SUPPLEMENT TO "QUALITY DISCLOSURE AND REGULATION: SCORING DESIGN IN MEDICARE ADVANTAGE"

BENJAMIN VATTER
Sloan School of Management, Massachusetts Institute of Technology

ADDITIONAL TABLES

TABLE I

DATA DESCRIPTIVE STATISTICS

	TM		MA			MA	
Panel A: MCBS - Enrollees					**Panel B: CMS - Plans**		
Female	0.527	(0.499)	0.555	(0.497)	Bid	9439.215	(1058.175)
Age	72.257	(10.234)	73.025	(8.994)	Benchmark	10146.541	(964.674)
Part B premium	103.003	(41.743)	111.724	(20.140)	Benefits	885.201	(478.974)
Income	49.454	(63.854)	44.417	(57.179)	MA Premium	351.918	(493.690)
ESRD	0.007	(0.082)	0.003	(0.057)	Rebate	765.506	(602.134)
Disabled	0.113	(0.316)	0.098	(0.298)	Part D premium	285.546	(291.295)
Health - Excellent	0.183	(0.387)	0.176	(0.381)	Market share	0.024	(0.035)
Health - Very good	0.306	(0.461)	0.327	(0.469)	**Panel C: Quality**		
Health - Good	0.284	(0.451)	0.290	(0.454)	Access	0.860	(0.085))
Health - Fair	0.144	(0.351)	0.148	(0.355)	Intermediate	0.712	(0.088)
Health- Poor	0.060	(0.238)	0.047	(0.212)	Outcomes	0.674	(0.274)
Risk score	0.853	(0.901)	0.461	(0.253)	Patient	0.867	(0.029)
Predicted spending	6963.60	(8810.28)	6792.20	(10177.86)	Process	0.686	(0.056)
Observations	58211		12416			139323	

Note: Panel A shows the means and standard deviations (in parentheses) of key variables in the MCBS sample of enrollees used for estimation. Health variables present self-assessed enrollee health status. Risk scores are computed based on linked claims and public CMS software. These numbers are scaled to exactly match the observed average risk score of each plan in estimation. Predicted spending includes inpatient, outpatient, physician visits, SNF, home health, and hospice. Observations are weighted by nationally representative multipliers. Panel B shows the means and standard deviations of key elements of the panel of plans used in the analysis. Panel C shows the mean group normalized quality within each category and its standard deviation (in parentheses). All values are Health-CPI adjusted to 2015 values.

TABLE II

ADDITIONAL EVIDENCE OF ENROLLMENT RESPONSES TO SCORING DESIGN

		(1)		(2)		(3)	
Star Rating	3	0.027	(0.004)				
	3.5	0.044	(0.004)				
	4	0.068	(0.005)				
	4.5	0.094	(0.006)				
	5	0.179	(0.009)				
High Rated (≥ 4)	Baseline			0.008	(0.001)	0.008	(0.001)
	$\times$ Post 2012			0.007	(0.001)		
	$\times$ Access contr.					-0.130	(0.017)
	$\times$ Patient contr.					0.205	(0.025)
	$\times$ Process contr.					-0.006	(0.023)
	$\times$ Intermediate contr.					-0.055	(0.011)
N		65263		416,399		416,399	
R^2		0.053		0.746		0.746	
Sample		MA Enrollees		All		All	

Note: This table presents the results of the policy knowledge test. The dependent variable is the choice indicator. The population is limited to new MA enrollees and to MA plans as in the reduced-form analysis conducted in the main text. The first two regressions examine whether consumers with any chronic condition respond more to star ratings when the weight of the Intermediate Outcome category increases. This category primarily contains measures related to chronic condition management. The last two columns explore whether diabetic consumers respond more to ratings when the weight given to diabetic quality measures changes. Diabetic quality measures are almost all contained in the Intermediate Outcome category. Errors are homoskedastic.

TABLE III

PLAN QUALITY RESPONSE TO DESIGN VARIATION AND ROBUSTNESS

	(I) - Main		**(II) - Uncensored**		**(III) - Quartiles**		**(IV) - Controls**	
Preexisting quality group (G_{lj})								
1 star			1.226	(0.0669)				
2 stars	0.592	(0.0601)	0.873	(0.0603)			0.630	(0.0630)
3 stars	0.163	(0.0425)	0.393	(0.0524)			0.185	(0.0458)
4 stars			0.250	(0.0469)				
2nd quartile					0.156	(0.0387)		
N	195575		198189		194914		36464	
R^2	0.588		0.594		0.589		0.701	

Note: This table presents the estimates and robustness of the triple-difference estimates of quality responses to design changes. Column (I) provides the main estimates, as detailed in the main text. The remaining columns, (II), (III), and (IV), serve as robustness checks, each addressing a specific aspect: (II) robustness to the domain censoring, (III) to CMSs' cutoffs, and (IV) to the selection of controls. The standard errors, indicated in parentheses, are clustered at the contract level.

TABLE IV

ESTIMATED DEMAND PREFERENCE COEFFICIENTS, DEMOGRAPHIC PREFERENCES FOR MA

	coeff	std. error		coeff	std. error
2009 cohort	-2.046	(0.119)	Disabled	-0.276	(0.141)
2010 cohort	-2.434	(0.187)	ESRD	-0.433	(0.356)
2011 cohort	-2.163	(0.171)	Employer sponsored	-1.251	(0.040)
2012 cohort	-1.530	(0.126)	Female	-0.026	(0.066)
2013 cohort	-1.271	(0.128)	Graduated high school	-0.035	(0.052)
2015 cohort	-0.615	(0.135)	High income	-0.385	(0.084)
Asian indicator	-0.071	(0.119)	High risk score	-2.360	(0.109)
Attended college	-0.006	(0.044)	Hispanic indicator	0.179	(0.068)
Black indicator	0.075	(0.062)	Medium income	-0.080	(0.079)
College degree or higher	-0.162	(0.044)	Medium risk score	-0.332	(0.077)
N	36447		loglikelihood	-5.403	

Note: This table shows the estimated coefficient on variables determining consumers' overall preference for enrollment in MA, which are elements of λ^l in consumers' indirect utility equation in the main text. These coefficients are estimated during the first stage of the estimator using constrained MLE. Each variable presented above interacts with an indicator that equals zero for TM and one for any MA plan. Cohort variables are indicators for each generation of new enrollees during their first year of enrollment. Risk scores and income are segmented according to terciles, with the baseline of the low-income low-risk score being omitted due to collinearity with the plan-market-year fixed effects included in the estimation. The regression also includes fixed effects that interact age groups and self-reported health status with firm identity for each of the top six firms in the market: United Healthcare, Kaiser, Humana, BCBS, Cigna, and Aetna. Standard errors are homoskedastic and corrected for multi-stage estimation.

TABLE V

ESTIMATED DEMAND PREFERENCE COEFFICIENTS, EFFECTS OF INSTRUMENTS

	OLS		Premium & Bid IV		Quality IV		All IVs	
Premium	-0.965	(0.296)	-1.361	(0.377)	-0.965	(0.296)	-1.361	(0.377)
Benefits	0.887	(0.343)	3.090	(0.498)	0.887	(0.343)	3.090	(0.498)
Expected quality								
Access	5.627	(0.210)	4.961	(0.126)	6.056	(0.137)	5.338	(0.160)
Intermediate	2.232	(0.057)	2.014	(0.033)	2.429	(0.062)	2.198	(0.096)
Outcome	6.760	(0.167)	5.430	(0.198)	6.615	(0.112)	5.493	(0.603)
Patient	4.583	(0.207)	4.175	(0.136)	4.342	(0.189)	4.052	(0.194)
Process	2.833	(0.077)	2.497	(0.071)	2.726	(0.066)	2.470	(0.265)
Additional benefits								
Dental cleaning	1.619	(0.060)	1.882	(0.077)	1.619	(0.060)	1.882	(0.077)
Dental exam	-1.697	(0.059)	-2.573	(0.116)	-1.697	(0.059)	-2.573	(0.116)
Dental x-ray	0.369	(0.018)	0.777	(0.053)	0.369	(0.018)	0.777	(0.053)
Drug deductible	-0.002	(0.000)	-0.001	(0.000)	-0.002	(0.000)	-0.001	(0.000)
Enhanced drug coverage	0.101	(0.014)	0.072	(0.018)	0.101	(0.014)	0.072	(0.018)
Fluoride treatment	-0.272	(0.014)	-0.536	(0.031)	-0.272	(0.014)	-0.536	(0.031)
Hearing aids	-0.006	(0.016)	-0.332	(0.041)	-0.006	(0.016)	-0.332	(0.041)
Hearing aids fitting	0.019	(0.022)	-0.164	(0.037)	0.019	(0.022)	-0.164	(0.037)
No part d coverage	-1.615	(0.017)	-1.816	(0.029)	-1.615	(0.017)	-1.816	(0.029)
Vision coverage	0.195	(0.014)	-0.028	(0.031)	0.195	(0.014)	-0.028	(0.031)
MA-demographic controls	Yes		Yes		Yes		Yes	
Market-MA FE	Yes		Yes		Yes		Yes	
Top firm-demographic controls	Yes		Yes		Yes		Yes	
N	36447		36447		36447		36447	
loglikelihood	-5.403		-5.403		-5.403		-5.403	
Price elasticity	-5.757		-9.380		-5.757		-9.380	
Premium elasticity	-0.652		-0.968		-0.652		-0.968	

Note: This table shows the coefficient estimates for the second stage of the demand model. Coefficients of the first stage, which capture preference heterogeneity for premiums, benefits, and MA-demographic controls, are independent of the chosen instruments and omitted. The first column shows the results without any instrument, the second with only premium and benefit instruments, the third with only quality instruments, and the last with all instruments. Standard errors are homoskedastic and corrected for multi-stage estimation. See Section II of this appendix for a discussion of the price elasticity.

TABLE VI

ESTIMATED INSURANCE MARGINAL COST COEFFICIENTS

	coeff	std. error		coeff	std. error
Quality (θ_q^c)			Enhanced drug benefits	-20.728	(3.698)
Access	30.077	(16.937)	Vision benefits	7.211	(3.123)
Intermediate	104.038	(12.883)	Dental cleaning benefits	48.314	(30.031)
Outcome	15.903	(3.886)	Dental exam benefits	-17.346	(29.786)
Patient	-215.450	(57.988)	Dental fluoride benefit	2.380	(3.123)
Process	-176.810	(28.043)	Dental X-ray benefit	1.326	(5.572)
Other components (θ_a^c)			Hearing aids	43.000	(3.513)
No part D	-107.381	(4.041)	Hearing aids fitting	7.174	(4.455)
Drug deductible	0.008	(0.014)			
New plan	-211.270	(51.362)			
N	28,966		R^2	0.531	

Note: This table shows the estimated coefficients on insurers' marginal insurer cost function. The estimated regression also includes fixed effect controls for plan types (HMO, PPO, Regional plans, and PFFS), county, year, and contract identifiers. Values are in dollars per unit-risk member per month. Standard errors are heteroskedasticity robust and corrected for two-step estimation following Murphy and Topel (1985).

TABLE VII

INVESTMENT COST SPILLOVER TERMS ($\mu''_{k,k'}$)

Term	Intermediate	Outcome	Patient	Process
Panel A: Insurance cost				
Access	0.219	-0.159	-0.389	-0.208
	(0.097)	(0.090)	(0.239)	(0.171)
Intermediate		-0.563	-1.227	-0.320
		(0.124)	(0.333)	(0.293)
Outcome			-0.794	-0.950
			(0.310)	(0.206)
Patient				-2.318
				(0.358)

Note: This table presents the estimated investment cost parameters associated with across-category spillovers in millions of dollars per hundred thousand Medicare beneficiaries per year. The regression includes common linear and quadratic terms shown in the main text and linear cost terms for the interaction between the top six firms' identities and each quality dimension, which are omitted for space. The estimation strategy imposes symmetry across categories. $N = 7,684$.

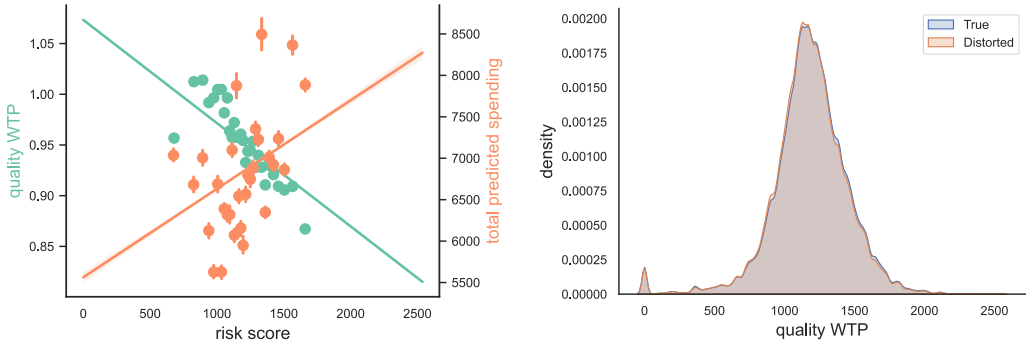
TABLE VIII

MARGINAL WELFARE VALUE OF QUALITY, COMPETITION, AND SCORING DESIGN

	(I)	(II)
HHI	0.069	0.044
	(0.011)	(0.013)
Category scoring contribution	-65.597	-39.738
	(12.750)	(16.551)
category-contract FE	Y	
category-state-firm FE		Y
N	7305	7305
R^2	0.776	0.595

Note: This table presents the results of regressing the marginal welfare value of quality for each contract, quality dimension, and year against the local HHI and the relative contribution of each dimension (category) to the score. Each observation is a contract-category-year. The mean HHI is 206.3, computed at the county-year-firm level, and includes the market share of TM.

ADDITIONAL FIGURES

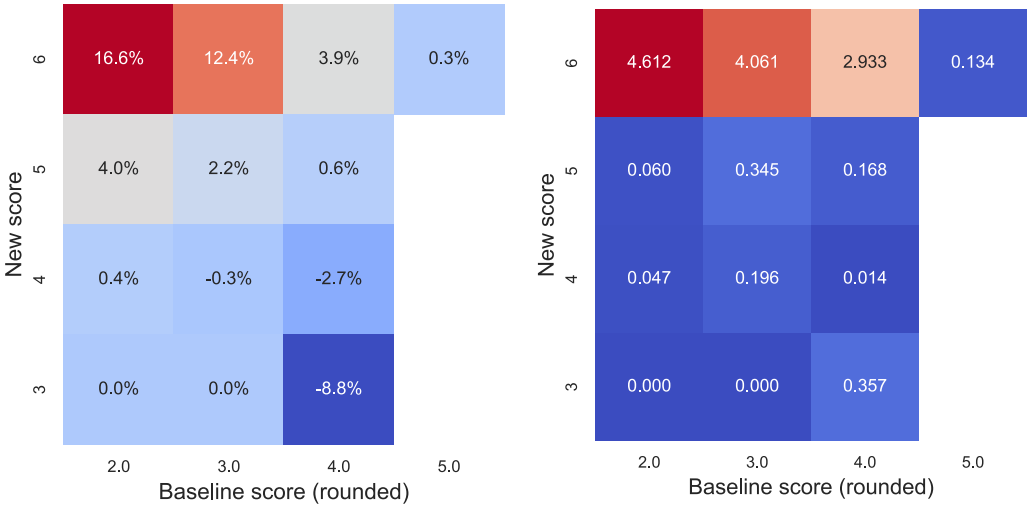


(a) Correlation between WTP and risk scores

(b) Distribution of WTP

FIGURE 1.—Individual risk scores and consumers' willingness to pay for quality

Note: Figure (a) displays a binned scatter plot of consumers' willingness-to-pay for quality (WTP) relative to their risk score and total predicted spending. WTP is computed by taking the average WTP for plan quality for each consumer among her MA plans. Figure (b) shows the distribution of WTP among the Medicare population (blue) and how risk scores distort this distribution from the perspective of insurers (red). The distortion is computed by resampling consumers based on their risk score, as perceived by insurers.

