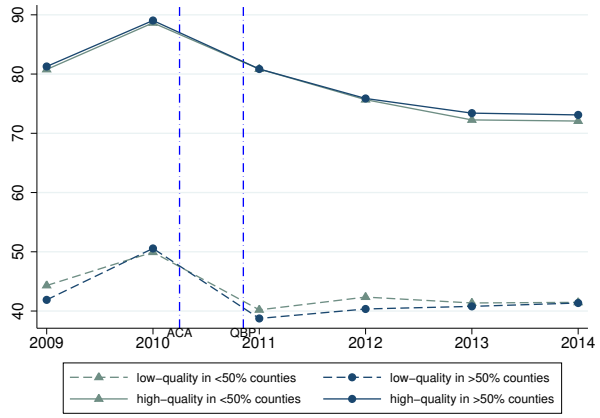
(f) Zero Drug Deductible, Enrollment



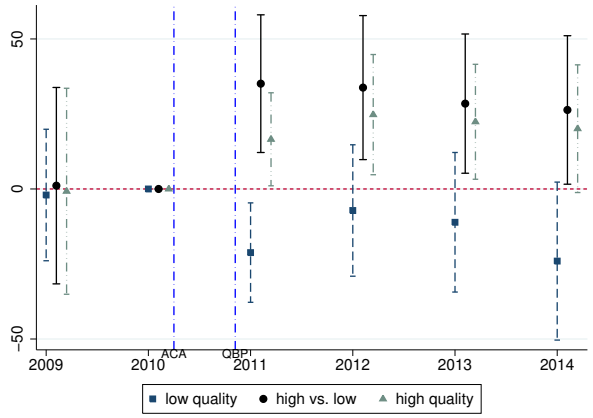
Notes. Figure plots the raw trends of zero-premium and zero-deductible plans in the left panels, and enrollments in these plans in the right panels. Specifically, outcome variables in the left panel are the percent of plans with zero premiums or drug deductible offered by the contract in a contract-county pair. Outcome variables in the right panels are the share contributed by the contract-county pair among total enrollment in zero-premium and zero-deductible plans offered by the contract. We restrict locations to counties in the lower or upper 15% tails of county risk score in the contract's service area. Each panel plots the premium levels across the 15% risk tails for an average low-quality contract (dotted lines) and an average high-quality contract (solid line).

Figure B7: Effect on the total premium (Part C and D), within-contract variation, event study

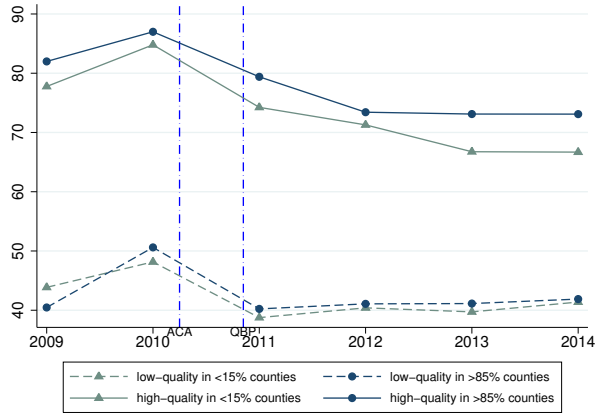
(a) Raw Trend



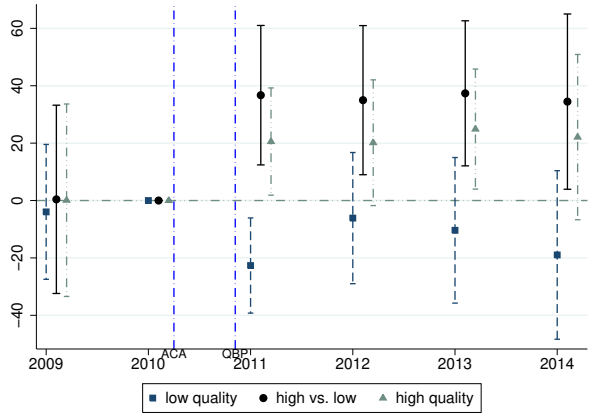
(b) Event Study



(c) Raw Trend, 15% Tails



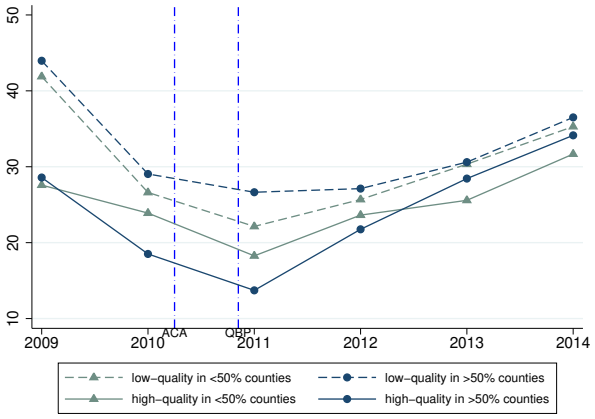
(d) Event Study, 15% Tails



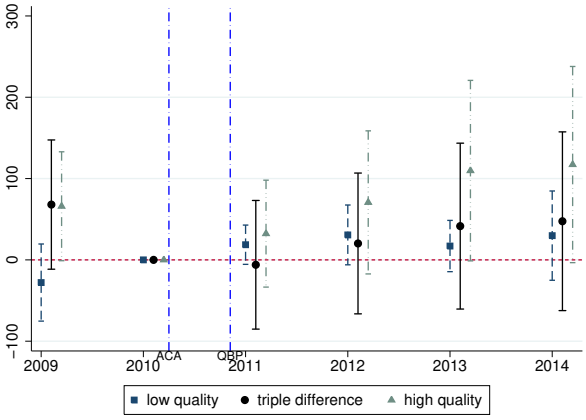
Notes. Figure plots the raw trends of total premiums in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality 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-quality contract (dotted lines) and an average high-quality contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 drug deductibles, within-contract variation, event study

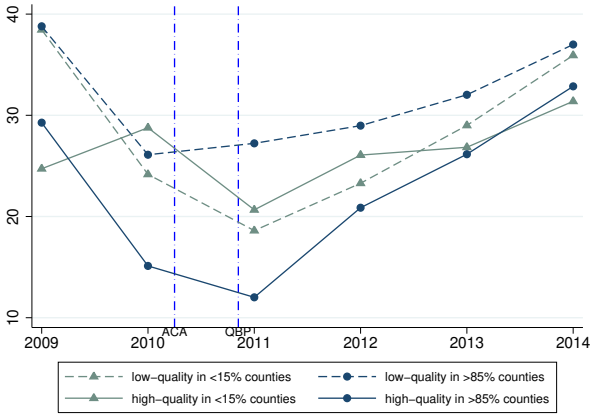
(a) Raw Trend



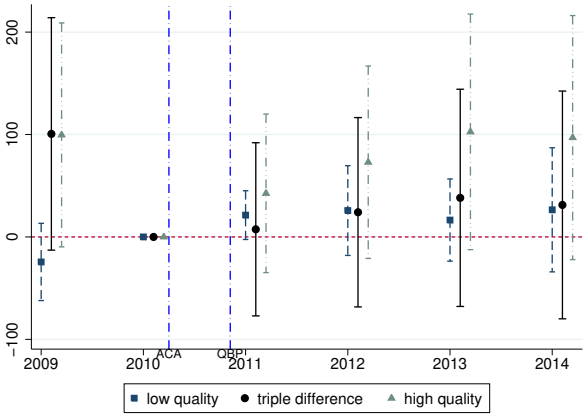
(b) Event Study



(c) Raw Trend, 15% Tails



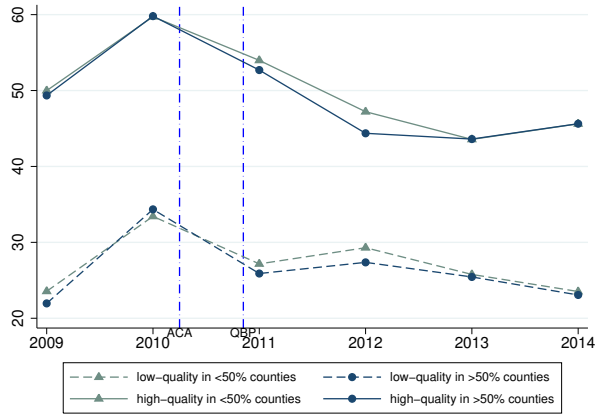
(d) Event Study, 15% Tails



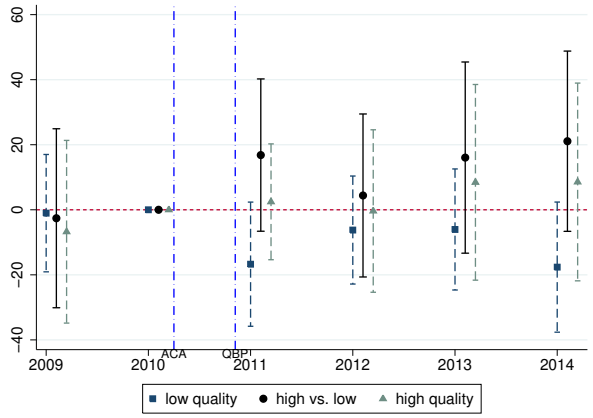
Notes. Figure plots the raw trends of drug deductibles in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. The raw trends in panel (a) plot the price levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality 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-quality contract (dotted lines) and an average high-quality contract (solid line). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 B9: Effect on Part C premiums, within-contract variation, event study, unweighted by enrollment

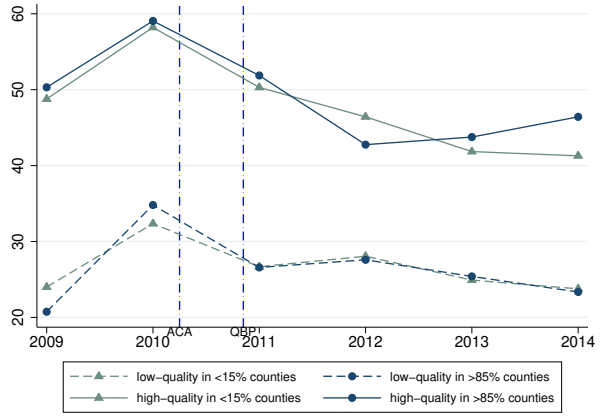
(a) Raw Trend



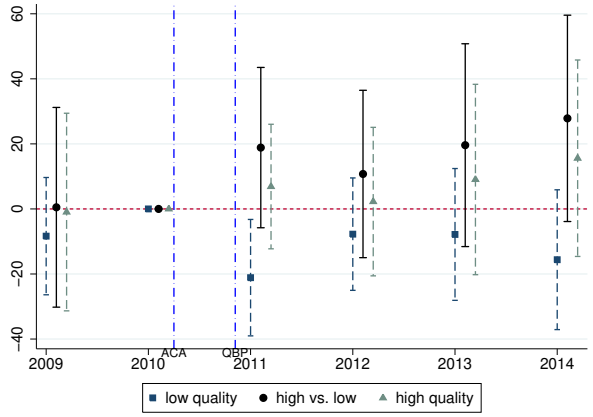
(b) Event Study



(c) Raw Trend, 15% Tails



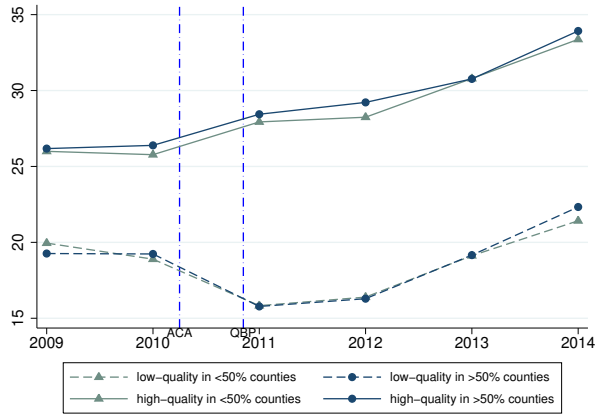
(d) Event Study, 15% Tails



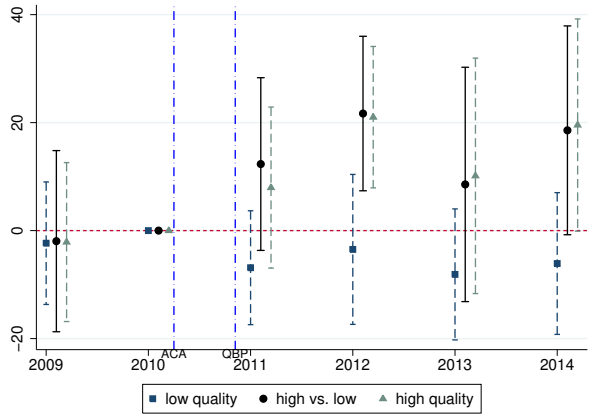
Notes. Figure plots the raw trends of Part C premiums in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score in the contract's service area, and plot premium levels across 15% tails for an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 B10: Effect on Part D premiums, within-contract variation, event study, unweighted by enrollment

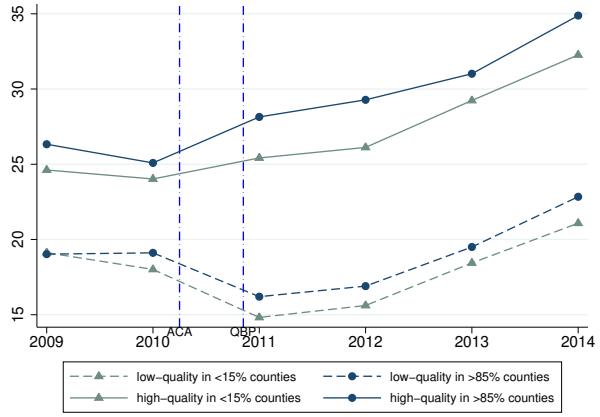
(a) Raw Trend



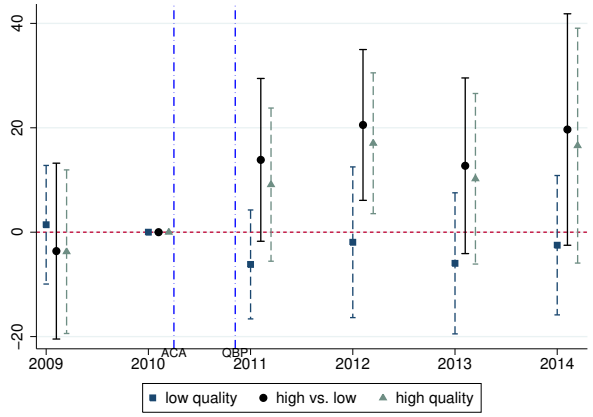
(b) Event Study



(c) Raw Trend, 15% Tails



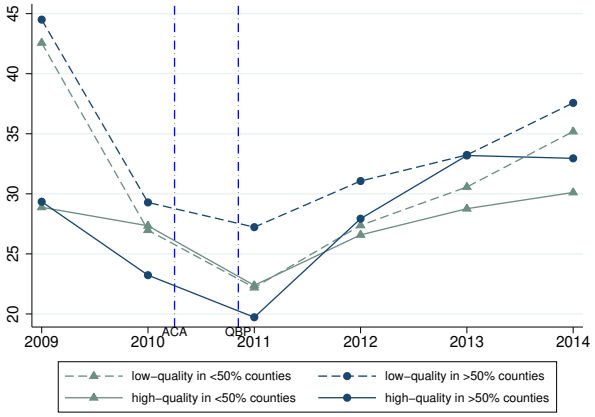
(d) Event Study, 15% Tails



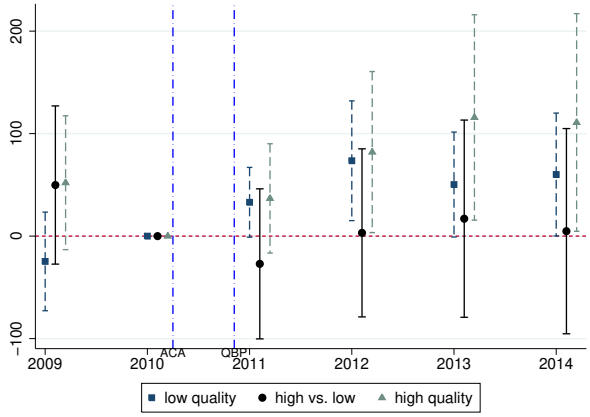
Notes. Figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. The raw trends in panel (a) plot the premium levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score in the contract's service area, and plot premium levels across 15% tails for an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 B11: Effect on drug deductibles, within-contract variation, event study, unweighted by enrollment