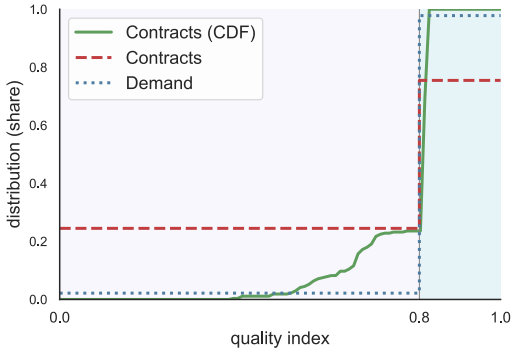


(a) Quality utility

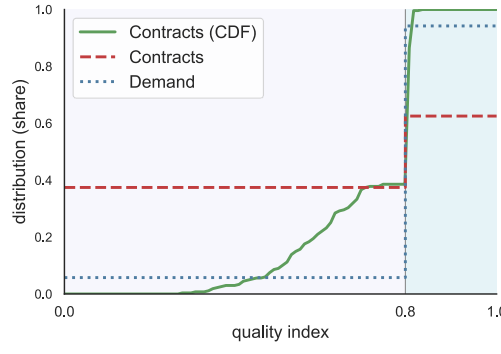
(b) Investment cost

FIGURE 2.—Quality and investment cost transition from baseline to optimal design equilibrium

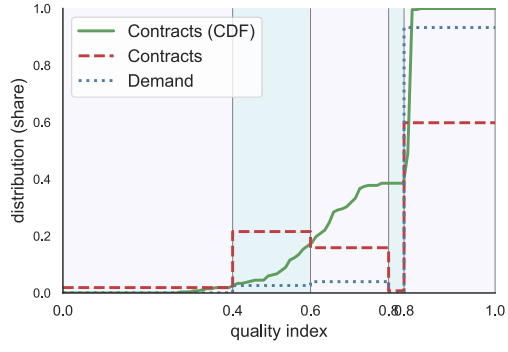
Note: These figures display the change in quality utility (a) and investment cost (b) for contracts that move from each score in the baseline to each score in the counterfactual equilibrium. The baseline scores have been rounded to the nearest integer (e.g., baseline 2.0 captures both 2.0 and 2.5), and the displayed numbers are group averages. Empty cells mark transitions that do not occur in expectation (over investment risk).



(a) Optimal certification



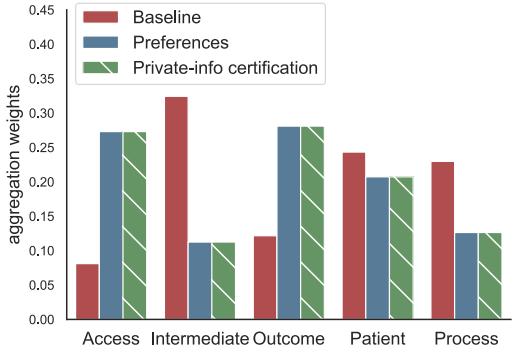
(b) CMS certification



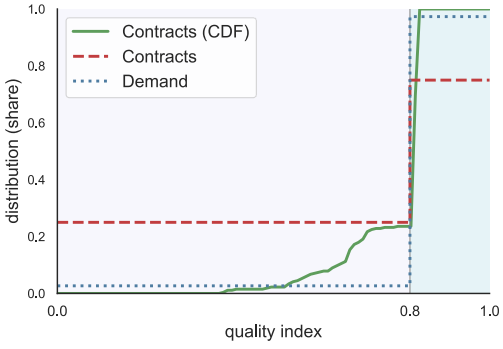
(c) CMS optimal

FIGURE 3.—Optimal certification design and CMS-weighted designs

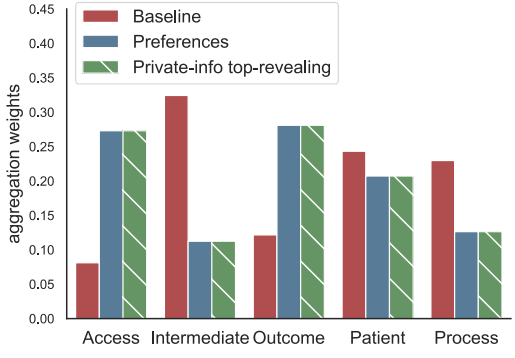
Note: These figures display the cutoff placements for the optimal certification design (a), the optimal CMS-weighted certification design (b), and the optimal CMS-weighted design with multiple scores (c). The aggregation weights for the optimal certification match consumers' preferences. Those for the CMS-weighted match the baseline aggregator shown in the main text.



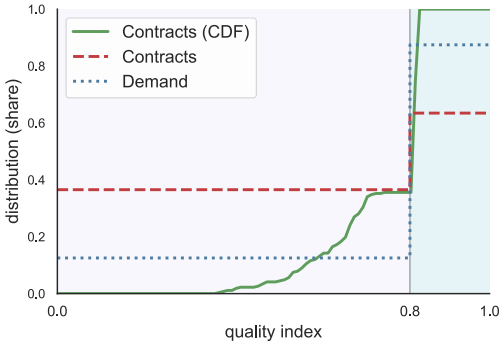
(a) Coarse design - aggregator



(b) Coarse design - cutoffs



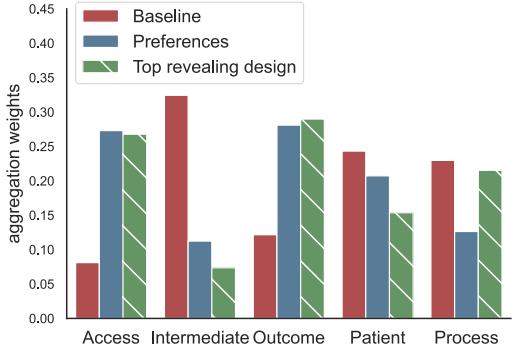
(c) Top revealing - aggregator



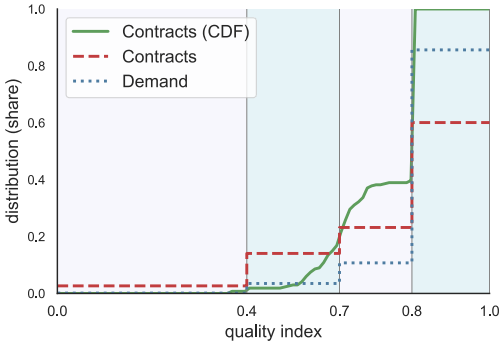
(d) Top revealing - cutoffs

FIGURE 4.—Optimal certification under cost type uncertainty

Note: These figures display the optimal design under regulatory uncertainty of firms' investment cost type. Figures (a) and (b) show the coarse certification cutoff design that only discloses whether scores exceed the threshold. Figures (c) and (d) report the results for a design that fully reveals the quality of plans that exceed the threshold.



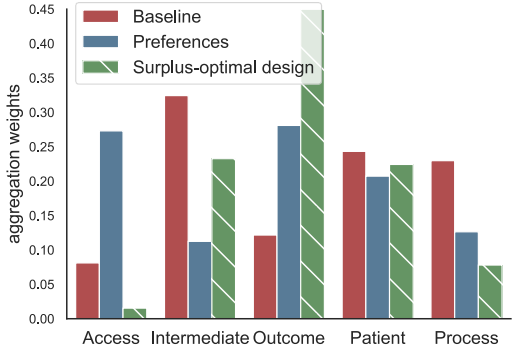
(a) Aggregator



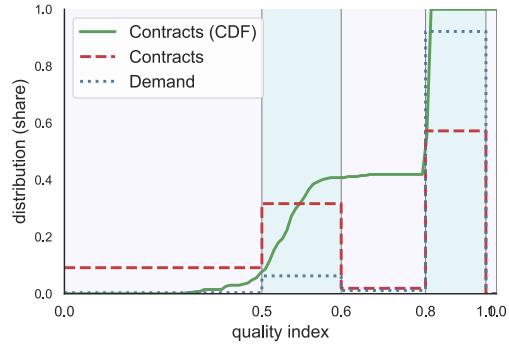
(b) Cutoffs

FIGURE 5.—Optimal top-revealing design

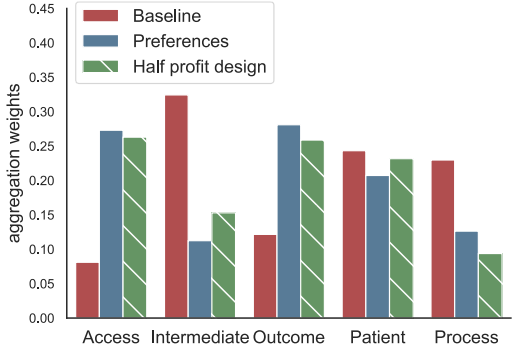
Note: These figures display the optimal top-revealing scoring design. Figure (a) shows this design's optimal aggregator for quality dimensions. Figure (b) presents the cutoff placement along the aggregated quality index.



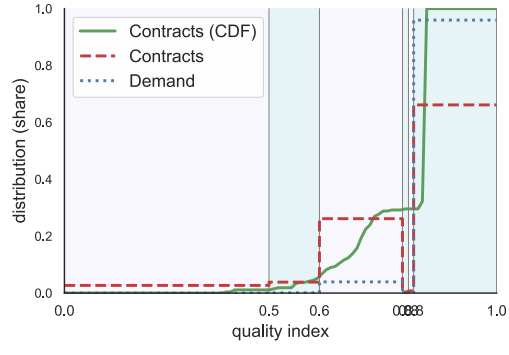
(a) $(\rho^F = \rho^G = 0)$ - Aggregator



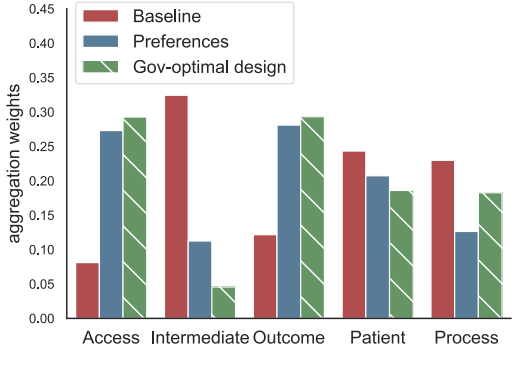
(b) $(\rho^F = \rho^G = 0)$ - Cutoffs



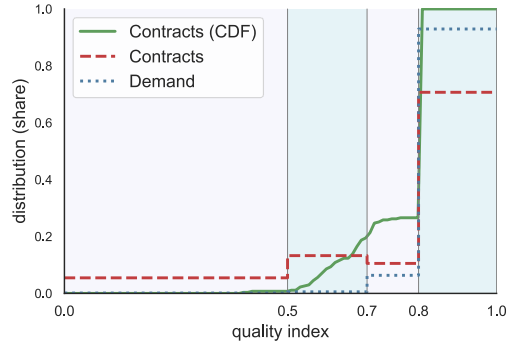
(c) $(\rho^F = 0.5, \rho^G = 0)$ - Aggregator



(d) $(\rho^F = 0.5, \rho^G = 0)$ - Cutoffs



(e) $(\rho^F = \rho^G = 1)$ - Aggregator



(f) $(\rho^F = \rho^G = 1)$ - Cutoffs

FIGURE 6.—Optimal designs under alternative regulatory objectives

Note: These figures display the optimal design under alternative regulatory objectives. Figures (a) and (b) show the result for an objective considering only consumers' surplus. Figures (c) and (d) consider an objective where the regulator values consumers' surplus twice as much as insurers' profits. Figures (e) and (f) consider an objective that maximizes the total welfare net of subsidization cost.

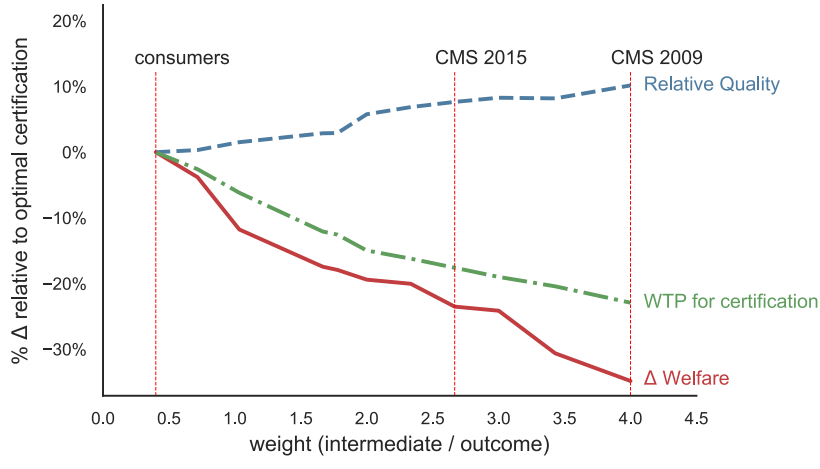


FIGURE 7.—Impact of regulatory preferences on market welfare

Note: This figure shows the consequences of skewing the relative weight given to the Intermediate Outcomes category relative to the Outcomes category in the scoring design. Starting with the optimal weighting (which matches consumers' preferences) and increasing the relative weight to match CMS's design across the years. For 11 points that span the interval, I solve for the constrained optimal certification design and report the changes in welfare relative to the optimal design (solid red), the change in consumers WTP for certified products (dashed-dotted green), and the average Intermediate Outcome quality relative to Outcome quality among products offered in the market, weighted by enrollment.

1. INSTITUTIONAL DETAILS AND DESCRIPTIVE EVIDENCE

1.1. *Pricing regulation*

Price and coverage regulation in MA operates through a process known as bidding. Every year, insurers submit insurance plan offerings, listing each plan's participating counties, cost-sharing attributes, actuary-certified estimates of expected expenses, plan-level estimates of administrative costs, and profit margins. The bidding procedure combines these data to form two components: First, the revenue required to cover expenses and margins related to standard and mandatory Medicare coverage, which CMS calls the *bid*. In the main text, I call this value the plan's *price*. Second, the revenue required to cover supplementary benefits, such as lower copays and maximum out-of-pocket amounts, which I refer to as the plan's additional *benefits*. These additional benefits are not optional and exclude dental, vision, hearing, or Part D prescription coverage.

The bidding process compares a plan's bid against a benchmark related to TM's cost. CMS computes plan benchmarks by averaging TM's fee-for-service costs in every county the plan operates in, using weights proportional to the plan's expected enrollment. CMS pays plans bidding above the benchmark an amount equal to the benchmark per enrollee, and enrollees pay the difference in what is known as the *basic* MA premium. For plans bidding below the benchmark, CMS pays their bid plus a rebate equal to a fraction of the difference.

Additional benefits allow MA insurers to offer more generous coverage than TM. However, to provide these benefits, insurers must fund them through either premiums or rebates. Specifically, insurers must use every dollar of rebates to either fund benefits or buy down non-MA enrollee premiums. The latter includes the part B premium CMS charges to every enrollee regardless of their choice between TM and MA and any part D prescription drug premium the plan might charge. Any additional benefits not funded by rebates are paid directly by the consumer under a *supplementary* MA premium.

Overall, the following equations summarize this regulation.

$$\begin{aligned}
\text{Rebate}_j &= \rho_j \max\{B_j - p_j, 0\} \\
\text{Premium}_j &= \underbrace{\max\{p_j - B_j, 0\}}_{\text{basic}} + \underbrace{\max\{b_j - \mathit{Rebate}_j, 0\}}_{\text{supplementary}} \\
\text{Payment}_j &= \min\{p_j, B_j\} + \mathit{Rebate}_j + \mathit{Premium}_j
\end{aligned}$$

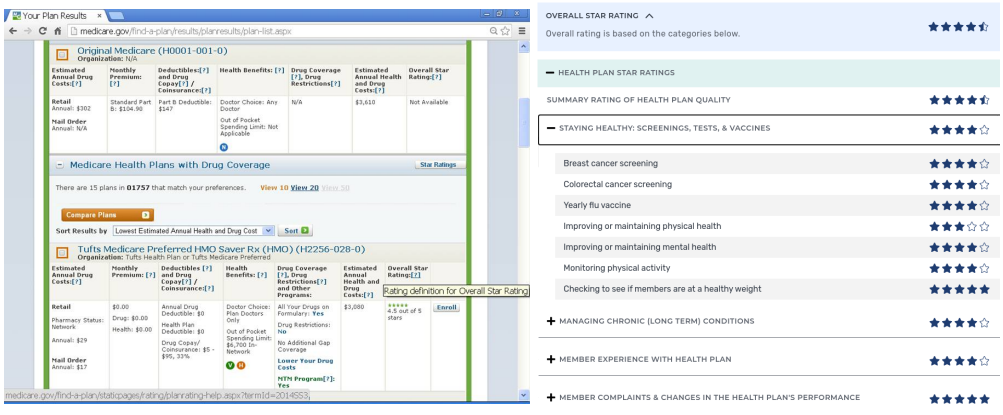
Where b_j is the plan's additional cost-sharing benefits, measured in cost-savings for the average unit-risk consumer, and ρ_j is the rebate share.[1] Put in perspective, per member per month, the average MA plan in 2015 (by enrollment) submitted a price of \$700, additional benefits equal to \$70, and faced a benchmark of \$782. Among plans with a non-zero premium (43.8%), the average was \$73.5, with 13.4% coming from the basic MA premium. More than half of enrollees chose a zero-premium plan (58%). However, 83.9% of plans had a zero basic premium, resulting in a rebate averaging \$63.8. Every MA plan offered additional consumer benefits, averaging an actuarial value for medical services of 87.2%.[2]

1.2. *Enrollment platform*

MA organizes plan offerings in a unified shopping platform and regulates contract characteristics. Figure 8a displays the view of the platform in 2015. The platform presented consumers

[1] The rebate share has varied over the years and, since 2012, depends on plans' rating in previous years.