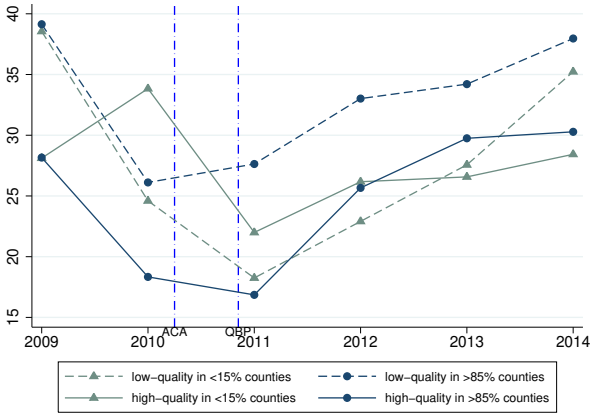
(a) Raw Trend



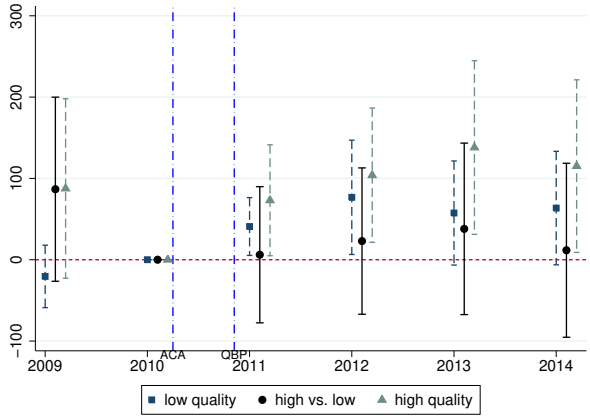
(b) Event Study



(c) Raw Trend, 15% Tails



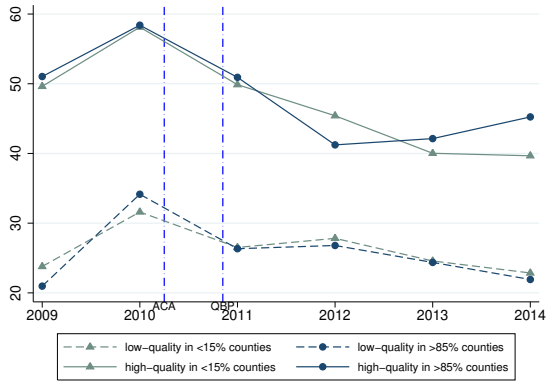
(d) Event Study, 15% Tails



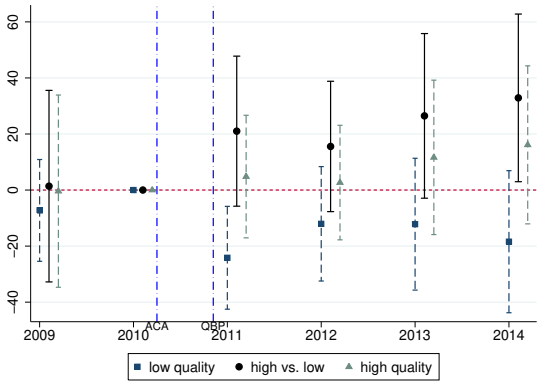
Notes. Figure plots the raw trends of drug deductibles in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. Different from the main analysis, contract-county prices are aggregated from plan prices taking simple averages, unweighted by enrollment. The raw trends in panel (a) plot the price levels above and below the median risk county within an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Panel (c) restricts the within-contract locations to the lower and upper 15% tails of county risk score in the contract's service area, and plot the price levels across 15% tails for an average low-quality contract (dotted lines) and an average high-quality contract (solid lines). Corresponding event study estimates in panel (b) and (d) show the within-contract variation 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 B12: Effect on median premiums and drug deductibles, within-contract variation, event study

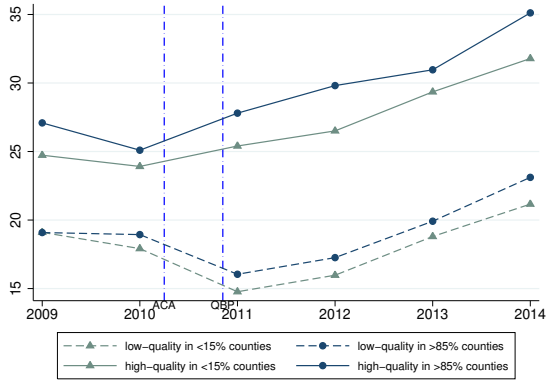
(a) Part C Premium, Raw Trend



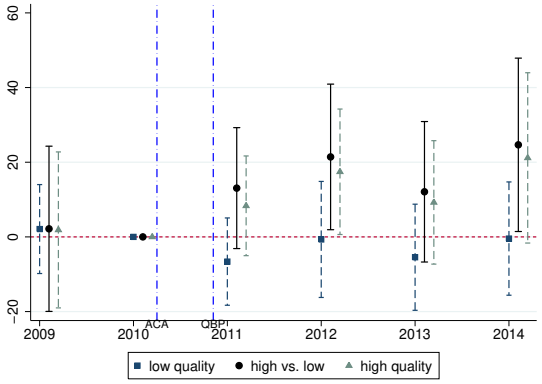
(b) Part C Premium, Event Study



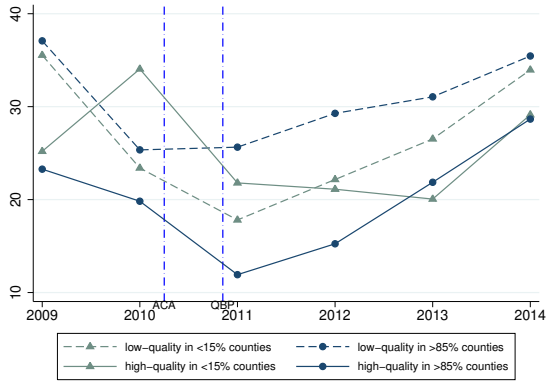
(c) Part D Premium, Raw Trend



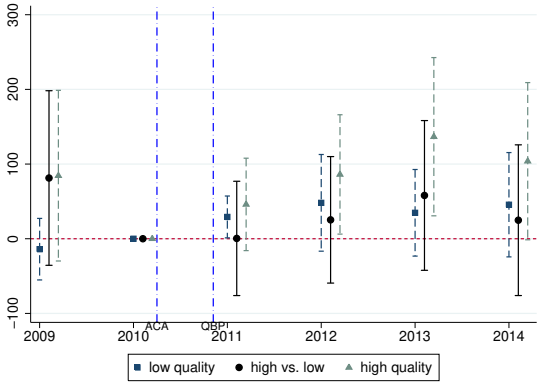
(d) Part D Premium, Event Study



(e) Drug Deductible, Raw Trend



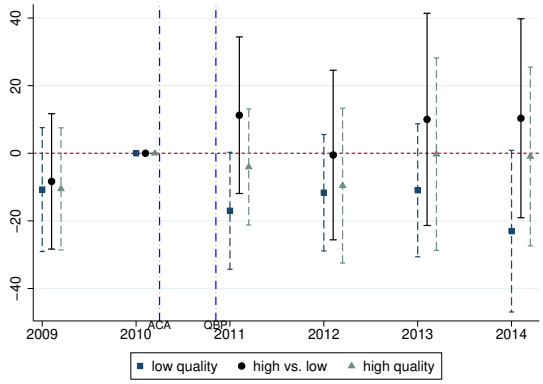
(f) Drug Deductible, Event Study



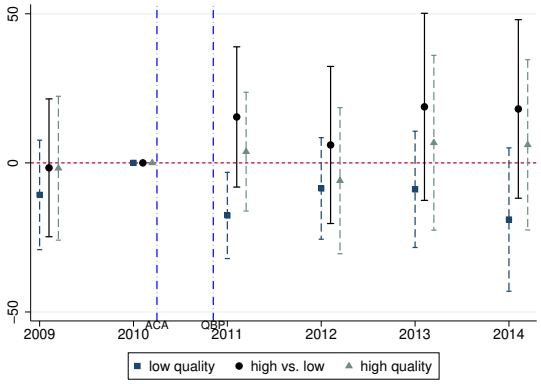
Notes. Figure plots the raw trends of premiums and drug deductibles in the left panels and event study estimates of the within-contract variation over county risk scores in the right panels. Different from the main analysis, we aggregate plan prices to the contract-county level using the median plan 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-quality contract (dotted lines) and an average high-quality contract (solid lines). Corresponding event study estimates in the right panels show the within-contract variation 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 B13: Effect on premiums and drug deductibles, within-contract variation, event study, distance to mean