[2] Actuarial values of MA are computed using public CMS software.



(a) Plan Compare website (b) Plan Compare rating details

FIGURE 8.—Medicare plan finder view

Note: Figure (a) shows a view of the Medicare Plan Finder platform from 2015. Figure (b) shows the detail presented to consumers after clicking on the details button in 2022. A similar view was offered in 2015.

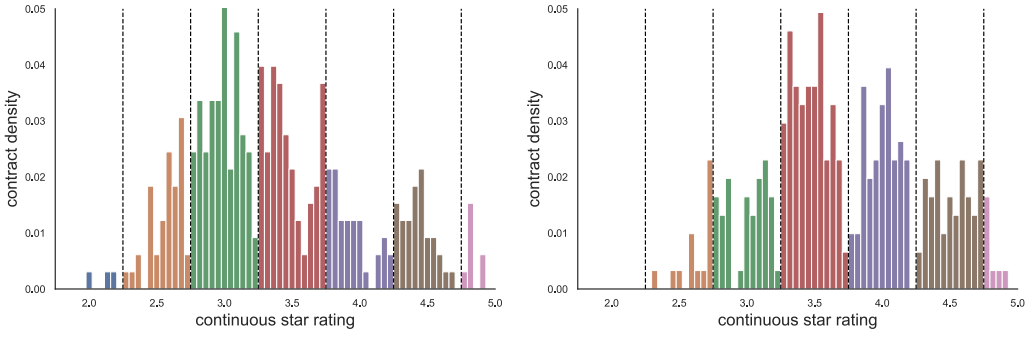
with the Original Medicare option at the top, and MA plans at the bottom. Each plan displayed its estimated deductibles, cost-sharing rules, and monthly premiums. The system showed the "Estimated annual health and drug costs" closely related to the benefit levels used in the main text. Also, the system included the MA Star Rating for the plan next to the enrollment button. Clicking on the question mark in the column name revealed the basic construction details of this rating, as in Figure 8b.

1.3. *Comparison to previous descriptions of the market*

There is some discrepancy in the literature regarding the bidding process. A recent release of information by CMS, containing the complete bidding data, software, and instructions sent to insurers for 2009-2015, informed the description above and solved some of these discrepancies. For example, bidding is often misdescribed as taking place at the plan-county level. However, insurers can segment their contracts across counties and submit separate bids for each segment. The fact that 99% of all plans offered in more than one county chose not to segment suggests the gains from bidding independently across counties are small, and so are the losses from treating competition as occurring at the county level.

Another distinction is that premiums are often described as the positive difference between bids and benchmarks. However, this is only the definition of the *basic* MA premium, which constitutes but 12% of the total MA premium paid by consumers. The standard model also describes rebates as the positive difference between benchmarks and bids. Thus, each plan should only have either a premium or a rebate. The data refute this model as 45% of plans have both premium and rebate.[3]

[3]Using the wrong model can have implications for previous work. In particular, previous work inferred bids from premiums or rebates. Depending on the scenario, these inversions can lead to erroneous bids and linkages between prices and subsidies.



(a) Year 2012 (b) Year 2015

FIGURE 9.—Distribution of underlying continuous aggregate star rating

Note: The figures show the histogram of aggregate contract quality before rounding. Every contract within a same-color bin obtains the same rating, denoted on the horizontal axis.

1.4. *The Star Rating Program*

Improving the quality of care and beneficiary general health is one of CMS' most important strategic goals (Centers for Medicare and Medicaid Services, 2016). To this end, CMS has undertaken several initiatives to gather and display the quality of MA plans. Following the Balanced Budget Act of 1997, CMS began collecting information on multiple quality measures through surveys and insurer reports. A summary of the gathered data was first presented to consumers in the November 1999 edition of *Medicare & You*, a handbook mailed annually to Medicare-eligible enrollees. The impact of this first implementation was noticeable, as studied by Dafny and Dranove (2008). In 2007, CMS began summarizing the quality information into five quality domains (e.g., "Helping You Stay Healthy"), with values described by one to five stars. In 2009, the star rating program took on its current form, with a single overall rating displayed to consumers next to the plan's name, premium, and cost-sharing attribute. The Supplementary Material presents the formulas used to construct plan scores, adjustment factors, and the adjustment to rebate shares that depend on previous scores.

1.5. *Data*

Details about the data construction are provided in the Supplementary Material. Table I displays descriptive statistics of the combined plan-county-level public data and the administrative MCBS individual-level data. Importantly, the MCBS reports individuals' chronic conditions and well-being. The list of reported conditions that are treated as indicators of chronic illness in this article includes having an enlarged prostate, hearing loss, AlzheimerâĂŹs, amputated limbs, arterial hardening, arthritis, diabetes, broken hips, cancer, heart failure, angina pectoris, CHD, high cholesterol, high blood pressure, skin cancer, chronic valve problems, dementia, depression, COPD/asthma/emphysema, hypertension, retardation, past myocardial infarctions, osteoporosis, ParkinsonâĂŹs, paralysis, mental disorders, arrhythmia, and past strokes.

1.6. *Understanding the design*

The Star Ratings' design consists of two fundamental components: the overall contribution of each category to the score and the cutoffs used to discretize the measures underlying each

category. Category contributions are easily observed as measure weights are uniform within a category: 3 for Outcome and Intermediate Outcomes, 1.5 for Patient and Access, and 1 for Process.[4] Therefore, it suffices to know the number of measures in each category, which is visible on the enrollment platform. CMS informs consumers of the scores and what they capture in a yearly booklet sent to every beneficiary called *Medicare & You* and on their website. Additionally, insurers often cover large changes to the design in news articles and discuss them in promotional material.

Measure-level cutoffs are less visible, but two features facilitate their understanding. First, their variation over time is minimal, which allows consumers to learn them. As Medicare beneficiaries are either retirees or disabled, they likely interacted significantly with their local healthcare markets. They would, therefore, understand the distribution of quality inputs for insurers. By looking at the relative distribution of ratings in their markets, consumers can use their experience to form a partial understanding of cutoffs. For example, suppose only a small fraction of plans have five stars, but consumers know that there are many high-quality providers, and the contribution of the Outcome category is large. In that case, it stands to reason that the average five-star cutoff within the Outcome category is very high.

The second facilitating feature is that, as discussed in Appendix III, cutoffs need not be understood individually. As cutoffs within a category are averaged and then discretized in the overall score, they are well approximated by cutoffs to aggregate quality. These cutoffs are fewer and benefit from the same stability as their measure-level sources.

1.6.1. *Testing consumers' understanding*

As noted in the main text, the ideas expressed above are partially testable. In addition to the two nuanced tests presented in the main text, Section IV, we can examine two simpler pieces of evidence. First, we can test whether consumers have any understanding of the variation in scoring policy by evaluating whether their choice behavior is affected by changes to the policy. Under the null hypothesis of ignorance, the derivative of the individual choice probability s_{ijt} with respect to any policy parameter ω_{kt} should be zero. To implement this, I leverage the largest change in the scoring policy caused by the change of measure-level weights in 2012. I ask whether consumers' responsiveness to star ratings changes as this new policy is introduced. To keep things simple, I test only whether their likelihood of buying a high-rated plan (with four or more stars) changes. As in the main analysis, I consider only potential new MA enrollees. To test the null hypothesis, I estimate the linear probability model (LPM) parameters associated with the regression:

$$y_{ijt} = \alpha \mathbb{1}\{t \geq 2012\}\mathbb{1}\{r(jt) \geq 4\} + \beta \mathbb{1}\{r(jt) \geq 4\} + \boldsymbol{x}'_{jt}\gamma + \delta_i + \epsilon_{ijt} \tag{1}$$

The coefficient of interest, α, tests whether the likelihood of choosing high-rated plans changed around the policy cutoff. The base effect, β, captures the baseline propensity to buy high-scoring plans in MA, and $\boldsymbol{x}_{jt}$ includes all other observed factors of the plan, including premiums and benefits. To match the structural demand model and reduce the number of demographic parameters to estimate, I include an individual fixed-effect δ_i.[5]

The second column of Table II shows the estimated parameters. As seen there, we can reject the null of no change with a p-value of less than 0.1%. Therefore, we can confidently reject the

[4] These are the weights starting in 2012. The only exceptions are newly introduced measures, which always get a weight of one.

[5] Replacing the model with a logit choice model and dropping the individual parameters results in economically similar effects as documented here.

hypothesis that consumers are entirely ignorant of the policy variation. Taking the idea further, we can test whether consumers' choice likelihood is correlated with changes to each category's contribution in the design. As discussed above, consumers on the platform are likely to observe these changes. I modify Equation (1) to include interactions of high ratings with each category's relative contribution.

$$y_{ijt} = \sum_{k \in \mathcal{K}} \alpha_k \omega_{kt} \mathbb{1}\left\{r(jt) \geq 4\right\} + \beta \mathbb{1}\left\{r(jt) \geq 4\right\} + \boldsymbol{x}'_{jt}\gamma + \delta_i + \epsilon_{ijt} \tag{2}$$

Above, $\mathcal{K}$ is the set of all categories, and ω_{kt} is the relative category contribution in year t. As contributions add up to one each year, I drop one of the categories (Outcome) and present the estimates for the rest in the third column of Table II. As shown, we can reject the null that specific elements of the scoring design do not influence consumers' choices.

Finally, we can refine the tests to examine whether consumers' responses to the design are consistent with rational behavior. These are the tests conducted in the main text. As shown, consumers with any chronic condition are more likely to buy a high-scoring plan in years when the design changes to reflect good chronic condition management. These results support the assumption that consumers are informed of key policy variations. Nevertheless, consumers' understanding of the policy variation is likely imperfect. These tests suggest that in the spectrum between ignorance and full information, consumers are closer to the latter. This motivates the reliance on the assumption of sophisticated consumers (informed choice) rather than naive ones (ignorance) for the main analysis. The alternative is, instead, used to present robustness in the final section of this Online Appendix.

1.7. *Quality and outcomes*

A natural concern is whether scores reflect "true" plan quality changes. To answer this question, we need to step back from the scores. The correlation between stars and any given quality outcome might be low, not because the actual quality measurement is off but because the design of the scores is poor and uninformative. Instead, we should focus on whether we are measuring "true" quality in the first place. For example, in 2015, the measurements included the degree to which beneficiaries are immunized, keep their blood pressure under control, obtain regular cancer screens, and maintain or improve bladder control or cognitive abilities. These are crucial for beneficiaries' quality of life, even if they have little impact on mortality. For example, fall risk is a major concern for healthcare practitioners and researchers because geriatric falls can lead to severe deterioration of health status (Masud and Morris, 2001, Howcroft et al., 2013) and cost billions of dollars in associated medical spending (Englander et al., 1996, Stevens et al., 2006).

To further test that CMS measures true quality, I combine the quality measurement data with individual-level data on claims and health status. I ask whether improvements in a dimension of quality q by a contract are associated with changes in health outcome y for its enrollees. Specifically, I regress outcomes on quality changes, controlling for beneficiary and contract identities. It is worth noting that quality measurements for plans are based on data from one or two years before they enter the scoring quality data. Hence, the populations for which the quality was measured are likely different from those I evaluate. Additionally, the outcomes that go into quality measurement are typically not from the MCBS. I focus on pairs of outcomes and measures that bear a clear relationship (e.g., plan vaccination rates and member vaccination status) and those that capture broad improvements in quality and outcomes (e.g., member rating of their health plan and self-assessed health improvement).

TABLE IX

ASSOCIATION BETWEEN PLAN QUALITY AND BENEFICIARY OUTCOMES

Outcome	**Any Falls** ∼	**Hurt From Falls** ∼	**Had Flu Shot**∼
Quality Measure	Reducing The Risk Of Falling	Reducing The Risk Of Falling	Flu Vax Rate
coeff (se)	-0.225 (0.110)	-0.224 (0.0885)	0.274 (0.123)
N	5355	1626	5457
R^2	0.603	0.545	0.842
Outcome	**Had Pneumonia Shot** ∼	**Had Mammogram** ∼	**Had Cancer Screens**∼
Quality Measure	Pneumonia Vax Rate	Breast Cancer Screening Rate	Quick Appointments
coeff (se)	0.0926 (0.107)	0.135 (0.197)	0.512 (0.168)
N	5457	5265	5457
R^2	0.916	0.814	0.678
Outcome	**Health Improved** ∼	**Health Improved**∼	**Health Improved**∼
Quality Measure	Rating Of Health Plan	Rating Of Health Care Quality	Getting Needed Care
coeff (se)	0.602 (0.300)	0.615 (0.383)	0.362 (0.326)
N	5457	5455	5434
R^2	0.583	0.582	0.582
Outcome	**Had Cholesterol Check**∼	**Health Improved (diabetic only)**∼	**# Adls Can Not Perform**∼
Quality Measure	Hypertension Adherence	Diabetes Treatment	Reducing All-Cause Readmissions
coeff	0.515 (0.264)	4.305 (1.314)	-4.911 (1.827)
N	2370	1097	2356
R^2	0.848	0.569	0.864

Note: This table shows results from regressing beneficiary outcomes on their chosen plan quality, including individual and contract-level fixed effects. Plan quality and outcome measurements come from different data sources and different populations due to the lag between measurement and scoring. Each cell in the table is a regression. Sample sizes vary as measurements are not available every year. The "any falls" column indicates whether the consumer had any unexpected falls in the current year. Hurt from falls is measured as the change between reports of being hurt from falls in the previous and current year. Flu and pneumonia shots, as well as cholesterol checks, are measured within the same year. Mammograms indicate whether the individual had a mammogram in the past four years. Cancer screens show whether male beneficiaries had prostate checks within four years and females had mammograms or pap smears within the same time. The column "health improved" indicates whether the individual reported that their health had improved this year relative to the last. "# ADLs can not perform" stands for the number of Activities of Daily Living that the individual can not perform in the current year, including getting into or out of bed, bathing, dressing, concentrating, cooking, eating, walking, conducting different levels of housework, shop, manage money, shower, socialize, use the telephone, and using the toilet. Standard errors in parentheses are homoskedastic.

Table IX shows the results for 12 quality outcomes and measures combinations. First, we see that members who enroll in contracts that have improved on the metric of reducing their population's risk of falling have, in fact, a lower risk of unexpected falls and of being hurt by such falls. Members in plans that improved vaccination rates and cancer screens are also, on average, more likely to perform better on these dimensions than before. Members of plans who improve their plan management, ratings of health care quality, or access to needed care are more likely to report health improvements. Plans that have improved their hypertension medication and review adherence programs are also more likely to have patients with higher reported cholesterol checks. Among people with diabetes, plans that improve diabetic treatment quality also have members who report improved health status changes. Finally, all-cause readmission rates are a typical measure of provider network quality (Keenan et al., 2008). Members in plans that improve their all-cause readmission rates report fewer difficulties in conducting activities of daily living (ADLs). Therefore, the results support the assumption that measured quality presents valuable improvements for consumers.

1.8. *Robustness of quality responses to scoring design*

The analysis done in section IV.C in the main text suggests a causal effect of scoring design on quality. Column I in Table III presents the estimates matching equation (4) in the main text.

This analysis censors the quality domain, dropping plans whose preexisting quality falls within the first or last score. This avoids inflating the results due to reversion to the mean. Column II shows the effect of removing this censoring, which, as expected, increases the effects. The analysis also relies on CMS's definition of the cutoffs, which might be subject to influence by insurers due to lobbying. Column III in the table shows that the results hold qualitatively if one compares the second to the third quartiles of preexisting quality rather than relying on cutoffs.

Recent research has raised concerns about staggered difference-in-differences designs similar to the one used in this analysis (Callaway and Sant'Anna, 2020, Baker et al., 2021, Goodman-Bacon, 2021). However, the structure of the treatments used in this work differs from the canonical example used in this literature. Unlike standard staggered differences-in-differences, where the same treatment is assigned to different units over time, measure entry can be seen as different treatments assigned to different units. However, regardless of this distinction, the concern remains that by aggregating the effect of different treatments, the pooled regression used in the main analysis might deliver a biased estimate of the treatment effect. To alleviate this concern, I structure the data behind this analysis as a *stacked regression estimator* (Baker et al., 2021). In this structure, the data is propagated so that each event and unit are directly matched with all their controls. Column IV in Table III shows robustness to randomly selecting 20% of controls rather than using all available observations. Additionally, Table X shows the estimated coefficient on the main analysis separately for each measure. The coefficient aggregates the effect on plans in groups 2 and 3 relative to 4 to reduce the number of estimates. The results show that all effects are positive, and most are significantly so. Therefore, the results are unlikely to be driven by a negative or non-convex weighting of the underlying individual events.

TABLE X

PLAN QUALITY RESPONSE TO DESIGN VARIATION - ROBUSTNESS

Measure	coefficient	std. err	Measure	coefficient	std. err
Access & performance	0.858	(0.213)	MTM program completion	0.123	(0.187)
BMI assessment	0.488	(0.131)	Medication reconciliation	0.772	(0.145)
Breast cancer screenings	0.133	(0.054)	Members leaving plan	0.256	(0.125)
Enrollment timeliness	0.426	(0.117)	SNP management	0.266	(0.272)
Improving bladder control	0.240	(0.094)			

Note: This table presents the coefficient estimated by measure in the triple-differences regression used to evaluate quality responses to design changes. The coefficient pools the response of plans with preexisting quality falling in 2 or 3 stars, relative to those of 4 stars. Standard errors in parentheses are clustered at the contract level.

1.9. *Scores as rankings*

Consumers who are imperfectly informed about scoring design might rely on the distribution of scores in their market to make additional quality inferences. In particular, consumers might use scores to rank their options, giving scores an ordinal interpretation. To evaluate this, I test whether the residual of individual demand in equation (2) of the main text systematically correlates with the local rank of a plan within its county. I define the local rank such that 1 indicates the highest scoring plan in the county (ties are assigned equal rank). A firm fixed effect is added to the residual regression to partially address selective entry by firms to counties with fewer competitors.[6] The estimated coefficient is -0.0013 with a standard error of 0.0012, indi-

[6]Removing firm fixed effects slightly increases the magnitude but does not change the result qualitatively.

cating that a higher-scoring plan is preferred. However, the effect is statistically insignificant, and its magnitude is negligible.

1.10. *Imperfect quality control*

MA features substantial cross-sectional variation in quality. Before their final discretization, the distribution of continuous ratings is well dispersed, as shown in Figure 9.[7] While it might appear contradictory for a firm to invest in quality in the interior of a rating interval—as consumers do not observe it—, there is a simple explanation for it. Investments in MA are contractual arrangements with providers and third-party services. Insurers can change their quality by restructuring their network and forming incentive contracts, but the final delivery of quality is rarely in their control. For example, an insurer can expand its physician network to reduce waiting times for primary care appointments. However, a harsh flu season can increase the burden on physicians and result in higher wait times than expected by the insurer when optimizing the network.

2. MODEL, IDENTIFICATION, AND ESTIMATION

2.1. *Pricing regulation and rebate allocation model*

CMS requires plans to allocate rebates among benefits or premium reductions (see Appendix I). Each plan determines a fraction of rebates to allocate to Part B premium reductions (κ^b_{jmt}), Part D reductions (κ^d_{jmt}), extra benefits (κ^e_{jmt}), and increasing consumers' coverage on standard Medicare-covered health care services ($1-\sum_{l\in\{e,d,b\}}\kappa^l_{jmt}$). Because premium reductions are payments directly to CMS or a transfer within the firm, they do not add to the firm's revenue. CMS strictly regulates improvement to standard Medicare coverage, requiring insurers to submit cost assessments based on CMS utilization models certified by actuaries. Because of this, I assume that plans offer these additional benefits at cost. The only remaining free source of rebate revenue is the κ^e_{jmt}. Therefore, the additional revenue of plan j in market m at year t, $R(p_{jmt}, \boldsymbol{z}_{jt})$ is given by the sum of its Part D premium (p^D_{jmt}) and any rebates allocated to additional benefits ($Rebate_{jmt}(p_{jmt})\kappa^e_{jmt}$). I assume that in the short run, each plan's rebate allocation fraction is an exogenous feature.[8]

The total premium consumers pay in the model (p^{total}_{jmt}) is a sum of their mandatory Part B premium (p^B_i), the MA plan's premium (p^C_{jmt}), and a Part D premium (p^D_{jmt}) if the plan bundles prescription drug coverage. The result is $p^{\text{total}}_{ijmt} = p^B_i - PR^B_{jmt} + p^C_{jmt} + p^D_{jmt} - PR^D_{jmt}$ where $PR^B_{jmt} = \kappa^b_{jmt}Rebate_{jmt}(p_{jmt})$ and $PR^D_{jmt} = \kappa^d_{jmt}Rebate_{jmt}(p_{jmt})$ are the rebate dollars allocated by the insurer to reduce consumers' part B and part D premiums.

Part B and D premiums are treated as exogenous attributes. The former is an individual-specific element, and the bundled prescription drug coverage plan determines the latter. The Part C premium is the sum of the basic and supplementary premiums. The basic premium is equal to the positive difference between bid (or pre-subsidy price) and benchmark (or subsidy kink), $(\max\{p_{jmt} - B_{jt}, 0\})$. The supplementary premium equals the fraction of additional Medicare cost-sharing plan benefits ($\bar{b}_{jmt}$) not financed by rebates, $\bar{b}_{jmt} - Rebate_{jmt}(p_{jmt})(1 - \sum_{l\in\{e,d,b\}}\kappa^l_{jmt})$.

[7] Star ratings of 1 and 1.5 are rarely observed. Ratings between 2 and 3 are less rare but not often provided by top insurers. Otherwise, large firms provide contracts covering the range of stars.

[8] For the 13% of plans without rebates in the data, I assume that all counterfactual rebates would go to cost-sharing standard benefits. This assumption's effect is minimal on consumer premiums and firm revenue and does not affect this stage's estimates.

The total benefits of a plan correspond to the sum of its mandatory TM benefits (b_0), additional Medicare cost-sharing benefits ($\bar{b}_{jmt}$), and extra benefits ($\kappa^e_{jmt} Rebate_{jmt}(p_{jmt})$). The first two types of benefits are treated as exogenous. The first is a regulatory level dictated by CMS, and the second is a product attribute. The extra benefit term varies with the rebate and depends on the plan's bid. Hence, both premiums and benefits are endogenous components of the model, albeit to different degrees.

2.2. *Modeled and unmodeled quality dimensions*

An important aspect of the model is that quality enters linearly in the indirect utility of choice, additively from the coverage generosity of the plan. Implicitly, this assumes that plan quality benefits consumers on the extensive margin of enrollment and care rather than on the intensive. The data reveals that most of the quality measured in MA lies in the extensive margin of care. For starters, out of the 46 positively weighted measures in 2015, only ten might be associated with the quality of services affecting the intensive margin of care, such as readmission rates and overall health improvements in the population. All others are related to care that can be easily provided in a single visit to a PCP, PT, or RN or are associated with customer service and insurer behavior. To further corroborate this, we could ask whether patients with higher expected utilization have stronger preferences for quality. This should be true in a model in which higher utilization consumers benefit more from quality. Table XI shows that consumers with higher predicted utilization or risk scores do not seem to have stronger preferences for highly rated products in various specifications. In light of this evidence, I chose to model quality in MA as an additive product attribute

Another relevant simplification of the model is that the only relevant quality measures are those assumed to be captured by the scores. To clarify, we can think about three types of quality dimensions a plan might have: those captured by the scores, those excluded from the scores but correlated with those included, and those excluded and uncorrelated. For example, the timeliness of needed care is included, which likely correlates with waiting time for primary care physicians, which has been excluded since 2012. These two are likely uncorrelated with the quality of addiction treatments covered by the insurer, which is neither measured nor included.

The demand model assumes that consumers agree with the regulator that these dimensions are related to more fundamental aspects of insurance, such as the quality of care or access to it. So, when consumers see a quality score that places some weight on the "access to care" category, they make an inference about their overall ability to access care. We can think about this as making an inference not only about the included timeliness of needed care but also about the correlated waiting time for physicians. The crucial assumption is that the scoring design does not determine this correlation between included and excluded-but-correlated qualities. Thus, when we recover consumers' preference for average access quality, we capture a stable metric of preference that we can then port to our counterfactuals. This assumption is consistent with the supply-side investment model and firms' incentives.

Regarding the excluded-and-uncorrelated quality dimensions, the model assumes that consumers' expectations of them are independent of the scoring design. This is consistent with the scoring and supply models by definition. The model decomposes the value of a network into its various components, such as its flexibility ("access"), the quality of its included hospitals ("outcome" and "intermediate" quality), and that of their physicians ("patient experience" and "process" quality). Consumers observe the score and make inferences about these aspects from it. Any aspect of the network that does not fall within these divisions is assumed to be uncorrelated with those that do and is captured in the fixed effect.

TABLE XI

HETEROGENEITY IN UTILIZATION AND RESPONSE TO STAR RATINGS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Highly rated (≥ 4) $\times$									
Predicted MA utilization	0.000376	0.00144	0.000974						
	(0.00445)	(0.00441)	(0.00442)						
Predicted any utilization				-0.00239	-0.000547	-0.00232			
				(0.00412)	(0.00408)	(0.00410)			
Risk score							0.0166	0.0205	0.0158
							(0.0142)	(0.0144)	(0.0143)
Highly rated (≥ 4)	0.204	0.183	0.199	0.228	0.199	0.227	0.196	0.181	0.196
	(0.0433)	(0.0429)	(0.0433)	(0.0419)	(0.0410)	(0.0419)	(0.0240)	(0.0243)	(0.0242)
Predicted MA utilization	-0.000637	-0.00183	-0.000332						
	(0.00419)	(0.00419)	(0.00420)						
Predicted any utilization				0.00264	0.00118	0.00245			
				(0.00401)	(0.00402)	(0.00402)			
Risk score							-0.00794	-0.00930	-0.00739
							(0.0105)	(0.0105)	(0.0105)
Product attr. controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Income and Demographic heterogeneity	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
WTP heterogeneity	No	No	Yes	No	No	Yes	No	No	Yes
N	55508	55508	55508	55508	55508	55508	55934	55934	55934
r2	0.0518	0.0523	0.0538	0.0518	0.0523	0.0538	0.0516	0.0521	0.0536

Note: This table shows the heterogeneity in response to star ratings across consumers with different levels of predicted medical utilization. The dependent variable is the enrollment indicator. The sample is restricted to new MA enrollees and MA plans. A high rating is a rating greater or equal to 4 stars. Predicted utilization includes log predicted normalized spending on all medical services. Predicted MA utilization limits utilization to inpatient, outpatient, and physician spending. Risk scores are computed based on linked claims and using public CMS software. Product attribute controls all observable product attributes, including the premium, benefits, and cost-sharing parameters. Demographic controls include health status, education, gender, race, age, and disability status. Income and demographic heterogeneity include interactions of high ratings with income and demographic variables. WTP heterogeneity includes interactions of premiums and benefits with consumer demographics at the same level as those used in demand estimation. Standard errors in parentheses are heteroskedastic robust.

2.3. *Consumers preference heterogeneity*

The MCBS provides exceptional data about consumers' demographics, chronic conditions, knowledge of the Medicare system, and other factors that might impact how they purchase healthcare. The demand estimation results provided in the article show that consumers are vastly heterogeneous in their preference for MA vs. TM and in their willingness to pay for quality. However, the model assumes they perceive and value quality signals equally across groups. To evaluate this assumption without imposing additional demand structure, I use the reduced-form exercise of Section IV of the main text. In particular, I evaluate all new MA consumers' responses to star ratings conditional on all other observable plan factors while accounting for heterogeneous preferences for the public outside option. I include two new factors in the regression: an interaction of a consumers' population group with a high-rating indicator (α_g^*) and a fixed effect for the population group (μ_g^G). The first variable captures the groups' heterogeneous preferences for higher-rated products, while the second accounts for differences in preferences for TM across groups.

$$\underbrace{y_{ijt}}_{\text{choice}} = \underbrace{\sum_g \alpha_g^* \mathbb{1}\left\{g(i) = g, r(jt) \geq 4\right\}}_{\text{het. quality preference}} + \underbrace{\sum_r \alpha_r \mathbb{1}\left\{r(jt) = r\right\}}_{\text{common quality preferences}}$$

$$+ \boldsymbol{x}_{jt}\boldsymbol{\lambda} + \mu_{m(i)} + \xi_t + \sum_g \mu_g^G \mathbb{1}\left\{g(i) = g\right\} + \epsilon_{ijt} \tag{3}$$

Table XII shows the results for 27 different binary population groups plus a first column for the entire population. The control variables are the same as in the descriptive analysis of the main text. The results show that, on average, consumers strongly prefer high-rated products. Consumers living in metropolitan areas and those who have completed college respond more to high-quality signals. In contrast, black consumers and the obese have weaker responses. There might be other weak preferences expressed in the results, but whose effect is either too small or too noisy to capture in this analysis. When put together (not shown), the main variable splitting heterogeneity across population groups is whether they reside in a metropolitan area. However, 98.2% of the relevant MCBS population lives in such an area, as MA is unavailable in many rural counties and the elderly population is scarcer in non-metropolitan areas.

These findings suggest that incorporating heterogeneity in quality preferences should not greatly impact the main analysis. If consumers interpret scores or value quality differently across population groups, this should be reflected in their responsiveness to scores, as measured in this exercise. There are, however, two situations under which this test could fail. First, if consumers have different beliefs and preferences about quality that exactly offset each other, resulting in the same aggregate behavior. Second, if consumers have heterogeneous underlying preferences for quality dimensions but have homogenous naive responses to scores. Both scenarios seem peculiar and farfetched but cannot be a priori ruled out.

2.4. *Model flexibility in quality provision*

The inefficiency in quality provision found in the main analysis plays an important role in the resulting scores. I consider a simplified version to show that the model is flexible enough to accommodate both over and underprovision of quality. To highlight the mechanisms, I ignore subsidies, investment risks, multidimensional quality, and multiproduct firm incentives. Consumers in the simplified model choose a plan to maximize an indirect utility $u_{ij} = \alpha_i p_j + \gamma \mathcal{E}[q|r_j, \psi] + \xi_j + \epsilon_{ij}$, where all terms are as in the main analysis. The utility of their outside option is normalized to zero, up to a logit error. Ignoring investment risk and focusing on single product firms, the insurer owning plan j chooses quality and prices to maximize $\pi_j(\boldsymbol{q}, \boldsymbol{p}, \psi) = D_j(\boldsymbol{q}, \boldsymbol{p}, \psi)(p_j - \theta' q_j - c_j) - \mu_j q_j^2$.

In this simplified model, under a scenario of full information and a monopolist firm, it is easy to find conditions under which quality is over or underprovided. In particular, by applying the logic of Spence (1975) and comparing the first-order conditions of the monopolist and the regulator, it is straightforward to show that quality will be overprovided if $\frac{1}{D}\int D_i \frac{\gamma}{\alpha_i} dF(\alpha_i) > \frac{\partial p}{\partial q}$, efficiently provided if this condition holds with equality, and underprovided if the inequality is reversed. On the left, $D_i \frac{\gamma}{\alpha_i}$ is consumer i's valuation for a marginal increase in quality, measured in units of premiums and weighted by her choice probability. Thus, the left-hand side of this equation is the weighted average valuation for marginal quality increases. On the right, we have the increase in price associated with a marginal increase in quality. This result is a direct analog to Spence's first proposition.

To illustrate that this condition can lead to different efficiencies of quality production, Figures 10a and 10b show simulations for two markets that differ only in their distribution of price preferences but generate opposing quality outcomes. Figure 10c shows that this extends to markets with more than one firm. In particular, it presents simulations for a duopoly market where consumers preferences are $\alpha_i \in \{1, 1.5, 3\}$, distributed with probability $[(1-x)/2, x, (1-x)/2]$, and $\gamma = 0.5$. Consumers have fixed unobserved preferences for firm 1 given by $\xi_1 = 2$, while their preference for the second firm, ξ_2, varies in the simulation. Firms have equal cost functions ($\theta = 0.2, \mu = 0.1, c = 0$). In the simulations, firm 1 always underprovides quality relative to the social optimum, but firm 2 might over or underprovide depending on consumers' preferences.