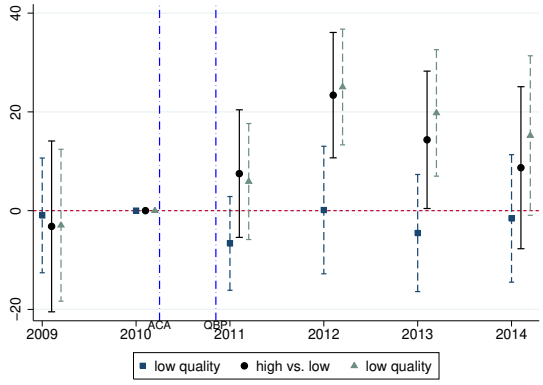
(a) Part C Premium



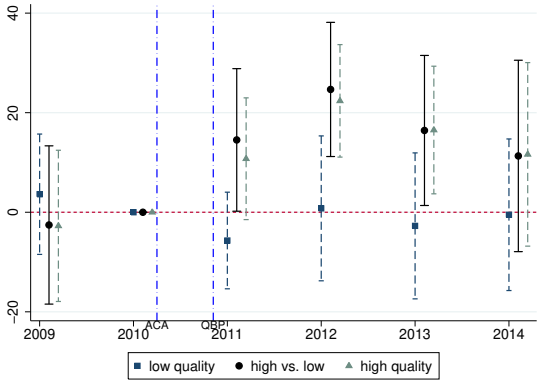
(b) Part C Premium, 15% Tails



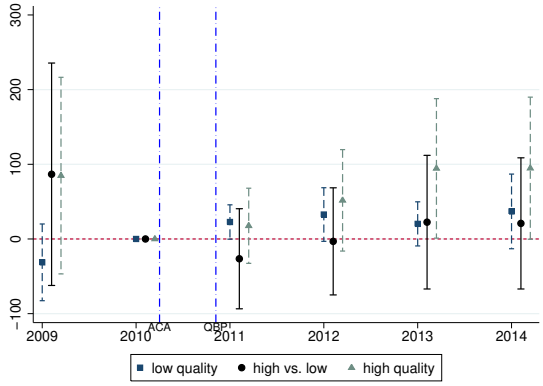
(c) Part D premium



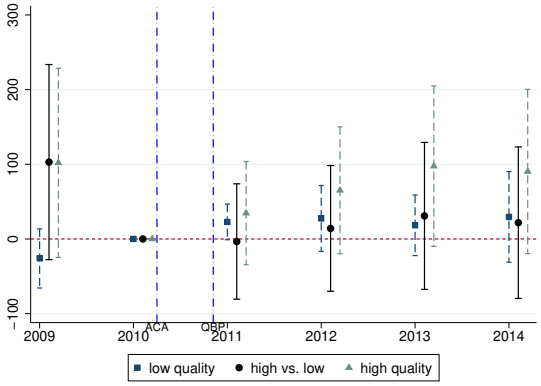
(d) Part D Premium, 15% Tails



(e) Drug Deductible

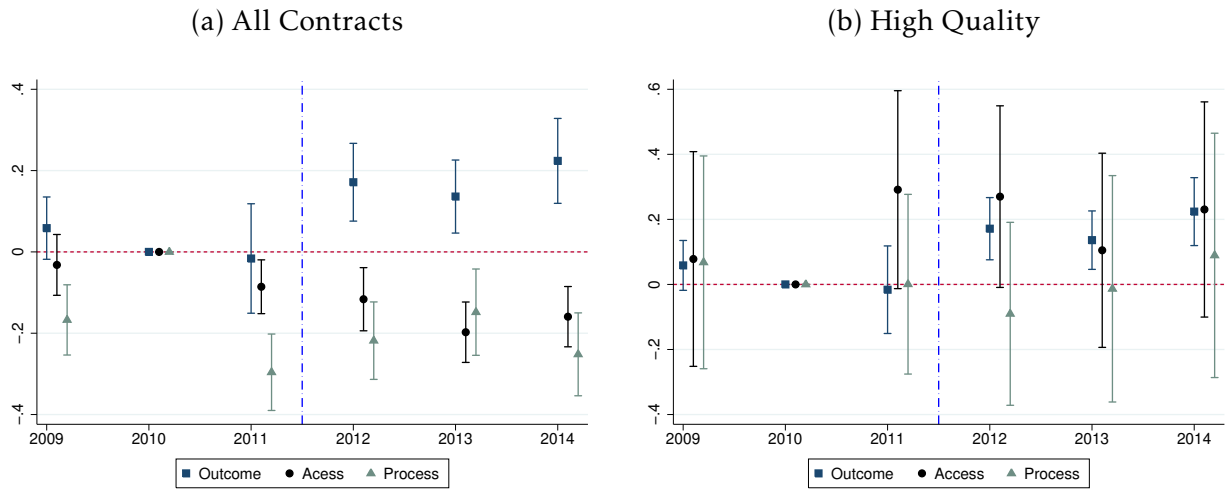


(f) Drug Deductible, 15% Tails



Notes. Figure plots the event study estimates of the within-contract variation 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 distance to the mean county risk in the service area, as opposed to the distance-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.

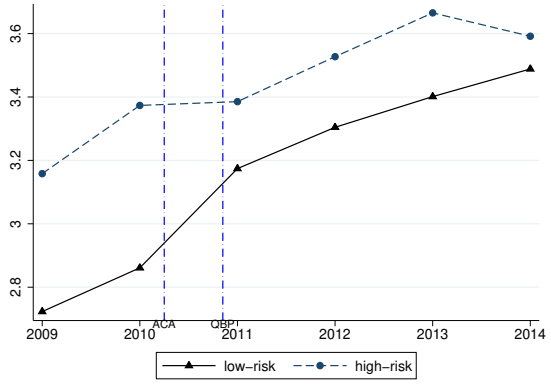
Figure B14: Weight increase and the contribution of measures to the overall star rating



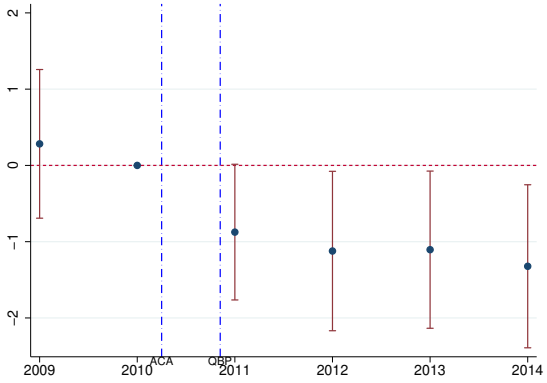
Notes. Figure shows the event study estimates of the relative contribution of outcome, access and process measures to the overall star rating. The weight increase occurred in 2012. The left panel plots the results from the full sample of contracts. The right panel shows separate results for high-quality contracts with at least a 4.0 star or above in the baseline. 95% confidence intervals are plotted based on robust standard errors clustered at the contract level.

Figure B15: Outcome ratings by baseline enrollee risk scores, event study

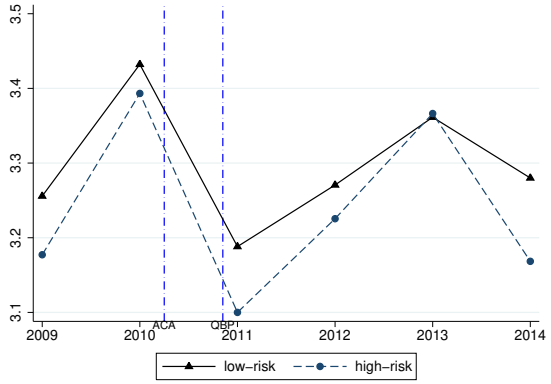
(a) Average Outcome Rating, Raw Trend



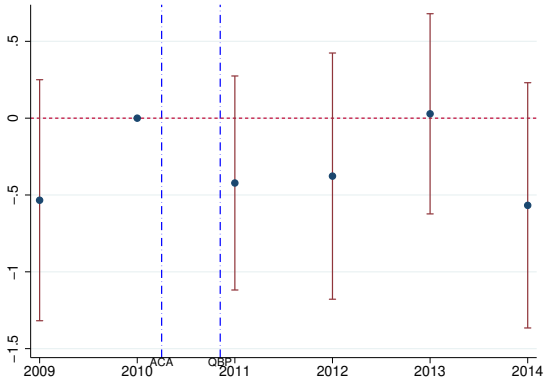
(b) Average Outcome Rating, Event Study



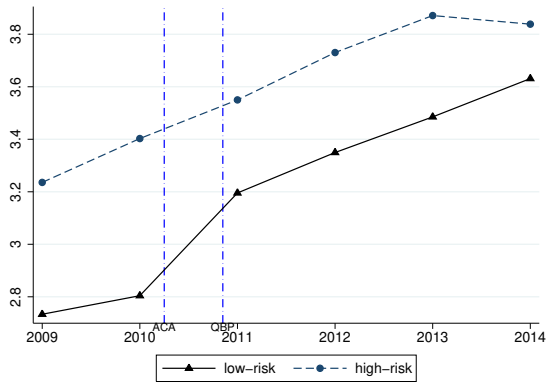
(c) Health Improved, Raw Trend



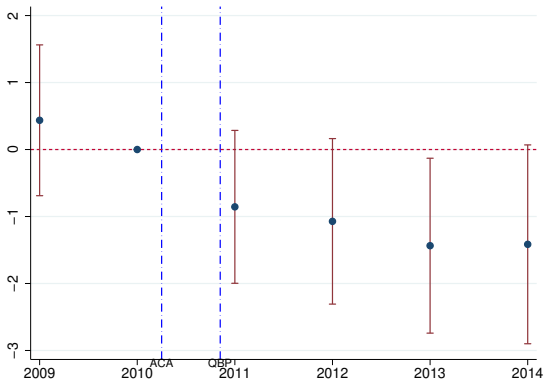
(d) Health Improved, Event Study



(e) Diabetes & Blood Pressure, Raw Trend



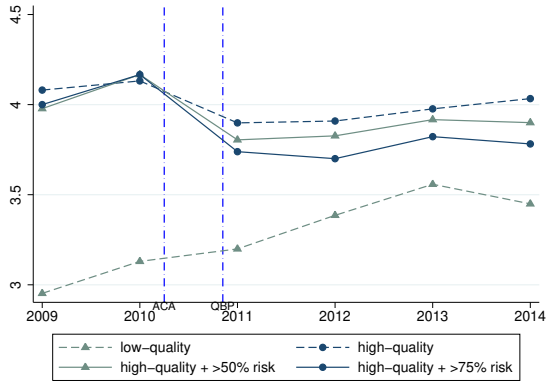
(f) Diabetes & Blood Pressure, Event Study



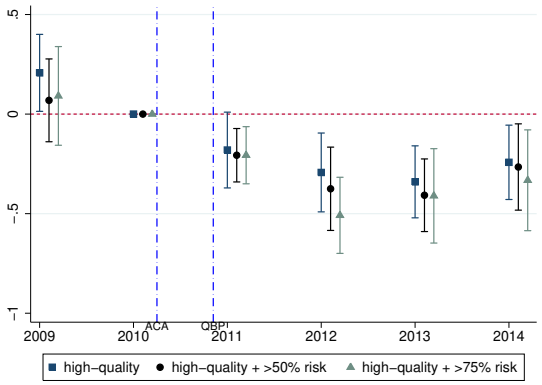
Notes. Figure shows the dynamics of outcome ratings by baseline enrollee risk scores. The raw trends in the left panels plot separate trends for binary groups of contracts above and below the median enrollee risk score (0.97) in the baseline. The right panels show event study estimates from difference-in-differences specifications over continuous variation 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 B16: Outcome ratings by quality and enrollee risk scores, event study