TABLE XII

HETEROGENEITY IN RESPONSE TO STAR RATINGS

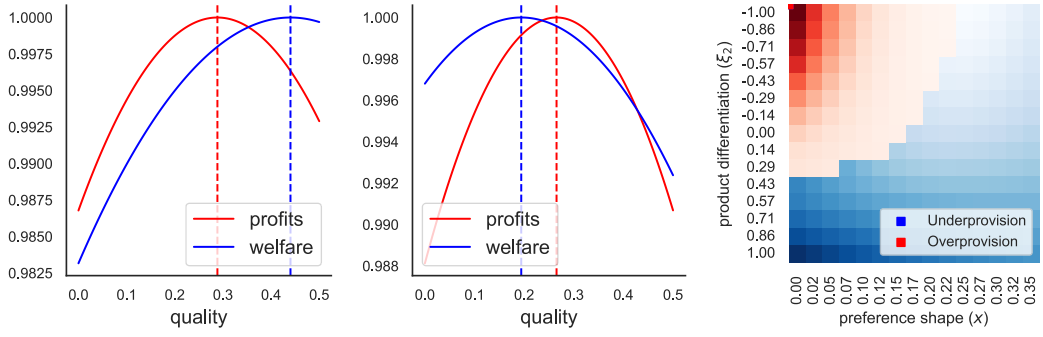
Group	All	In metropolitan	Medicare is easy to understand	Read handbook
Group x high rating	0.203	0.0506	0.00902	0.00306
	(0.0231)	(0.0138)	(0.00581)	(0.00567)
Group	Satisfied with Medicare info	College degree or higher	Graduated high school	Is asian
Group x high rating	0.00845	0.0125	0.00418	-0.0406
	(0.00665)	(0.00633)	(0.00724)	(0.0232)
Group	Is black	Is hispanic	Low income subsidy	Attended college
Group x high rating	-0.0217	-0.0102	-0.00886	0.00538
	(0.00939)	(0.01000)	(0.00810)	(0.00582)
Group	Has any chronic condition	Has complex chronic condition	Diabetic	Depressed
Group x high rating	0.000655	-0.000174	-0.00362	-0.00574
	(0.00902)	(0.00726)	(0.00662)	(0.00592)
Group	Smoker	Obese	Frequent faller	Has urinary control
Group x high rating	-0.00438	-0.0133	-0.00133	0.00133
	(0.0111)	(0.00631)	(0.0107)	(0.00601)
Group	ADL impaired	ADL challenged	Has routine checks	Married
Group x high rating	-0.0102	-0.00728	-0.000752	0.00820
	(0.00588)	(0.00785)	(0.00776)	(0.00595)
Group	Female	Predicted Utilization	Predicted MA utilization	Risk score
Group x high rating	-0.00289	-0.00239	0.000376	0.0181
	(0.00593)	(0.00412)	(0.00445)	(0.0143)

Note: This table shows the estimates of α_g^* in Equation (3). Each cell represents an independent regression under an alternative group definition. Chronic conditions include an enlarged prostate, hearing loss, Alzheimer's, amputated limbs, arterial hardening, arthritis, diabetes, broken hips, cancer, heart failure, angina pectoris, CHD, high cholesterol, high blood pressure, skin cancer, chronic valve problems, dementia, depression, COPD/asthma/emphysema, hypertension, retardation, myocardial infarction, osteoporosis, Parkinson's, paralysis, mental disorders, arrhythmia, and stroke. Complex chronic conditions exclude those treatable by medication: arthritis, high cholesterol or blood pressure, and arrhythmia. Frequent fallers suffered from three or more unexpected falls the previous year. Obesity marks BMI above 30. ADL stands for Activities of Daily Living (e.g., dressing, eating, bathing, walking, continence). Impairment and challenge refer to the inability or difficulty of executing at least one ADL. Predicted utilization includes log predicted normalized spending on all medical services. Predicted MA utilization limits utilization to inpatient, outpatient, and physician spending. All regressions have a sample size of 55934, except the last three, which have 55508 observations, given limitations on linked claims. Standard errors in parentheses are heteroskedastic robust.

In particular, it overprovides quality when consumers' WTP for quality is more heterogeneous and when ξ_2 is smaller.

2.5. *Demand identification - Instruments*

Insurers' knowledge of consumers' unobserved (to the econometrician) preferences for plans renders their bidding decisions endogenous in the demand estimates. As bids affect premiums and benefit levels, both factors must be instrumented when estimating demand. I use a benchmark instrument corresponding to the leave-one-out average of benchmark rates among all other counties where the plan is offered and a rebate instrument corresponding to minus the rebate share for plans bidding below the benchmark and one for those above. The benchmark instrument captures bidding incentives for the plan, as all its counties' benchmarks determine its effective benchmark and thus its subsidization. The exclusion restriction for this instrument



(a) Underprovision (b) Overprovision (c) Duopoly

FIGURE 10.—Efficiency with heterogeneous price preferences

Note: Figures (a) and (b) display scenarios of under and overprovision in monopolistic markets. Their common configuration is $\gamma = 0.5, \xi = 2.0, \theta = 0.2, \mu = 0.1, c = 0$. Price preferences take on values in $\{1.0, 1.5, 3.0\}$. In (a) the probability of each point is $[0.2, 0.6, 0.2]$, and in (b) $[0.4, 0.2, 0.4]$. Welfare and profits have been normalized to have a maximum of 1 for display only. Figure (c) illustrates the efficiency of quality in a duopolistic market. Red blocks indicate underprovision by the second firm, and blue blocks indicate overprovision. Lighter color tones indicate lower welfare losses from quality distortions.

assumes that yearly changes in consumers' unobserved preferences for each plan change independently across counties. The second instrument corresponds to the derivative of the firm marginal revenue from a marginal increase in its bid. Below the benchmark, a dollar increase in the bid translates to a loss equal to the rebate share, while above the benchmark, it translates to a dollar gain. Regulation implies that these increases must accompany changes in consumer benefits, thus linking the endogenous component and the instrument. The exclusion restriction for this instrument is that the regulator did not change the rebate shares of firms due to changes in consumers' unobserved preferences and that, conditional on the leave-one-out benchmark IV, variation in bidding above or below the benchmark is not driven by variation in unobserved preferences.

To examine the appropriateness of the instruments, I use three variables that describe the local TM market: the average risk score of TM enrollees in the county, their rate of enrollment in Part B, and their average cost. Each of these might be connected with the average consumer's preferences for insurance and, thus, with their unobserved preference for certain products. In addition, all three are excluded from the demand model. Panel A of Table XIII shows that MA bids are correlated with these excluded variables, which creates a potential for endogeneity. Panel B, however, shows that the instruments have a negligible correlation with these factors. This provides evidence supporting the exclusion assumption: the instruments are connected to bids via changes in regulation in other markets, which is not a function of variation in local market preferences for insurance plans. Table XIV shows the first stage of the instruments, showing their relevance.

Quality—as price and benefits— correlates with consumers' unobserved preference for plans. To address the resulting endogeneity problem, the demand estimation strategy relies on ratios of quality and scoring weights as instruments for scores. While the exclusion restriction on the scoring design instruments might seem natural, as the regulator is unlikely to consider variation in plan-specific preferences when designing scores, the quality ratio instruments might seem more peculiar. After all, scores are endogenous because firms' investments are made with some knowledge of consumers' unobserved preferences. To better understand why the ratios would be excluded from changes in consumers' unobserved preferences while the levels are not, we can look back at firms' investment problems, as stated in Section V.B.2 of

TABLE XIII

SUGGESTIVE TEST OF DEMAND INSTRUMENTS' EXCLUSION

	Risk score	Part B rate	FFS cost
Panel A: Endogeneity			
Log bid	0.123	-0.0498	386.3
	(0.00885)	(0.00382)	(14.64)
N	24235	24235	19862
R^2	e 0.522	0.490	0.620
Panel B: Exclusion			
Benchmark IV	-0.000525	0.000281	-0.0834
	(0.000275)	(0.000129)	(0.512)
Rebate IV	0.000997	-0.000130	3.899
	(0.00109)	(0.000551)	(1.706)
N	24235	24235	19862
R^2	0.517	0.486	0.594

Note: This table presents estimates of regressions that highlight potential endogeneity sources for bids in the demand model (Panel A) and suggest that instruments satisfy the exclusion restriction (Panel B). The dependent variables correspond to characteristics of the local TM market: the average risk score of TM enrollees, the share of TM enrollees enrolled in Part B, and the average FFS cost per enrollee. All regressions include contract-year fixed effects and non-bid-related product characteristics, as in the demand estimation model. Standard errors are heteroskedasticity-robust.

TABLE XIV

DEMAND ESTIMATION FIRST STAGE

	Premium			Benefits		
	I	II	III	I	II	III
Instruments						
Benchmark	-0.007	-0.002	0.007	0.0139	0.009	-0.008
	(0.001)	(0.001	(0.001)	(0.001)	(0.001)	(0.001)
Rebate	0.656	0.609	0.398	-0.317	-0.275	-0.162
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.002)
Other endogenous						
Benefits	0.782	0.836	1.288			
	(0.008)	(0.008)	(0.009)			
Premium				0.390	0.405	0.462
				(0.003)	(0.003)	(0.003)
Product controls	No	Yes	Yes	No	Yes	Yes
Contract-year FE	No	No	Yes	No	No	Yes
N	29033	29033	28863	29033	29033	28863
R^2	0.566	0.622	0.887	0.452	0.536	0.896

Note: This table reports the first-stage estimates of the second step of the demand estimation. All regressions include controls for other observable product characteristics and market fixed effects. Standard errors in parentheses are heteroskedasticity robust.

the main text. As noted, we can view firms' investment problem as deciding each contract's target rating and then picking a mix of qualities that minimizes their cost of attaining that rating. We can formalize the problem by first ignoring investment risk and the impact of quality on insurance marginal cost. The firm's problem then consists of choosing for each contract a rating r_c and then picking a point along the boundary of the set $\Psi_r = \{\boldsymbol{q} \in \mathcal{Q} | \psi(q) = r\}$. The optimal point for a firm along this boundary is the one that minimizes its cost $\boldsymbol{q}_c^* = \arg\min_{\boldsymbol{q} \in \partial \Psi_r} I_f(\boldsymbol{q})$. Given that the regulator's policy is (mostly) linear in category quality (as discussed below), the

boundary $\partial\Psi_r$ can be traced by following the ratios of any two dimensions of qualities, holding others fixed. Reintroducing investment risk and quality marginal cost does not change the fact that the chosen ratios are not a function of consumers' unobserved preferences for the contract, conditional on the choice of the target rating r_c.

We can conduct tests similar to the ones done for premiums and benefits to assess the validity and usefulness of these instruments. To examine the potential for endogeneity, I regress factors excluded from the demand equation on star-year fixed effects. These fixed effects capture consumers' preferences for scores each year and, thus, the source of endogeneity. To examine the exclusion restrictions, I regress the same excluded demand factors on the full set of instruments. As the source of endogeneity is assigned at the contract level and the excluded variables present variation at the plan-market level, I only include one observation per plan year to minimize overweighing the effect of more common plans. Given the large set of estimates associated with each regression, Table XV shows only a summary of the relevant aspects of the results. Panel A shows that the share of star-year fixed effects with a statistically non-zero (at 1%) coefficient varies between 3.5% and 57%. The joint test of star-year fixed effects being statistically zero can be rejected for two of the excluded factors at any reasonable confidence level. Therefore, star ratings correlated with the excluded demand factor, producing a potential for endogeneity. Panel B shows that the share of quality ratio instruments with statistically non-zero (at 1%) coefficients varies between 0% and 11% and that all joint tests can be rejected at 0.1% or higher. Therefore, the instruments do not appear to correlate systematically with the excluded factors. Panel C summarizes the first stage of the instruments, showing that 93.3% of all quality and scoring weight ratios are statistically significant when replacing the dependent variable for contracts' star ratings. The null of the joint significance test can be rejected at any reasonable confidence level.

2.6. *Star Ratings as monotone partitional scores*

The Star Ratings result from a weighted average of monotonic measure-level step functions, rounded to the nearest half. As the class of monotone partitional scores is closed under addition and positive scalar multiplication, the weighted sum of measure-level scores is monotone partitional. It is also easy to verify that rounding does not break the scores' monotonicity. Hence, the MA Star Ratings are monotone partitional.

However, CMS's design involves hundreds of measures and cutoffs that vary over time. Using it fully would require identifying consumers' beliefs and insurers' investment costs for each quality measure. This task is both untractable and likely not an accurate representation of reality. For example, it seems implausible that an insurer can invest in improving the rate at which physicians review their patients' medication while decreasing the rate at which the same physicians assess their patients' pain. CMS's quality categories capture this correlation and dependence across measures, forming a natural lower-dimensional space to evaluate quality.

The baseline design is, intuitively, easy to approximate at the average-category quality level. The star rating of a contract is the weighted sum of many step functions, rounded to the nearest half-star. Therefore, a smooth function could approximate each step function and have most of its fitting error rounded out. As the weights are category-specific and the distribution of quality within a contract is similar across firms, we could use a single smooth function per category to map average category-level investment to the total category score.

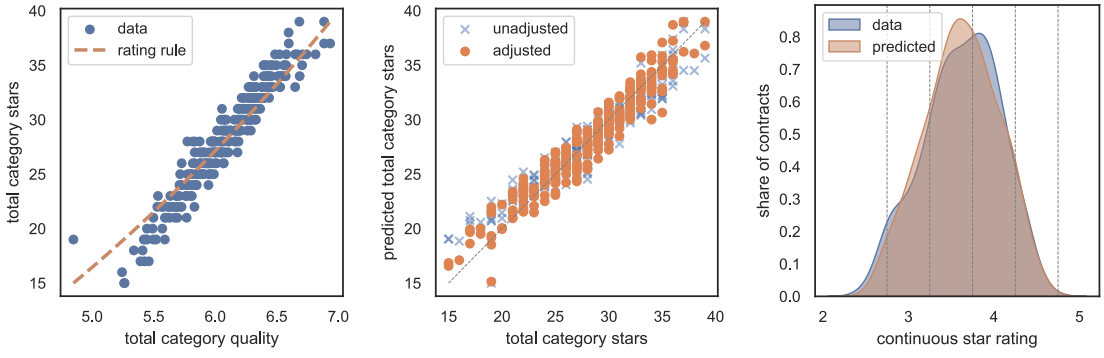
The approximation procedure follows this intuition: First, for each category year, I fit a bounded polynomial that maps the sum of each plan's qualities in a category onto the total sum of measure-level stars; Second, I capture minor systematic differences across firms by regressing the approximation error of step 1 on indicators for plan-type, state, firm, and year.

TABLE XV

SUGGESTIVE TEST OF SCORES INSTRUMENTS' EXCLUSION AND FIRST-STAGE

	Risk score	Part B rate	FFS cost	Star rating
Panel A: Endogeneity				
Share significant (1%)	3.571%	57.143%	53.571%	
joint p-value	22.400%	0.000%	0.000%	
N	14006	14006	11524	
R^2	0.480	0.431	0.554	
Panel B: Exclusion				
Share significant (1%)	0.000%	11.111%	0.000%	
joint p-value	89.300%	0.105%	26.700%	
N	14006	14006	11524	
R^2	0.480	0.429	0.552	
Panel C: First-Stage				
Share significant (1%)				93.333%
joint p-value				0.000%
N				16689
R^2				0.984

Note: Panels A and B of the table summarize the results of regressions that test for potential sources of endogeneity of scores in the demand estimation and the potential exclusion of instruments. The dependent variables correspond to factors excluded from the demand equation, but that might correlate with consumers' unobserved preferences. The share significant row presents the fraction of star-year fixed effects (panel A) or contract quality ratios (panel B) that are significant above 1% in the regressions. The joint p-values row shows the joint-significance test results for the star-year fixed effects (panel A) and quality ratios (panel B). Panel C shows results associated with the first stage of the instruments, where the dependent variable is each contract's star rating. The share of significant components and joint tests is among all instruments. All regressions include contract fixed effects and controls for other observable plan characteristics. Regressions are executed at the contract-market-year level.



(a) Step 1: Intermediate 2015 (b) Step 2: Intermediate 2015 (c) Step 3: Overall 2015 rating

FIGURE 11.—Scoring rule approximation

Note: These figures illustrate the scoring rule approximation used to reduce the complexity of the problem.

I also include the number of measures in the category and rating adjustment factors. The predicted value is added as a dispersion adjustment. Finally, I adjust category weights to maximize the fit, solving a least-squares problem subject to the constraint of positive weights. Figure 11 illustrates the three steps involved in this procedure.

When comparing this procedure's predictions with the data for each year, the R^2 ranges between 0.91 and 0.946. The model's maximum absolute error is only half a star, and 78% of plans get the same star rating as in the data. This remaining error is added to the adjustment factor so that the model predictions are exact in the baseline.

This approximation rule is only used three times in the paper. First, when estimating investment costs. The approximation error introduced there is relatively small as the rule is exact in the baseline, and only marginal changes are considered. Second, when estimating consumers' beliefs and preferences for quality under the informed choice assumption. This approximation is used to compute the quality domain that can achieve each rating in the baseline. It is mostly harmless and can be thought of as obtaining the quality partition associated with each score. Finally, I use it to simulate the baseline in the counterfactual analysis. This is where the approximation is truly leveraged, as I compute scores under the baseline rule for counterfactual quality outcomes.

The above approximation suggests that we can view the Star Ratings as the rounded weighted sum of five monotonic continuous functions, each taking average category-level quality as an argument. The sum of monotonic functions on different domains is strictly monotonic on the product domain, and the rounding maps the values to finitely many signals. Therefore, the Star Ratings at the category level are monotone partitional.

2.7. *Demand identification - proof of Proposition 1 in multiple dimensions*

The extension to multiple dimensions of quality follows from the following lemma.

LEMMA 1: *Let f, g be two distinct, continuous, strictly positive densities supported on $[0,1]^n$.$\forall w \in \mathbb{R}^n_+$ strictly positive, $\exists a < b < c \in \mathbb{R}_+$ such that either $\mathbb{E}_f[w'x|w'x \in (a,b)] \geq \mathbb{E}_g[w'x|w'x \in (a,b)]$ and $\mathbb{E}_f[w'x|w'x \in (b,c)] \leq \mathbb{E}_g[w'x|w'x \in (b,c)]$ with one of the inequalities strict, or the analogous statement hold with the roles of f, g reversed. Also, there exists another $a, b \in \mathbb{R}_+$ such that $\mathbb{E}_f[w'x|w'x \in (a,b)] = \mathbb{E}_g[w'x|w'x \in (a,b)]$*

PROOF: We can normalize one of $\boldsymbol{w}$ elements to 1 as all comparisons can be rescaled relative to it. Let $w_1 = 1$, and note that f and g define distributions over the random variable $y = \boldsymbol{w}'\boldsymbol{x}$. To find the implied distribution, define the linear map $\boldsymbol{y} = (\boldsymbol{w}'\boldsymbol{x}, x_2, \ldots, x_n) = W\boldsymbol{x}$, where W is an identity matrix with its first row replace by $\boldsymbol{w}$. Using this and the standard change-of-variables method, the distribution of $\boldsymbol{y}$ induced by the prior f is

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{1}{|det(W)|} f(W^{-1}\boldsymbol{y}) \quad \boldsymbol{y} \in \{\boldsymbol{y} \in \mathbb{R}_+ \times [0,1]^{n-1} : \sum_{i=2}^{n} w_i y_i \leq y_1 \leq \sum_{i=2}^{n} w_i y_i + 1\}$$

and so integrating over the marginals of $\boldsymbol{y}$

$$f_y(y) = \int_{V(y)} f(y - \sum_{i=2}^{n} w_i v_i, \boldsymbol{v}) d\boldsymbol{v} \quad V(y) = \{\boldsymbol{v} \in [0,1]^{n-1}, \boldsymbol{w}'\boldsymbol{v} \leq y\}$$

Where I used that $W = I + \boldsymbol{u}\boldsymbol{v}'$ where $\boldsymbol{u} = [1, 0, \ldots, 0]$ and $\boldsymbol{v} = [0, w_2, \ldots, w_n]$, which by the matrix determinant lemma implies that $|det(W)| = 1$. The implied distribution by g is defined analogously. Thus, we only need to prove that $f_y(\cdot)$ and $g_y(\cdot)$ satisfy the conditions of Lemma 1 from the main text appendix. First, note that both are positive and supported on $[0, \sum_{i=1}^{n} w_i]$. However, any closed interval of $\mathbb{R}$ is isomorphic to $[0,1]$, so continuity is the only property left to verify.

Let $\{y_n\} \in [0, \sum_i w_i]$ be a convergent sequence to y. Define the extension $\tilde{f} : [-\sum_{i=2}^n w_i, 1] \times [0,1]^{n-1} \to \mathbb{R}_+$ such that $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x})$ if $x_i \geq 0$, and $\tilde{f}(\boldsymbol{x}) = 0$ otherwise. Note that $\mathbb{1}\{\boldsymbol{v} \in V(y)\} f(y - \sum_{i=2}^n w_i v_i, \boldsymbol{v}) \leq \tilde{f}(y - \sum_{i=2}^n w_i v_i, \boldsymbol{v})$ for every $y, \boldsymbol{v}$. Also, $\tilde{f}$ is integrable as f is. Continuity follows from the dominated convergence theorem. *Q.E.D.*

Finally, we can state the proof of Proposition 1 in the multidimensional case.

PROOF: Fix a scoring rule slope w. Lemma 1 delivers partitions used for baseline and contradiction. With this, obtain ψ and $\tilde{\psi}$ such that $\mathbb{E}_{f_0}[w'q|r,\psi] < \mathbb{E}_{f_1}[w'q|r,\psi]$ and $\mathbb{E}_{f_0}[w'q|r',\psi] \geq \mathbb{E}_{f_1}[w'q|r',\psi]$ and as before

$$\gamma_0'(\mathbb{E}_{f_0}[q|r,\psi] - \mathbb{E}_{f_0}[q|\tilde{r},\tilde{\psi}]) = \eta_{rt} - \tilde{\eta}_{\tilde{r}} = \gamma_1'(\mathbb{E}_{f_1}[q|r,\psi] - \mathbb{E}_{f_1}[q|\tilde{r},\tilde{\psi}])) \implies w'\gamma_0 > w'\gamma_1$$

$$\gamma_0'(\mathbb{E}_{f_0}[q|r',\psi] - \mathbb{E}_{f_0}[q|\tilde{r},\tilde{\psi}]) = \eta_{r't} - \tilde{\eta}_{\tilde{r}} = \gamma_1'(\mathbb{E}_{f_1}[q|r',\psi] - \mathbb{E}_{f_1}[q|\tilde{r},\tilde{\psi}])) \implies w'\gamma_0 \leq w'\gamma_1$$

Which delivers the contradiction. *Q.E.D.*

2.8. *Connection to Abaluck and Gruber (2011)*

The demand estimates suggest consumers value a dollar in expected benefits similar to a $2.25 reduction in premiums. As benefits are related to expected spending given each plan's cost-sharing attributes, this result appears opposite to the findings of Abaluck and Gruber (2011) for Medicare part D. This conflict, however, is artificial. First, the cost-sharing benefits computed by CMS are for a unit-risk enrollee in TM. They are, therefore, related not to the expected OOP spending Abaluck and Gruber use but to their "Cost Sharing" index. Their estimations (Table 1) find that consumers value this index more than premiums.[9] Second, CMS's computation of benefit ignores moral hazard in spending. As consumers are more generously subsidized in MA than in TM, the benefit metric is underestimated, which inflates the estimated coefficient. Finally, the level and the variance of spending in MA differs from that of Part D, making the comparison difficult. To evaluate my findings against estimates from a similar insurance market, I use the model and estimates from Handel et al. (2015) and CMS's OOP calculator to compute the value of every plan in the ten largest counties in Texas during 2015. I find that, on average, a $1 increase in benefits for consumers equals a $5.6 reduction in premiums, an even larger preference for benefits than in my findings.

2.9. *Demand pre-subsidy plan price elasticity*

The regulatory environment requires that when firms increase their bids (i.e., their pre-subsidy plan price), they must also reduce their benefits or increase their premiums. For plans bidding below the benchmark, a marginal increase in bids reduces their rebates by a factor ρ_j. The rebate allocation model described above indicates that premium reduction benefits for consumers would fall by $\rho_j(\kappa^b_{jmt} + \kappa^d_{jmt})$ and that coverage benefit generosity would fall by $\rho_j(1 - \sum_{l \in \{e,d,b\}} \kappa^l_{jmt})$. If the firm was financing some of its rebates using supplementary premiums, the decrease in benefits lowers those costs. For plans bidding above the benchmark, a marginal increase in the bid translates to an equivalent increase in premiums. As most plans bid below the benchmark and finance benefits through rebate dollars and supplementary premiums, firms' effective pre-subsidy price elasticity is substantially different from the

[9] To quote them, "Individuals are willing to pay a price in premiums for desirable plan characteristics, but this price is insufficiently sensitive to their individual circumstance" (Abaluck and Gruber, 2011).

post-subsidy premium elasticity. That is, the price elasticity corresponds to $\theta_i p_{jmt}(1 - s_{ijmt})$ where $\theta_i = \alpha_i \frac{\partial p^{total}_{jmt}}{\partial p_{jmt}} + \beta_i((1 - \sum_{l \in \{e,d,b\}} \kappa^l_{jmt}) - \alpha_i(\kappa^b_{jmt} + \kappa^d_{jmt}))\frac{\partial Rebate_{jmt}}{\partial p_{jmt}}$, with the partial derivatives depending on whether the firm is bidding above or below the benchmark.

2.10. *Quality selection and manipulation*

It is reasonable to wonder whether insurers can affect quality measurements by selecting or attracting certain consumers. Arguably, one might expect that some beneficiaries comply better with preventive care or report a greater degree of satisfaction systematically. If insurers can selectively attract such consumers via other product attributes, then differences in quality across plans might be reflective not of true quality investments but of selection.

This problem was recently studied by Fioretti and Wang (2021), who used the introduction of the Quality Based Payment (QBP) program in MA in 2012. They ask if the QBP leads insurers to select consumers to enhance their scores and, if so, by which mechanism. Their main result is a four percentage point increase in the risk score of previously high-rated plans (pre-2012) in counties that operate in areas with below-median service risk. Therefore, the effect is small in magnitude and in the share of plans it affects. The authors find that this degree of selection affects overall ratings among their affected groups by an average of 0.23 stars; the average impact would be rounded out in the standard MA design.

Another way of examining the impact of selection is to look at the effect of adjustments made by CMS over time. In 2017, CMS introduced adjustments for population differences across contracts regarding their share of disabled, low-income, and dual-eligible populations. In 2018, CMS expanded the CAHPS case-mix adjustment to its HOS data. CMS and stakeholders determined these two changes to be particularly relevant to account for differences across plans' populations. If selection for quality purposes was prevalent before this change, we should expect stark changes in the ability of contracts to obtain scores. Figure 12 shows the distribution of changes in contract scores from one year to the next between 2014 and 2018. As can be seen, the distributions do not look different. The share of contacts without any change decreases slightly, which could be due to increased variability induced by the risk adjustment process. However, if plans selected their population before the policy, we should expect a large increase in the probability of decreasing ratings. The small change reflected in the picture is consistent with the small magnitude documented by Fioretti and Wang (2021). Overall, the figure suggests that selection might not have played a substantial role in delivering quality, at least at the margin identified by CMS.

There are various reasons why this extent of selection is limited. First, CAHPS measures are adjusted for population differences and so are a subset of other measures that CMS has identified as crucial to adjust, such as Plan All-Cause Readmissions from HEDIS. Second, selecting consumers on one dimension might deteriorate performance on others. For example, increasing premiums to select high-income, educated consumers to improve adherence metrics might decrease performance on outcome measures if the marginal benefit of care is higher in lower-income, underserved populations. Finally, selecting quality-altering populations might be less profitable and orthogonal to selecting over-compensated populations in risk adjustment (Brown et al., 2014).

2.11. *Investment risk*

The following describes the identification and estimation strategy of investment risk.

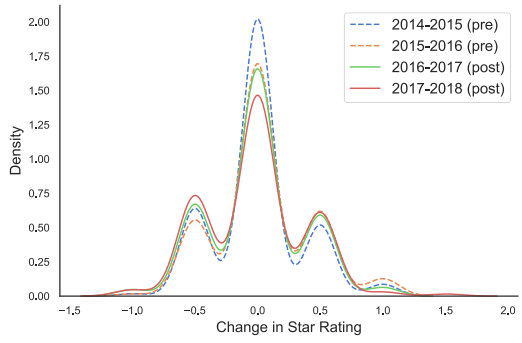


FIGURE 12.—Within-contract change in score before and after risk adjustment

Note: This figure displays the probability with which a non-exiting contract changes its score from one year to the next. The dashed line indicates the period before CMS incorporated extensive risk adjustment across quality measurement. Solid lines show the years following the adjustment.

2.11.1. *Identification*

Quality in MA is the outcome of firm contracts with internal agents (e.g., preventive care staff) and external providers (e.g., hospital networks), interacting with local shocks to enrollee's compliance and health. I express this idea through a mapping between the quality and investment, $q_{ckt} = \Phi_k(x_{ckt} + \epsilon^M_{m(c)kt} + \epsilon^F_{f(c)kt})$. In this expression, $\Phi_k(\cdot)$ is a known strictly increasing function, $\epsilon^M_{m(c)kt}$ is the market-level shock, and $\epsilon^F_{f(c)kt}$ the firm-level shock.[10] Market distortions include a harsh flu season or a community vaccination drive affecting the performance of all insurers. Firm distortions capture events such as cost shocks to provider contract negotiations or firm-level congestion in following up with patients. Given this structure, identification follows from standard results in the non-parametric measurement error literature (Schennach, 2016).

PROPOSITION 2: *(Identification of investment risk) Let $\boldsymbol{z}$ denote the vector of observables firms use to form beliefs about rival actions. Assume ϵ^M and ϵ^F are independent of each other, and $\boldsymbol{z}$, mean zero, symmetric, and have well-defined densities with nowhere-vanishing Fourier transforms. Then, the distributions of the errors are identified and given by*

$$f_{\epsilon^M_k}(\epsilon) = \mathcal{F}^{-1}\left(\mathbb{E}_z\left[\frac{|\mathcal{F}(f)_{\Delta^F \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/2}}{|\mathcal{F}(f)_{\Delta^M \Delta^F \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/4}}\right]\right)$$

$$f_{\epsilon^F_k}(\epsilon) = \mathcal{F}^{-1}\left(\mathbb{E}_z\left[\frac{|\mathcal{F}(f)_{\Delta^M \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/2}}{|\mathcal{F}(f)_{\Delta^F \Delta^M \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/4}}\right]\right) \qquad \forall k$$

Where $\mathcal{F}(\cdot)$ is the Fourier transform, Δ^M is the within-market across-firms difference operator, and Δ^F is the within-firm across-market difference operator.

[10] In estimation, I take $\Phi_k(x) = \Phi(x)(1 - \underline{q}_x) + \underline{q}_k$ where $\Phi(\cdot)$ is the standard normal CDF, and $\underline{q}_k$ is the minimum value of quality k that firms can produce. For example, an insurer cannot contract with a hospital to act in a way that would actively harm patients. The choice of $\Phi_k(\cdot)$ is not fundamental as it only provides an interpretation for the abstract unobserved investment.

PROOF: Denote $y_{jk} \equiv \Phi_k^{-1}(q_{jk})$ the normalized quality. The model implies that $y_{jk} = x_{jk} + \epsilon^M_{mk} + \epsilon^F_{fk}$, where x_{jk} is the unobserved investment. As this result is independent of the category, its index is omitted in what follows. Since the mean of x_j equals that of y_j, we can assume that all inputs are normalized to have mean zero. By the definition of z, $x_j|z$ is independent of $x_{j'}|z$ if different firms offer j and j'. Moreover, $y_j|z = x_j|z + \epsilon^M_j + \epsilon^F_j$.

Let (m, m') be two markets where firm f operates. Taking the within-firm difference we get, $\Delta^F y_f|z = \Delta^F x^*_f|z + \Delta^F \epsilon^M_{(m,m')}$. If there are multiple firms in (m, m'), this problem is analogous to a non-parametric unobserved measurement error problem with $N > 1$ repeated observations (Schennach, 2016). Applying the standard deconvolution to this class and using symmetry and mean-zero delivers $\mathcal{F}(f)_{\epsilon^M_{(m,m')}} = |\frac{\mathcal{F}(\Delta^F y_f|z)}{|\mathcal{F}(\Delta^M \Delta^F y_f|z)|^{1/2}}|^{1/2}$ where $\Delta^M \Delta^F y_f|z$ is the result of taking the difference of $\Delta^F y_f|z$ across two of the firms overlapping in (m, m'). Integrating the distribution of z on both sides and taking the inverse Fourier transform delivers the desired result. The result for the firm-level shock is analogous, taking first the difference within-market across firms. *Q.E.D.*

The intuition behind this result is that conditioning on firms' information set when investing is akin to conditioning on their investment choices. Consequently, any residual correlation in quality across firms within a market is driven by market-level quality shocks. Any residual correlation across markets within a firm is due to firm-level shocks.

2.11.2. *Estimation*

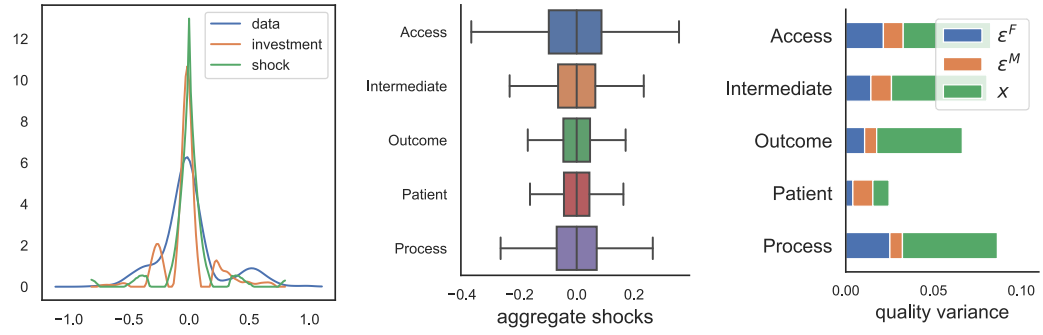
The quality shock distributions are estimated non-parametrically by solving the conditional deconvolution outlined in the proof above.[11] In the set of observables used by firms, I include benchmarks, rebate fractions, bundled services, plan types, market sizes, and means, variances, and correlations between the same set of variables for rivals. Additionally, I include indicators for the presence of the top ten firms (by all-time enrollment) in the market. Overall, the vector contains over a hundred attributes observable by firms when investing. However, firms likely use only a few of these to form beliefs as rivals' targets are only relevant insofar as they affect demand.[12] This observation suggests a sparse relationship between quality and the conditioning set that the estimator leverages. The estimation proceeds in three steps.