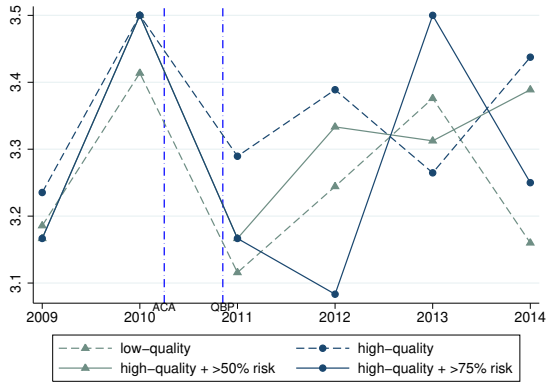
(a) Average Outcome Rating, Raw Trend



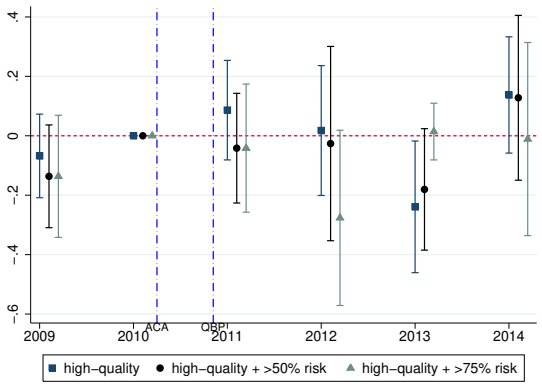
(b) Average Outcome Rating, Event Study



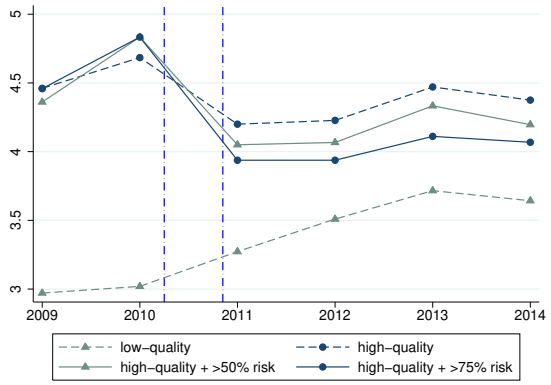
(c) Health Improved, Raw Trend



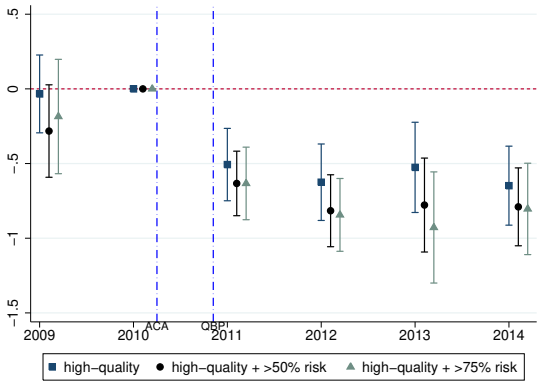
(d) Health Improved, Event Study



(e) Diabetes & Blood Pressure, Raw Trend



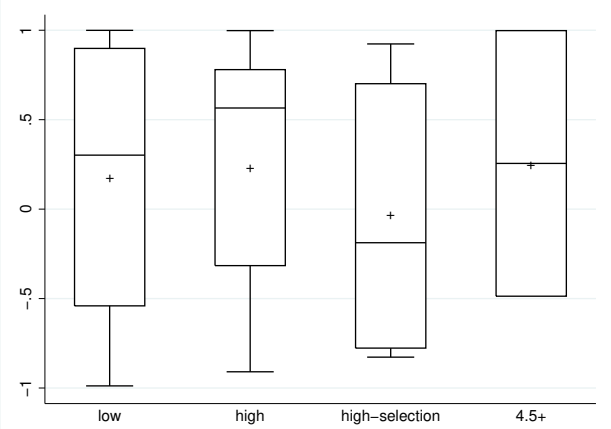
(f) Diabetes & Blood Pressure, Event Study



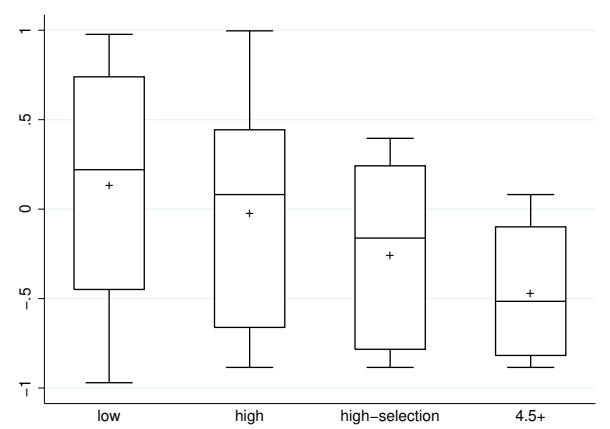
Notes. Figure shows the dynamics of outcome ratings by baseline quality and enrollee risk scores. High-quality contracts are restricted to those rated 4.5 stars and above in the baseline, or the higher-rated contracts. The left panels plot separate trends for sub-groups of higher-rated contracts with enrollee risk scores above the median (0.977) and in the upper 25% (greater than 1.041) of high-quality contracts, and for low-quality contracts. The right panels show the event study estimates from difference-in-differences specifications comparing ratings in (sub-groups of) higher-rated contracts with the low-quality controls. Panel (a) and (b) look at the average ratings 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 B17: Within-contract correlations between risk scores and overall star ratings

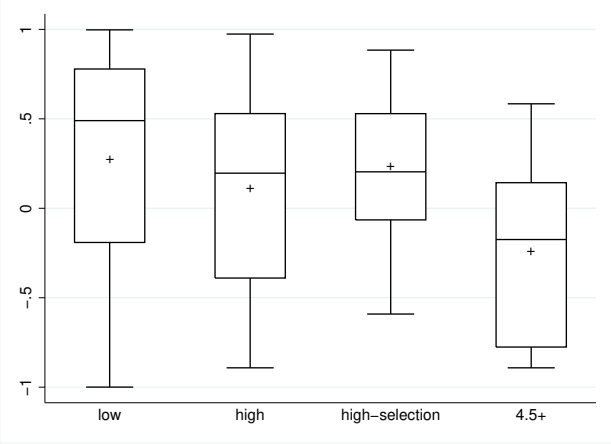
(a) $Corr(Star_t, Risk_{t-3})$



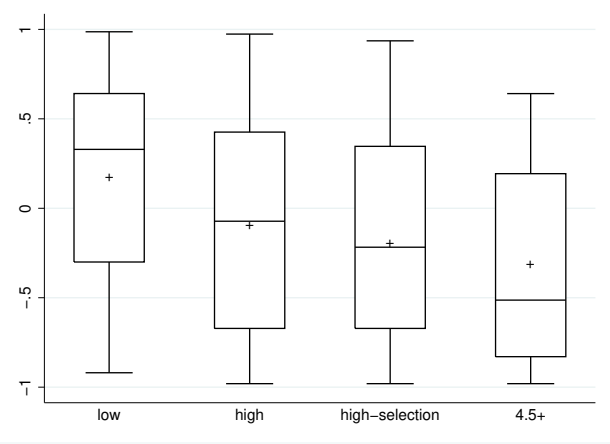
(b) $Corr(Star_t, Risk_{t-2})$



(c) $Corr(Star_t, Risk_{t-1})$



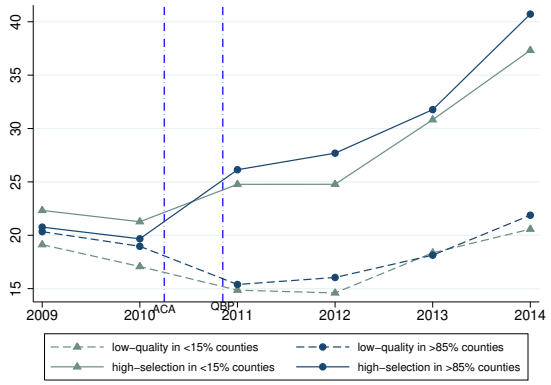
(d) $Corr(Star_t, Risk_t)$



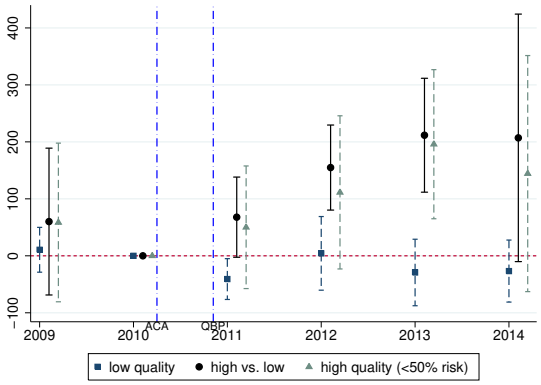
Notes. Figure plots the distribution of correlation coefficients between overall star ratings in year t and risk scores in year $t-3$ (panel a), $t-2$ (panel b), $t-1$ (panel c), and year t (panel d). The coefficients are calculated within contracts over time, grouped by the baseline contract quality. The middle line represents the median coefficient. The lower and upper edges of the box bound the 25th-75th percentile. The antennas reach the nearest observations outside. The plus (+) marks the mean coefficient.