First, I assign *shock markets* to plans and form the data set used for estimation. I select markets with at least two firms that overlap in another market. Markets that fail to satisfy this condition are not helpful for estimation. Most contracts offered in these markets are found only in one market or have the vast majority of their demand in one. However, there is a fraction for which the assignment is less clear. To avoid specifying this manually and arbitrarily, I use the Resident-Matching algorithm with both markets' and contracts' matching preferences set according to observed enrollment. This way, these contracts are assigned to markets with less loss of valuable data. I then compute the observable market characteristics that enter the conditioning vector $\boldsymbol{z}$.

Second, I used the data to estimate the conditional distribution of differences in realized quality (i.e., the distributions of $\Delta^F x_f|z$, $\Delta^M x_f|z$, $\Delta^M \Delta^F x_f|z$). To estimate the non-parametric conditional density, I use the estimator of Izbicki and Lee (2017). Specifically, I use the FlexZ-Boost implementation of Dalmasso et al. (2020), which combines this estimator with the XG-Boost Machine Learning algorithm (Chen and Guestrin, 2016). I use a cosine basis with a

[11] This estimator does not build on previous estimates. Therefore, to offset the slower convergence rate of this class of non-parametric estimators (Horowitz and Markatou, 1996), I use the complete 2009-2019 data.

[12] For example, knowing that one competes against Humana, which systematically commands a significant market share, is probably enough to render the attributes of all smaller rivals irrelevant.



(a) Deconvolution (b) Shock distributions (c) Variance decomposition

FIGURE 13.—Quality deconvolution

Note: Figure (a) displays the deconvolution of the Outcome category. Figure (b) shows the distribution of investment risk by category. Figure (c) shows the fraction of observed variance in quality attributed to investments and the two types of shocks.

maximum of 30 elements to minimize the mean squared error of predictions. I use a 20% sample of random permutations of the data to train the algorithm and the remainder to tune it. This estimator returns an estimate of each density at a collection of points. I ensure the symmetry of the distributions, as assumed in the theorem, by averaging the recovered density and its reverse-order values. Figure 13a illustrates how the deconvolution separates the observed quality distribution into investment and shock distributions.

This estimation recovers the ten distributions modeled: two shock types for five categories, illustrated in Figures 13b and 13c. The results show that shocks account for 39.6% of the variation in observed quality. Patients' assessments are the noisiest, as insurers cannot contract for better reviews, and, correspondingly, most of the variance is due to market-level shocks. Firm-level shocks are most important in Process measures, as they are insurer-labor intensive and involve monitoring patients and their care.

2.12. *Firms' beliefs about rivals*

Firms in the model hold rational expectations about rivals' unobserved investment actions. These investments, however, are not in the data. Nevertheless, each insurer is affected by its rivals' investments only to the extent that they affect realized quality, the distribution of which is observed. While we could identify and use the distribution of rival qualities to evaluate each firm's expectations, this would be challenging to compute. First, because rival qualities are not independent, and second, because each firm would have to integrate over the realizations of hundreds of rival plans along five quality dimensions.

These challenges are alleviated using a lower-dimensional sufficient statistic for rivals' actions. Specifically, each firm is affected by its rivals exclusively through their effect on the joint distribution of the individual preferences of each consumer i and the average of rival utility values $v_i = \frac{1}{|\mathcal{J}_{-f}|}\sum_{j\in\mathcal{J}_{-f}} \exp(\tilde{u}_{ij})$. Since the demand model has only a handful of consumer types, this distribution is of lower dimensionality and can capture rich correlation patterns across rival investments.

To implement this idea, I estimate the joint distribution of $(\alpha_i, \beta_i, \boldsymbol{l}_i, v_{if})$, where $\boldsymbol{l}_i$ is the consumers lock-in status vector. I consider only groups of (α_i, β_i) that are statistically different from one another in the demand estimates. Additionally, I group lock-in statuses into either

"previously enrolled in MA", "previously enrolled in TM", or "new enrollee".[13] This results in 28 consumer types. Next, I use the demand estimates to compute v_{if} for each consumer, firm, and market and fit a lognormal probability distribution for each consumer type across markets.[14] I allow the parameters of the fitted distributions to depend on market characteristics. I use the same observables as when estimating investment risk, allowing the mean parameter to vary with the mean of these variables for rivals and analogously for the variance.[15] Finally, I combine the 28 marginals by fitting a Gaussian copula to the logarithm of rival values. I then draw 100 samples from the copula and use the inverse CDF of the fitted marginals to map these back to market-firm-level draws. At the end of the process, I obtain 100 independent draws for each firm-market and consumer type. These draws describe the estimated distribution, which I contrast with the aggregate rival-values data. The mean squared error of the fitted empirical CDF is $4.827 * 10^{-5}$.

It is worth noting, however, that the simplification of this approach comes at a cost. The dimensionality of v_i is smaller than that of the set of rival targets only if I do not condition on rivals' vector of non-quality attributes. However, the demand estimates indicate that quality, premiums, and benefits largely dominate v_i. The two last items are unknown to firms when investing in quality. Hence, the extent of firms' uncertainty goes beyond investments, which I capture through uncertainty over v_i. Nevertheless, the loss is in the connection between these beliefs and the equilibrium ensuing. Thus, this approach is an approximation that slightly extends the assumption regarding firms' beliefs to include uncertainty about other product characteristics determined after quality.

2.13. *Investment costs*

To estimate investment costs, I combine the estimated distribution of investment risk, the empirical distribution of quality in the data, and firms' investment optimality condition. Next, I explain how I derive the regression equation used in the main text and how the conditional expectation of marginal revenue can be calculated from identified distributions. Throughout, I omit time indices and denote $\boldsymbol{y}_f = \Phi^{-1}(q_f)$, the mapping of observed quality back into investment space; y_{jk} is the expected investment of contract j in category k given the data. Additionally, I denote $\pi_f \equiv \sum_m \int \mathbb{E}[V_{fm}(\boldsymbol{q}_f, \boldsymbol{q}_{-f})] dF(\boldsymbol{q}_f|\boldsymbol{x}_f)$ the expected insurance profit of the firm.

Assuming, for now, the differentiability of π_f, we can consider the marginal investment profit of firms as a random variable in the data, and decompose it as $\frac{\partial \pi_f}{\partial x_{jk}} = \mathbb{E}[\frac{\partial \pi_f}{\partial x_{jk}}|\boldsymbol{y}_f] + \nu_{jkt}$. Where ν_{jkt} is the mean zero conditional on the observed $\boldsymbol{y}_f$. Rearranging and using firms' optimality conditions, we obtain the regression equation used in the main text.

Now note that we can re-express π_f as

$$\pi_f = \int\int\int V_f(\Phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f), \boldsymbol{q}_{-f}) f(\boldsymbol{q}_{-f}|\boldsymbol{\epsilon}_m, \boldsymbol{z}) f_{\epsilon_F}(\boldsymbol{\epsilon}_f) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{q}_{-f} d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_f$$

[13] The insurer only uses v_{if} to predict the effect of changing quality on its future profits. This implies that dimensionality reductions of the parameter space are harmless as long as they do not systematically alter the likelihood that consumers will adopt or drop a plan as its quality changes.

[14] The data's distribution is remarkably similar to a log-normal with a mass-point near zero. The mass occurs when the firm is nearly a monopolist, facing only small rival plans. These markets are plentiful yet small in enrollment and their contribution to firms' profits.

[15] The average mean-square error (MSE) resulting from predicting the empirical cumulative distribution of rival values across all marginals was 0.00022.

$$= \int_{\epsilon_m} \int_{\epsilon_{-k}} \sum_{r \in R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})} \int_{e_r(x_{jk})}^{e_{r+0.5}(x_{jk})} \tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f | \boldsymbol{\epsilon}_m, r) f_{\epsilon_F}(\boldsymbol{\epsilon}_f) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_f$$

where $V_f = \sum_m V_{fm}$ is the sum of profits over markets, $\boldsymbol{z}$ are the observables that firm f uses to form beliefs about rival actions, $\boldsymbol{\epsilon}_m$ is the vector of market-level shocks in all markets and categories, and $\tilde{V}_{fj}(\cdot|\boldsymbol{\epsilon}_m, r) = \int V_f(\Phi(\cdot), \boldsymbol{q}_{-f}|r_j = r) f(\boldsymbol{q}_{-f}|\boldsymbol{\epsilon}_m, \boldsymbol{z}) d\boldsymbol{q}_{-f}$ is the firm's expected profit conditional on market shocks and a score, taking expectations over rivals' realizations. In the second equality, one dimension of firm-level shocks ϵ_{fk} was selected to partition the integral into segments that maintain the rating of x_{jk} constant. $R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})$ is the set of ratings that plan j can reach through different firm-level shocks in category k, given the other values of integration. As the shocks have full support, this set includes all possible scores assigned by the regulator. The integration limits associated with each partition are denoted by $e_r(x_{jk})$ and equal to $-\infty$ and ∞ for $r = 1$ and $r = 5.5$, respectively.[16]

We can evaluate the derivative by using the Leibniz integral rule and the envelope theorem

$$\frac{\partial \pi_f}{\partial x_{jk}} = \int_{\epsilon_m} \int_{\epsilon_{f,-k}} \sum_{r \in R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})} \Big(\tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + [\boldsymbol{\epsilon}_{f,-k}, e_r(x_{jk})] | \boldsymbol{\epsilon}_m, r) f_{\epsilon_{Fk}}(e_r(x_{jk})) -$$

$$\tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + [\boldsymbol{\epsilon}_{f,-k}, e_{r+1}(x_{jk})] | \boldsymbol{\epsilon}_m, r) f_{\epsilon_{Fk}}(e_{r+1}(x_{jk})) \Big) f_{\epsilon_{F,-k}}(\boldsymbol{\epsilon}_{f,-k}) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_{f,-k}$$

$$- \int_{\epsilon_m, \epsilon_f} \mathbb{E}[\sum_{m'} \gamma_{jm} D_{jm'}(\boldsymbol{p}_m(\boldsymbol{q}_m), r(\boldsymbol{q}_m)) | \boldsymbol{q}_f = \Phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f), \boldsymbol{z}] \theta_k \phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f) dF(\boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_f)$$

This expression, while complicated, has a simple interpretation. The first set of integrals corresponds to the change in profit due to increasing the probability of higher ratings and decreasing that of lower ratings. The second set of integrals is the change in profits associated with changing the marginal cost in the third stage. If service marginal costs were independent of quality, this term would be zero. Note that all of the densities involved in this equation are identified. The only unknown vector is $\boldsymbol{x}_f$.

The final step is in expressing $\mathbb{E}[\frac{\partial \pi_f}{\partial x_{jk}} | \boldsymbol{y}_f]$ as a function of identified densities. A change of variables shows that $\int_{-\infty}^{\infty} \frac{\partial \pi_f}{\partial x_{jk}}(\boldsymbol{x}_f) f(\boldsymbol{x}_f | \boldsymbol{y}_f) d\boldsymbol{x}_f = \int_{-\infty}^{\infty} \frac{\partial \pi_f}{\partial x_{jk}}(\boldsymbol{y}_f - \boldsymbol{e}) f_{\epsilon_F + \epsilon_M}(\boldsymbol{e}) d\boldsymbol{e}$ which can be evaluated given only the identified distribution of shocks and realized qualities.

2.14. *Challenges in adjusting investment cost standard errors*

The standard errors presented in Table 3, Panel B, in the main text do not account for the fact that the marginal return to investment incorporates estimates from previous stages of the game. In principle, the standard errors presented in the table should account for the added asymptotic variance introduced by these estimates. Doing so, however, proves challenging. This section outlines three key challenges that make adjusting the standard errors particularly difficult.

The first two challenges are theoretical. First, demand estimates affect the bidding stage equilibrium, over which the insurer must integrate when considering the marginal returns to

[16] For interior values, the boundaries are equal to $e_r(x_{jk}) = \Phi_k^{-1}(\psi_k^{-1}(r - 0.25 - \sum_{k' \neq k} \psi_{k'}(\Phi_{k'}(x_{jk'} + \epsilon_{m(j)k'} + \epsilon_{f(j)k'})) - \omega_j) - x_{jk} - \epsilon_{f(j)k}$. The inverse ψ_k^{-1} might not be unique. In this case, for the lower integration limit, we take $\inf \psi_k^{-1}$ and the supremum for upper integration limits.

investment. Changes to the demand parameters can result in changes to the bidding stage equilibrium and, thus, non-local changes in insurance profits. This challenge makes it difficult to derive an analytic expression for asymptotic variance for the estimator.

Second, the estimator includes investment-matching moments that effectively act as equality constraints on the estimated parameters. In principle, this can result in parameters lying on the boundary of the feasible set, which makes bootstrap methods unreliable (Horowitz, 2019).

Finally, the third challenge is computational. An insurer's marginal profitability of investment requires integrating investment risk for each of its insurance products over each dimension of quality. In order to have a precise value for the integral, I use a very large and dense quasi-Montecarlo grid. I leverage the fact that each integral value is independent and apply extensive parallelization to compute it. Relying on nearly one hundred CPUs, the integral requires over a week to compute and results in a large grid on disk. Applying subsampling-based methods for inference would, therefore, require large amounts of processing power and substantial disk space.

3. SCORING DESIGN

3.1. *Decomposition of monotone partitional scores*

The empirical design methodology relies on a decomposition of monotone partitional scores into a polynomial aggregator and a cutoff function. Formally, a monotone partitional score is an injective mapping ψ from a convex and compact quality space $\mathcal{Q} \subset \mathbb{R}^n$ into a finite ordered set $(\mathcal{A}, \geq_A)$, such that $q \geq q' \implies \psi(q) \geq_A \psi(q')$. A polynomial aggregator is a polynomial function mapping $\mathcal{Q}$ to $\mathbb{R}$, and a cutoff function is a weakly monotonic step function from $\mathbb{R}$ to $\mathcal{A}$. The following proposition proves the decomposition.

PROPOSITION 3: *Let $\psi : \mathcal{Q} \to \mathcal{A}$ be a monotone partitional score. Then $\forall \epsilon > 0$, $\exists m > 0$, a polynomial aggregator of order m, P_m, and a cutoff function K_m such that $\psi^* = P_m \circ K_m$ satisfies $||\psi^* - \psi||_{L_1} < \epsilon$.*

PROOF: Fix $\epsilon > 0$. Without loss of generality, $Q = [0,1]^n$ and $\mathcal{A} \subset [0,1]$. Assume $n > 1$ and that ψ takes on more than one value; otherwise, the proof is trivial. Every weakly monotone finite score partitions $[0,1]^n$ into a collection of finitely-many disjoint sets $\mathcal{M}$, such that ψ is constant over set $M \in \mathcal{M}$. Without loss, assume that the Lebesgue measure of every $M \in \mathcal{M}$ is positive; otherwise, the scoring set can be ignored under the L1 norm. Note that the boundary between every pair of contiguous sets is a set Δ_M of points, or equivalently, a line segment. Consider a small $\delta > 0$ such that the δ-neighborhood of each boundary $N(\Delta_M, \delta)$ do not overlap. Define the continuous function $f_\delta(x)$ as equal to ψ everywhere but in the boundary of the neighborhoods and equal to the linear interpolation between the steps of ψ across the boundary. Note that

$$||\psi - f_\delta||_{L1} = \int_{[0,1]^n} |\psi(x) - f_\delta(x)|dx =$$

$$\sum_{\Delta_M} \int_{N(\Delta_M,\delta)} |\psi(x) - f(x)|dx < \sum_{x' \in M} [\psi(\Delta_M+) - \psi(\Delta_M-)]\frac{\delta}{2}$$

Where $\psi(\Delta_M+)$ and $\psi(\Delta_M-)$ are the values of ψ above and below the boundary. Now as f_δ is continuous, by the Stone-Weierstrass theorem, there exists a polynomial P_m such that for all $x \in [0,1]$ we have $|P_m(x) - f_\delta(x)| < \frac{\epsilon}{2}$. Thus, picking $\delta < \frac{\epsilon}{\sum_{x' \in M}[\psi(\Delta_M+) - \psi(\Delta_M-)]}$ we have

that $||P_m(x) - \psi|| = ||P_m - f_\delta + f_\delta - \psi|| \leq ||P_m - f_\delta|| + ||f_\delta - \psi|| < \epsilon$. Finally, note that we can pick K_m to shrink the approximation error further. *Q.E.D.*

3.2. *Empirical scoring design implementation*

Each firm's insurance profit, $V_f(\psi, \boldsymbol{q})$ varies with $(\psi, \boldsymbol{q})$ only through their effect on scores, marginal costs, and the value that their rivals might offer in the market. Therefore, we can rewrite profits as $\tilde{V}_f(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$ where $\boldsymbol{w} = \boldsymbol{\gamma}' \mathcal{E}[q|\psi(\boldsymbol{q})]$ is consumers' expected quality-utility, $\boldsymbol{c} = C(\boldsymbol{q}, \boldsymbol{z})$ are marginal costs, and $\boldsymbol{v}$ are rivals' value as defined in Appendix II (see Firms' beliefs about rivals). As the space of qualities is convex and compact and marginal costs are continuous functions of quality, $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$ lie on a convex compact set. Moreover, the implicit function theorem implies that $\tilde{V}_f$ is differentiable and, as marginal costs and demand are bounded, its first derivative is bounded. Hence $\tilde{V}_f$ is Lipschitz continuous, which implies strict bounds on the approximation error from linear interpolation within a finite grid on the space of $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$, and that as the number of grid points increases (uniformly over the domain), the linear interpolation of $\tilde{V}_f$ converges uniformly to the true value. This grid, $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v}) \mapsto \tilde{V}$, and its interpolation, is the key to simplifying the computation of the designer's problem. Crucially, the grid is computed only once, and each point can be solved independently in parallel.