Figure B18: Effect on Part D premium, within-contract variation over health-adjusted diabetes prevalence rates, high-selection and higher-rated contracts (≥4.5 stars), event study

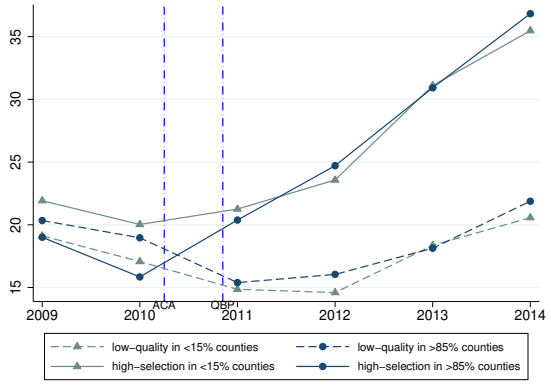
(a) High-Selection (≥4.5 Stars, <50% Risk), Raw Trend



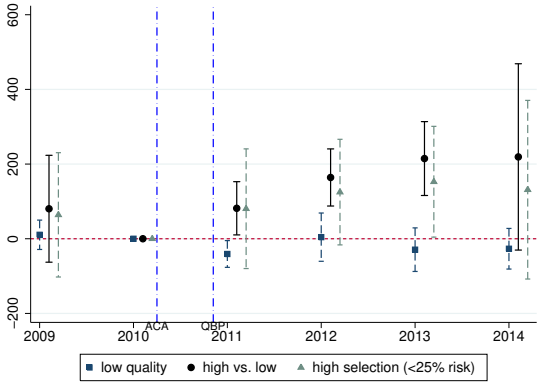
(b) High-Selection (≥4.5 Stars, <50% Risk), Event Study



(c) High-Selection (≥4.5 Stars, <25% Risk), Raw Trend



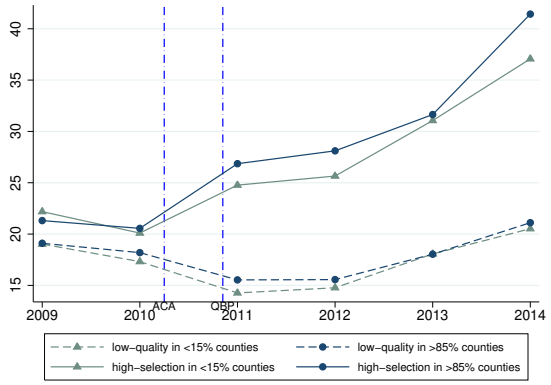
(d) High-Selection (≥4.5 Stars, <25% Risk), Event Study



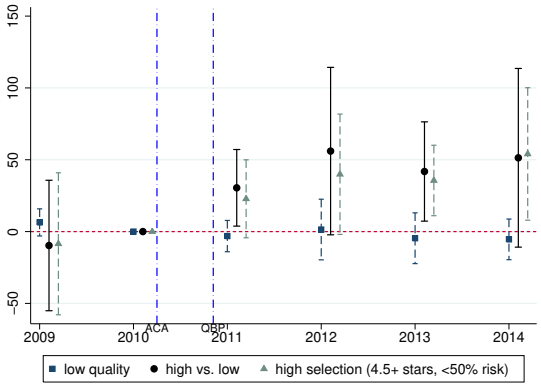
Notes. Figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract variation over county differences in health-adjusted diabetes prevalence rates in the right panels. The health-adjusted prevalence rate multiplies the raw prevalence rate by the coding-adjusted 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. We restrict high-quality contracts to higher-rated contracts (4.5 stars and above), and plot the price levels across the 15% tails within an average low-quality contract (dotted lines) and an average higher-rated, 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 variation 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 B19: Effect on Part D premium, within-contract variation over health-adjusted hypertension prevalence rates, high-selection and higher-rated contracts (≥4.5 stars), event study

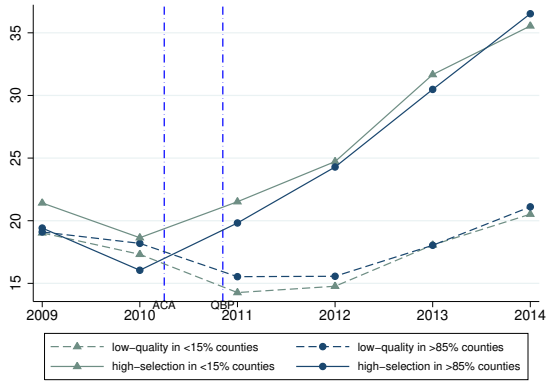
(a) High-Selection (≥4.5 Stars, <50% Risk), Raw Trend



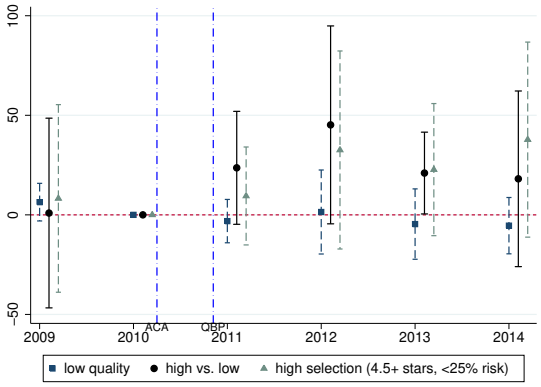
(b) High-Selection (≥4.5 Stars, <50% Risk), Event Study



(c) High-Selection (≥4.5 Stars, <25% Risk), Raw Trend



(d) High-Selection (≥4.5 Stars, <25% Risk), Event Study



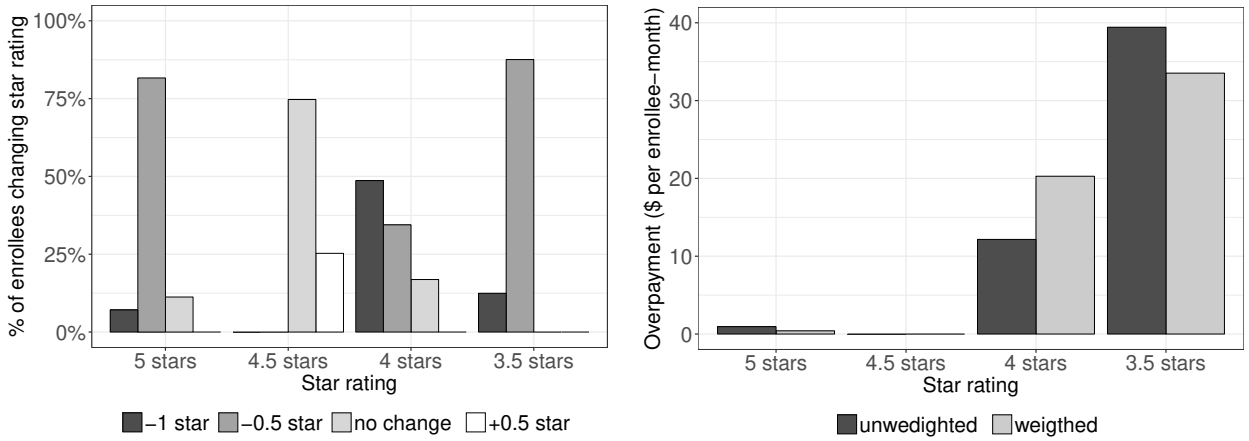
Notes. Figure plots the raw trends of Part D premiums in the left panels and event study estimates of the within-contract variation over county differences in health-adjusted hypertension prevalence rates in the right panels. The health-adjusted prevalence rate multiplies the raw prevalence rate by the 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. We restrict high-quality contracts to higher-rated contracts (4.5 stars and above), and plot the price levels across the 15% tails within an average low-quality contract (dotted lines) and an average higher-rated, 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 variation 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 B20: 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. 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-quality contract using a weighted average of low-quality 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 the payments across 2014 star ratings. 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 behavioral assumption is supported by our empirical analysis of bidding and pricing strategies at the contract level. 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.

C Data Appendix

C.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 C1 summarize the number of plan-county observations in the raw files, and the remaining sample after dropping regional PPOs and Part C only plans.

Payments to plans and rebates to enrollees 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 quality data due to small enrollments, a fixed star rating is applied to all such

contracts to determine benchmark and rebate bonuses.[52] 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 baseline contracts rated 3.0 stars and above continue to receive quality ratings over the sample period.[53] Data on measure ratings and overall ratings are available at https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html. The crosswalk file linking plans and contracts over time is available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MCRAdvPartDEnrolData/Plan-Crosswalks.html.

We merge in enrollment counts at the plan-year-county level from monthly enrollment counts from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MCRAdvPartDEnrolData/Monthly-Enrollment-by-Contract-Plan-State-County.html. Annual enrollment sums over enrollee-months over a 12-month period. However, exact counts are masked for counties with fewer than 10 enrollees. We include the full range of service areas when constructing the within-contract variation 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 achieved at least a 4.0 star overall rating in 2009 or 2010. 149 are rated at least 3.0 stars but less than 4.0 stars in both years. The remaining contracts have at least one rating below 3.0 stars. 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 variation in prices. Details of the county variables are provided

[52] 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.

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

Table C1: 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 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			
baseline below 3.0 stars			54			
baseline low quality			135			
baseline high quality			55			
high-selection (<50% FFS risk)			27			
≥4.5 stars			19			
≥4.5 stars, <50% FFS risk			11			

Notes: Table shows the step-by-step construction of the estimation sample from yearly Landscape Files. The main analysis further excludes contracts with less than 3.0 star rating in either year of 2009-2010. Counts of distinct contracts in sub-groups of high-quality contracts are listed at the end.

below.

C.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.[54] 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 variation in prices. The first four rows of Appendix Table C2 summarizes the FFS risk scores and costs by county.

Table C2: 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	0.96	0.002	1,465	Per capita income (k)	31.87	0.20	1,450
Per capita FFS Cost (k)	8.40	0.035	1,465	Per capita transfer income (k)	7.33	0.039	1,450
price adjusted (k)	8.08	0.031	1,465	Non-White (%)	11.85	0.36	1,465
price-risk adjusted (k)	8.62	0.023	1,465	Some college (%)	37.68	0.27	1,465
Diabetes (%)	8.88	0.056	1,465	HHI	0.55	0.005	1,461
Hypertension (%)	37.68	0.13	1,465	Low-quality HHI	0.76	0.007	1,164
Hospital re-admission (%)	17.86	0.070	1,462	High-quality HHI	0.88	0.009	502
Preventable hospitalization (%)	7.05	0.065	1,454				

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

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

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 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-specific HHI for markets at the level of county-quality pairs.

D Distributional Impacts: Additional Evidence

Following the discussion in Section 7.2, we provide additional evidence on premium and market share changes in the upper and lower 15% tails of county risk. We focus on the extreme tails of county risk because the within-contract variation tends to imply decreasing (increasing) premium in very low-risk (high-risk) counties. In the intermediate range of county risk, premium can either increase or decrease depending on the within-contract ranking of county risks and the distribution of contracts across counties.

Table E1 estimates the effect on premiums separately for the two risk tails.[55] We expect the effects to be asymmetric since high-quality contracts in the lower risk tail are predominantly high-selection contracts,[56] whereas in the upper risk tail, high-selection contracts account for only 10% of the high-quality sample. Consistent with larger within-contract variation in high-selection contracts, high-quality premiums decreased significantly with county risk in the lower tail (column 2), but not by much in the highest-risk counties (column 5). Figure E1 shows the event study.

Table E2 estimates the premium variation over pooled county risk scores across 15% tails in column 1-3, and the effects on market shares in column 4-6. These estimates underlie the diverging patterns of market shares shown in Figure 10. We focus on high-selection contracts in the table. Premiums in high-selection contracts increased significantly with county risk scores, with larger increases in the lower risk tail. The increase in low-quality premiums is negligible. In both quality, market shares decreased with county risk, and decreased much more in high-quality contracts. Moving from the bottom 15% to the top 15% risk tail increases high-quality premiums relative to low-quality by $6.51 after the QBP demonstration, and decreases high-quality market shares by 12.96 percentage points. Figure E2 shows the event study.

D.1 Contemporaneous Quality

Estimating aggregate effects using a sample of continuing contracts and their baseline ratings can be problematic for two reasons. First, future ratings of continuing ratings

[55] We use the specification

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

where outcomes vary at the level of contract (c), county (l) and year (t). $risk_l$ is the county risk score.

[56] Counties in the lower 15% tail (risk scores below 0.872) are included in the service area of 31 high-quality contracts and 57 low-quality contracts. Of the 981 contract-county-year observations in the high-quality sample, 903 are contributed by high-selection contracts.

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

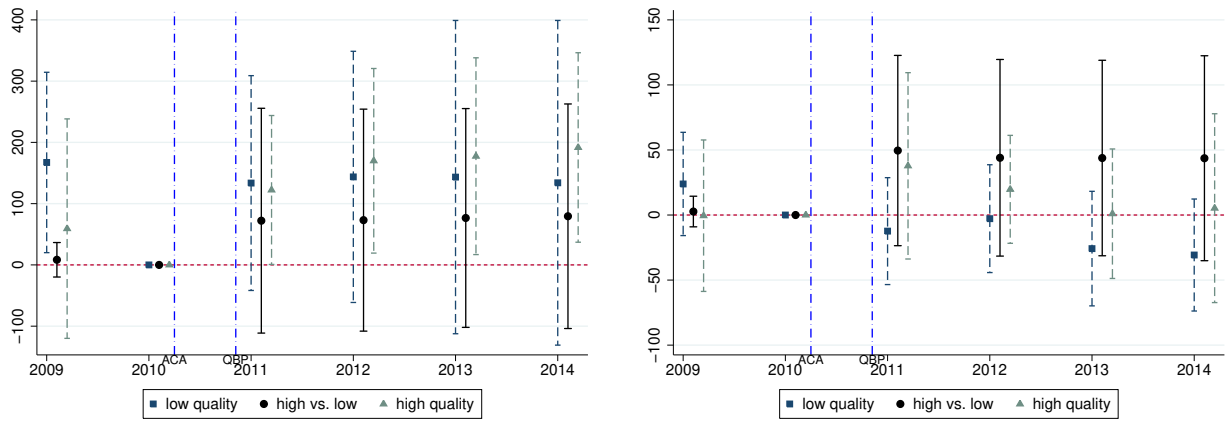
	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			71.47			44.05
			(93.52)			(37.27)
Risk · Post	59.94	134.74**	58.77	-29.16*	20.28	-32.23**
	(85.75)	(54.83)	(77.81)	(15.03)	(31.51)	(14.88)
High · Post			-53.14			-55.93
			(82.43)			(52.23)
Risk · High			-54.81			-107.69*
			(94.79)			(59.36)
Counties		<15% risk			>85% risk	
Contracts	low	high	all	low	high	all
y mean	49.75	93.63	67.73	40.11	72.45	47.03
R^2	0.83	0.82	0.86	0.81	0.76	0.80
N	1,264	981	2,293	3,883	1,057	4,946

Notes: Table estimates the premium variation over county risk scores in the lower 15% risk tail (<0.87205) in column 1-3, and in the upper 15% risk tail (>1.0295) in column 4-6. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

Figure E1: Premium variation over county risk scores, lower and upper 15% risk tails

(a) Lower 15% Risk Tail

(b) Upper 15% Risk Tail

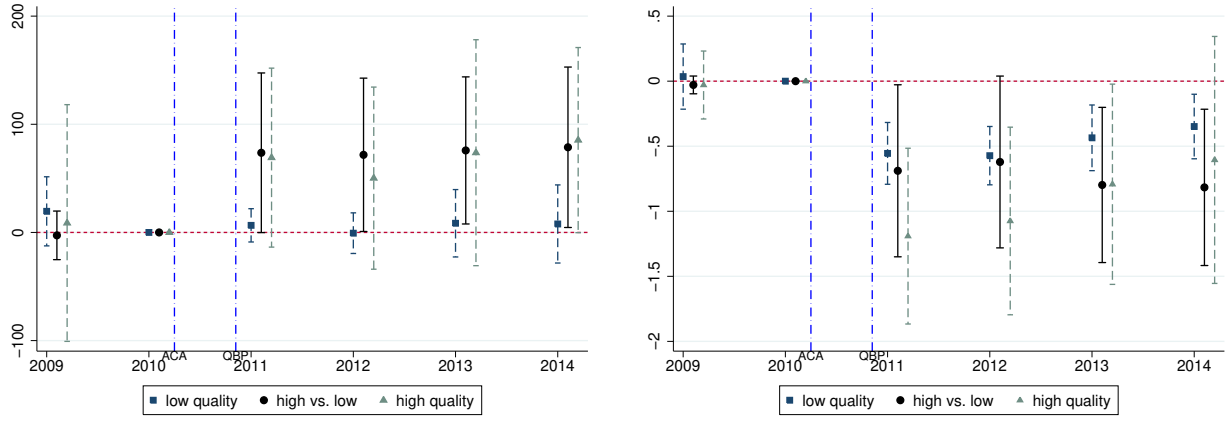


Notes. Figure plots the event study estimates of the premium variation over county risk scores in the lower 15% risk tail (<0.87205) in panel (a), and in the upper 15% risk tail (>1.0295) in panel (b). In each panel, the dotted lines on the left (right) plot the variation over time for low- (high-) quality contracts, and the solid line in the middle shows the differential effect on high-quality contracts. 95% confidence intervals are plotted based on robust standard errors clustered two-way at the level of contracts and counties.

Figure E2: Effect on premiums and market shares across the 15% risk tails

(a) Premium

(b) Market Share



Notes. Figure plots the event study estimates of the variation in premiums (panel a) and market shares (panel b) over county risk scores across the 15% risk tails (below 0.87205 or over 1.0295). We restrict high-quality contracts to the high-selection contracts below the median service area risk (0.975) of high quality. 95% confidence intervals are plotted based on robust standard errors clustered two-way at the level of contracts and counties.

Table E2: Effect on premiums and market shares across the 15% risk tails

	(I)	(II)	(III)	(IV)	(V)	(VI)
	Premium			Market Share		
Risk · High · Post			56.64			-0.61**
			(36.57)			(0.30)
Risk · Post	0.58	65.89*	0.41	-0.51***	-0.93**	0.53***
	(0.95)	(35.52)	(9.79)	(0.089)	(0.36)	(0.084)
High · Post			-43.99			0.62**
			(36.55)			(0.31)
Risk · High			-109.54*			-0.10
			(60.01)			(0.53)
Counties		15% tails			15% tails	
Contracts	low	high +	all	low	high +	all
Service area risk		<50%			<50%	
y mean	42.48	94.31	50.68	0.27	0.53	0.31
R^2	0.72	0.81	0.75	0.36	0.63	0.40
N	5,151	1,011	6,213	5,151	1,011	6,213

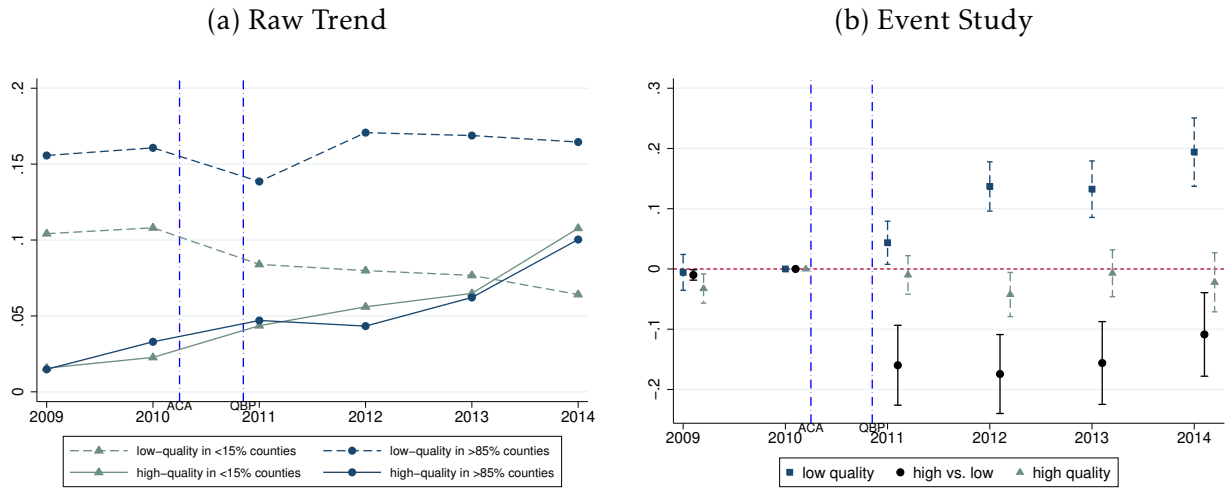
Notes: Table estimates the variation in premiums (column 1-3) and market shares (column 4-6) over county risk scores across the 15% risk tails (below 0.87205 or over 1.0295). We restrict high-quality contracts to the high-selection contracts below the median service area risk (0.975) of high quality. Standard errors clustered two-way at the level of contracts and counties in the parenthesis.

can differ from the baseline values. Second, contracts initially missing star ratings – such as new contracts – are not included in the analysis even if most obtain quality ratings later in the sample period. This section shows that the aggregate effects are robust to using contemporaneous quality ratings in the estimation sample and in the full sample of Medicare Advantage contracts. We first show results in the estimation sample.

D.1.1 Estimation Sample

Figure E3 shows the market share changes when high and low quality are defined using contemporaneous ratings. In the lower 15% risk tail (gray lines), high-quality market shares narrowed with and overtook low-quality market shares during the sample period. At the same time, low-quality market shares increased substantially in the upper risk tail (dotted lines). By comparison, high-quality market shares increased more in the lower risk tail, although the difference is not statistically significant. Table E3 shows the estimated effects.

Figure E3: Effects on market shares, contemporaneous quality, 15% risk tails



Notes. Figure plots the raw trends (panel a) and the event study estimates (panel b) of market shares by high and low quality based on contemporaneous ratings. Contracts receiving a 4.0 star rating or above are classified as high quality. We construct the quality-level market shares in a balanced panel of county-quality-years where counties with masked enrollment data in some but not all years receive zero market shares in years missing enrollments. The raw trends in panel (a) plot the levels of market shares across the 15% tails of county risk scores (below 0.8523 or above 1.0225). Event study estimates in panel (b) show variation in markets shares over continuous risk scores across the 15% tails. 95% confidence intervals are plotted based on robust standard errors clustered at the level of counties.

Table E3: Effects on market shares, contemporaneous quality

	(I)	(II)	(III)	(IV)	(V)	(VI)
Risk · High · Post			-0.21***			-0.14***
			(0.032)			(0.035)
Risk · Post	0.15***	-0.037**	0.16***	0.12***	-0.0031	0.13***
	(0.016)	(0.016)	(0.017)	(0.018)	(0.016)	(0.019)
High · Post			0.25***			0.19***
			(0.030)			(0.033)
Risk · High			-0.16***			-0.17***
			(0.022)			(0.025)
Counties		all			15% tails	
Contracts	<4.0	≥4.0	all	<4.0	≥4.0	all
y mean	0.12	0.050	0.086	0.13	0.049	0.088
R^2	0.62	0.69	0.40	0.62	0.68	0.43
N	17,236	17,236	34,508	5,070	5,070	10,164

Notes: Table estimates the variation in market shares by high and low quality over county risk scores. Quality-level market shares are based on contemporaneous ratings, and contracts with a 4.0 star rating and above are classified as high quality. We construct market shares for a balanced panel of quality-county-years where counties with masked enrollment data in some but not all years receive zero market shares in years with missing enrollments. Column 1-3 are estimated over the full sample of counties. Column 4-6 restrict counties to those in the 15% tails of county risk scores (below 0.8523 or above 1.0225). Robust standard errors clustered at the level of counties in the parenthesis.

D.1.2 Full Sample of MA Contracts

For a comprehensive view of the quality distribution in the Medicare Advantage market, this section constructs market shares using all contracts listed in the Landscape Files. The key difference from the estimation sample is that regional Preferred Provider Organization (PPO) plans, Part-C only plans, and contracts with missing quality ratings are retained in this sample. We define high-quality status as having at least a 4.0 star rating in the contemporaneous rating, and construct high-quality market shares for county-years with at least one Medicare Advantage plan listed in the Landscape Files.

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

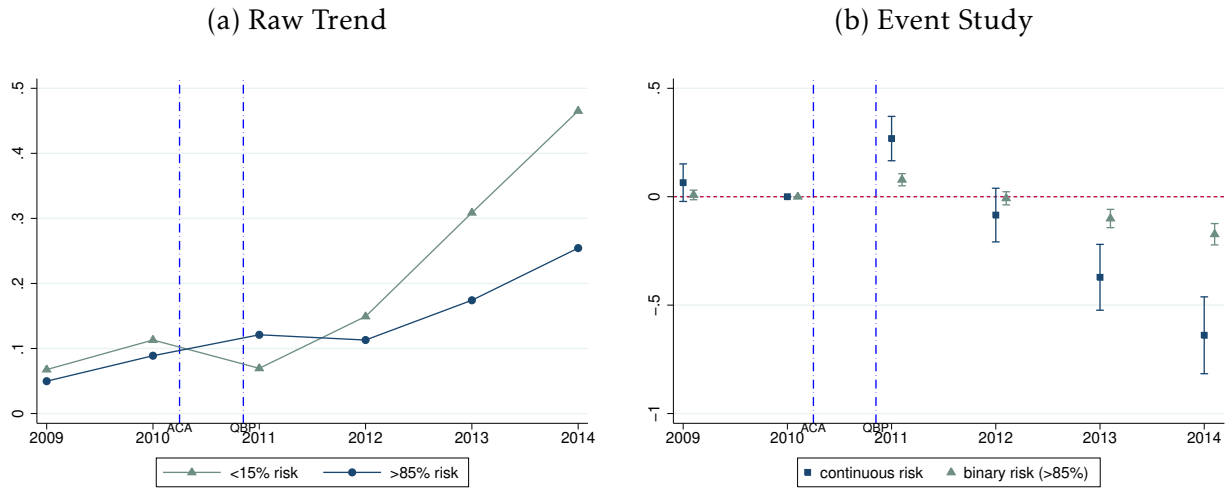
	(I)	(II)	(III)	(IV)
Risk · Post	-0.84***	-0.073	-0.16***	-0.037***
	(0.27)	(0.12)	(0.048)	(0.013)
Risk	continuous county risk scores			>85%
Counties	<15%	>85%	15% tails	15% tails
y mean	0.19	0.13	0.16	0.16
R^2	0.50	0.65	0.54	0.54
N	2,410	2,639	5,049	5,049

Notes: Table estimates the change in the high-quality market share where quality is defined using the contemporaneous ratings. Contracts receiving a 4.0 star rating and above are classified as high quality. High-quality market shares are constructed from contract enrollments for county-years with at least one Medicare Advantage plan listed in the Landscape Files. We restrict counties to the bottom 15% of county risk scores (below 0.8597) in column 1, and to the top 15% county risk scores (above 1.0233) in column 2. Column 3-4 estimate market share changes over pooled county risk across the 15% tails. In column 4, the variable Risk is a binary indicator of the top 15% tail, and the estimate gives the discrete change in the high-quality market share when moving from the bottom to the top 15% risk tail. Robust standard errors clustered at the level of counties in the parenthesis.

Table E4 shows that high-quality market share decreased significantly with county risk across the 15% risk tails: a ten percentage point increase in the county risk score lowered high-quality market share by 1.6 percentage points (column 3). Moving from the bottom to the top 15% risk tail, high-quality market share decreased by 3.7 percentage points

(column 4). By contrast, the corresponding estimate in the estimation sample is small and insignificant (column 5 of Table E3), suggesting that much of the effect in the full sample is driven by new contracts receiving higher quality ratings in lower-risk counties. In the bottom risk tail, high-quality market shares decreased by 8.4 percentage points per ten percentage point increase in county risk score (column 1), driving the average effect across 15% tails.

Figure E4: Effects on high-quality market shares, contemporaneous quality, 15% tails



Notes. Figure plots the raw trends (panel a) and the event study estimates (panel b) of high-quality market shares based on contemporaneous ratings. Contracts with at least a 4.0 star rating are classified as high quality. High-quality market shares are aggregated from contract enrollments for county-years with at least one Medicare Advantage plan listed in the Landscape Files. The raw trends plot high-quality market shares in the bottom (below 0.8597) and the top (above 1.0233) 15% tails of county risk scores, respectively. The event study plots the yearly effects over pooled risk scores across the 15% tails, showing variations over continuous risk scores on the left, and discrete changes in high-quality market shares when moving 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.

Figure E4 shows that the growth of high-quality market share accelerated in low-risk counties in 2012-2014 (panel a). By 2014, moving from the bottom to the top risk counties would decrease high-quality share by as much as 17.3 percentage points (panel b). The dip in 2011 reflects the rating computation change when the overall star rating combining the Part C and Part D ratings was first introduced. The new rating requires a larger number of measure-level ratings, some of which cannot be computed due to insufficient data supplied by small-enrollment contracts. The disruption is not visible in the estimation sample (Figure E3) and does not affect the main analysis of continuing contracts differentiated by baseline quality ratings.

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.