Using the grid of equilibrium insurance profits, we can evaluate the total welfare of the market at any scoring rule ψ through the following steps. First, given ψ, compute consumers expected quality at each score and interpolate across the computed $\boldsymbol{w}$ to form a new grid $(\boldsymbol{w}(\psi), \boldsymbol{c}, \boldsymbol{v}) \mapsto \tilde{V}$. Second, initialize firms' beliefs about rivals to some initial distribution $G_f(\boldsymbol{v}_{-f})$ (e.g., the empirical distribution of baseline rival values). Third, find each firm's optimal investment choices. As $\tilde{V}_f$ is precomputed, this consists of solving

$$\max_{\boldsymbol{x}_f} \sum_{n=1}^{N_w \times N_c} \sum_{l=1}^{N_v} \tilde{V}_f(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n, \boldsymbol{v}_l) \mathbb{P}_F(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n | \boldsymbol{x}_f) G_f(\boldsymbol{v}_l) - I_f(\boldsymbol{x}_f, \boldsymbol{\mu}_f)$$

where $\mathbb{P}_F(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n | \boldsymbol{x}_f)$ is the probability of expected quality-utility and marginal costs as determined by the investment choice and the estimated investment risk distribution, F. In the sum, N_d denotes the number of grid points over each dimension d in the grid. This problem is differentiable and can be solved independently and in parallel for each firm. Third, update firms' $G_f(\boldsymbol{v}_{-f})$ to be consistent with these choices and iterate back to point two until convergence. Finally, I use the equilibrium investments to compute expected profits and the equilibrium bids (by-products of computing the grid of expected insurance profits) to compute expectations over consumers' surplus.

Given that these expectations require integrating the investment risk of firms, it involves a high dimensional integral. The dimensionality is large because it involves integrating the investment risk of each category. To make this computation effective and precise, I compute the distribution of the vector of shocks and integrate over it. It is simple to compute this distribution using convolutions of the estimated non-parametric shock distributions. This approach reduces the dimensionality of the shocks in half, as firm and market shocks are aggregated, allowing me to use more precise quadrature techniques for multivariate distributions rather than rely on unguided multidimensional quasi-Montecarlo techniques.

I make three simplifying assumptions when solving this procedure. First, I hold the lognormal shape restriction on firms' beliefs about rival actions fixed, changing only their means to satisfy rational expectations. Second, I hold consumers' prior fixed, only changing their posterior belief according to ψ. Third, I grid potential investment choices to accelerate finding optimal investments. I interpret the first two choices as an expression of the short-run effect

of changing the scoring rules. Evaluating the long-run effects would involve specifying how a firm's belief distribution is formed and a process for adjusting consumers' priors. Additionally, it would require considering the dynamic effects of quality investment, which is beyond this work's scope. The third choice is purely for computational convenience.

The methodological steps above make evaluating the designer's objective feasible. However, exploring the space of solutions is burdensome as the objective is not everywhere differentiable and features flat regions where changes in cutoffs or weights do not affect firms' choices. To find the global constrained optima of this function, I use the algorithm of Malherbe and Vayatis (2017). This algorithm is guaranteed to find the global optima of Lipschitz functions of unknown finite Lipschitz constant over a compact convex space with a non-empty interior.

3.3. *Limitations and potential extensions of the methodology*

This article's empirical scoring design methodology can solve problems beyond those considered here. To recap, the method searches the space of all coarse monotone partitional scores by splitting policies into aggregators and cutoffs. It addresses the computational burden associated with evaluating the expectation of welfare given the stochastic realization of equilibrium quality by precomputing a grid of subgame outcomes and then averaging them according to the distribution implied by the evaluated policy.

As done in the main analysis, the methodology can be easily extended to search over designs that fully reveal quality within scores (e.g., top-revealing design), designs that consider preference heterogeneity over quality, and those that incorporate regulatory uncertainty over costs. As shown below, when discussing the robust scoring problem, the method can also solve the scoring design problem, which is subject to uncertainty over consumers' preferences. In addition to what has been shown, it is easy to adapt the method to search over constrained but discontinuous aggregators. For example, it is easy to search over the space of minimum-quality certifications by simply holding the aggregator fixed as the minimum function. It is also simple to use the method to explore designs that incorporate only some quality dimensions or a restricted amount of scores.

The main limitation of the method is its restriction to deterministic scores. It cannot be used to evaluate stochastic assignments such as rankings or contests. It also cannot solve scoring designs whose cutoffs are a function of quality realizations, such as those adopted by CMS starting in 2016. There are two reasons for this limitation. First, such designs are not ex-ante determined and, therefore, cannot be decomposed into aggregators and cutoffs independent of the market realization. Second, stochastic policies result in different types of equilibria than those evaluated here, which would require a different set of tools to handle. In particular, contests might often result in no equilibrium in pure investment strategies.

3.4. *Bounds on government spending*

The designs evaluated in the paper assume that consumers who substitute into MA are as risky as the average TM consumer. There is, however, abundant evidence that this might not be the case. Table XVI revisits the changes in estimated total subsidy spending per MA enrollee without selection and with the estimates of Curto et al. (2021) that switchers into MA are 2.3% less expensive than stayers. The adjusted estimates account for differential selection into MA across consumers of different risk scores and within risk scores across consumers of differential spending.

The results show a substantial difference in estimated governmental spending with and without adjustment. These numbers, however, have to be evaluated with caution for three reasons.

TABLE XVI

GOVERNMENTAL SUBSIDY SPENDING CHANGES ACROSS DESIGNS

	Full info	Optimal	Certification	CMS-Cert	CMS-Full	Top-Revealing (9)
Unadjusted	1.154	12.571	15.485	-2.403	-5.573	11.740
Adjusted	89.111	160.450	142.163	55.306	50.298	163.897

Note: This table shows the change in governmental subsidy spending across the different proposed designs relative to the status quo. The numbers are per Medicare beneficiary year. The unadjusted numbers match those in the main text and assume that any consumer substituting into or out of TM leaves the county's average TM risk score unchanged. The adjusted numbers compute governmental spending at the individual level and inflate the cost of consumers remaining or substituting into MA by 2.3%, conditional on risk score.

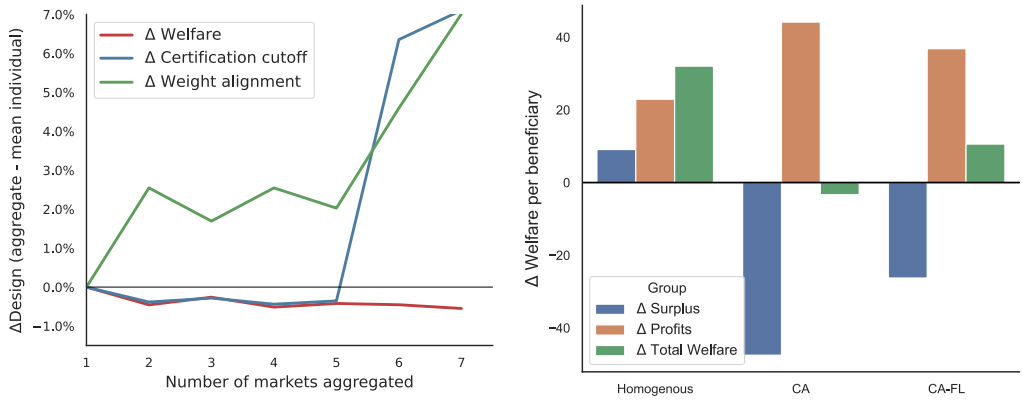
First, they assume that counterfactual selection into MA in this simulation matches the estimated selection on the margin, estimated by Curto et al. (2021). It is well possible that selection would happen differently in these counterfactuals, as what attracts consumers to MA differs from what drives selection in Curto et al. Second, there is evidence that advantageous selection into MA is decreasing (Newhouse et al., 2015). The estimates in Curto et al. (2021) are derived from an older sample than that used in this article. Therefore, selection in 2015 might be different than these estimates. Finally, a significant part of the cost increase stems from high-risk consumers in low-served markets switching back from low-quality MA plans to TM. The advantageous selection estimates do not speak to this margin of selection out of MA.

3.5. *Preference heterogeneity*

The optimal design problem aggregates the preferences of heterogeneous consumers to form a uniform national scoring policy. Consumers and firms, however, differ across markets and might benefit from different designs. To explore the impact of this national aggregation, I isolate seven markets and compute the optimal certification design for each individually and then in increasingly larger groups. The seven markets are selected to be disconnected from each other in the graph formed by contract presence across markets. This ensures that spillovers or complementarities across markets do not drive the optimal design that emerges from aggregation. Figure 14a shows the results.

As optimizing at the market level can only improve welfare relative to the aggregate design, aggregation always produces a welfare loss. This loss, however, is mild, with about 0.5% of total welfare lost due to preference aggregation. The loss also increases very slowly as aggregation grows, suggesting that the total surplus loss from aggregation is small. The optimal certification design, however, does seem to change substantially with aggregation. In particular, as more markets are added, the optimal aggregation weights become better aligned with consumers' quality preferences, and the cutoff for certification increases. This change is driven by increased heterogeneity across firms, which can be understood using Figure 5b of the main text: As more heterogenous iso-cost curves are drawn to meet the scoring threshold, the increased investment cost imposed by aligning BE with DC on some firms (as in the drawing) will be offset by a lower cost on others (those whose tangency point are on AD).

To complement this exercise, I evaluate the potential impact of allowing some consumers to have heterogenous quality preferences. I simulate this alternative scenario by picking three large states, Florida, Massachusetts, and California, and gradually changing consumers' preferences in these markets to match the regulator's weights. The CMS-equivalent preferences are scaled so consumers' total WTP for maximum quality remains unchanged. For each scenario, I solve for the optimal certification design across the three states and show the welfare changes in Figure 14b. The first set of bars shows the welfare change if all markets have the same homogenous preferences as in the baseline. As can be seen, the welfare gains of the optimal certification in these states are smaller than the average gains presented in the main text,



(a) Geographic heterogeneity (b) Simulated preference heterogeneity

FIGURE 14.—Effect of heterogeneity on scoring design

Note: These figures illustrate how the optimal design depends on consumers' preference heterogeneity. The exercise underlying Figure (a) considers seven non-overlapping markets and computes the optimal certification design for each market individually and as they are aggregated into a single policy. The red line shows the percent of welfare lost from aggregation. The blue and green lines show how the mean certification threshold and the alignment of weights with consumers' preferences change with aggregation. Alignment is computed as one minus the squared root of the mean squared error between weights and preferences. Figure (b) shows the welfare gains from optimal certification per member in Massachusetts, California, and Florida only. The homogenous set considers welfare changes under optimal certification and homogenous preferences as in the main analysis. The second set of bars shows the welfare change under optimal certification if only consumers in California have preferences that exactly match CMS's weighting scheme, and those in Massachusetts and Florida have the estimated homogenous preferences. The final set of bars shows the results if consumers in California and Florida have CMS-like preferences and only those in Massachusetts have the estimated preferences.

as these markets tend to be more competitive and, therefore, have lesser distortions in quality. When only one large state (CA) has CMS-aligned preferences, the optimal certification scheme reduces total welfare. This is because of the welfare loss from heterogeneous market aggregation, as discussed above, and because a single cutoff cannot provide efficient matching between heterogeneous consumers and products. However, when California and Florida have CMS-aligned preferences, their scale is enough to tip the balance in favor of a specific design, increasing total welfare. Nevertheless, this design requires some redistribution across states, as the average consumer is worse off, driven by those in Massachusetts.

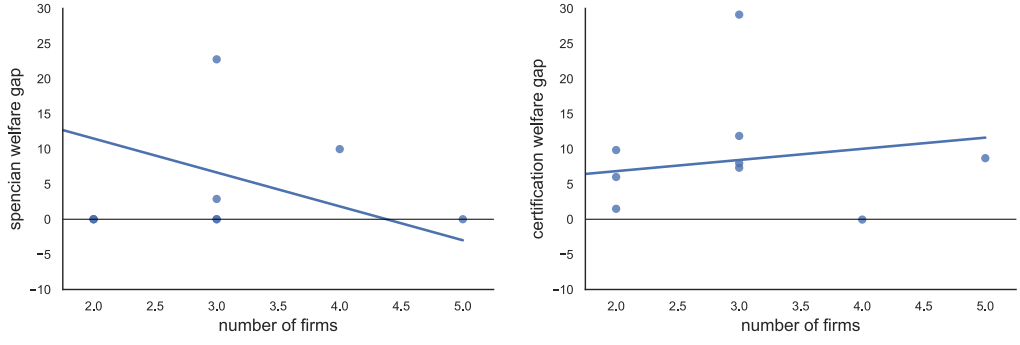
Overall, these results show that preference heterogeneity might have important implications for the optimal granularity of the design and the cost of imposing a uniform scoring policy for the entire country. Given estimates of consumers' preference heterogeneity, the methodology developed here could be used to find the optimal scores for the population. The answer in this scenario likely entails different scoring systems for different markets, each respecting the fundamental properties of the optimal scoring design outlined in the main text. Given limited evidence of preference heterogeneity in the MA setting and the lack of variation in the data to speak to the impact of heterogenous scoring policies across markets, a further evaluation of this heterogeneity would likely entail an unwarranted and unguided extrapolation.

3.6. *Competitive effects*

The main results suggest that the optimal design's main avenue for improving welfare is tackling the Spencian distortion and inducing additional quality investments. The spencian distortion, however, is a product of market power and thus should theoretically vanish as more firms compete for enrollment in the market. This poses the question of whether there is a de-

gree of competition, after which coarsening information through scores is no longer valuable. To explore this, I solve for the optimal certification design for 79 markets of different levels of competition. I also compute the full information and the second-best outcomes for each market, defining the latter as the outcome produced if the regulator could dictate each firm's quality and then fully reveal it to consumers.[17] I then compare the difference between the second-best welfare and the full information outcomes, which captures the Spencian distortion. I also compare the wedge between optimal certification and full information outcomes, which captures the gains from coarsening information.

The results, shown in Figure 15, indicate that as the number of firms in the market grows, the Spencian distortion vanishes. At about 4.4 firms per market, the wedge between the second-best and full information closes as firms' incentives to underinvest in quality have competed away. However, this increased efficiency in quality production does not eliminate the value of coarsening information. The wedge between optimal certification and full information outcomes grows moderately as more firms are added to the market. Even without market power distortions on quality, certification can reproduce similar outcomes while reducing vertical differentiation and, thus, market power over prices. Markets that support more firms also support more product variety on non-quality dimensions and larger enrollment pools. These factors increase the value from intensifying price competition among the firms, offsetting the moderate losses in information and quality incurred by certification. Therefore, while the Spencian distortion and the quality-regulation value of coarsening information vanish with competition, markets that support more competition can still benefit from the market-power regulating value of scores.



(a) Spencian gap

(b) Certification gap

FIGURE 15.—Effects of added competition on scoring design value

Note: These figures show the results (binned) from solving for the optimal certification design, full information benchmark, and second-best outcomes for a collection of 79 markets of different numbers of competitors. The second-best is defined as the welfare-maximizing (consumer surplus + insurer profits) choice of quality for each firm under full information. Figure (a) shows the additional welfare from the second-best relative to full information in dollars per member year. This corresponds to the Spencian distortion. Figure (b) shows the welfare gains of optimal certification relative to the welfare value of full information on the same scale.

[17] I call the regulator-dictated quality outcome the second-best as it does not allow the regulator to dictate prices nor optimally regulate matches between insurers and enrollees.

TABLE XVII

SCORING DESIGN UNDER THE POLICY-IGNORANCE ASSUMPTION

	Δ Consumer surplus	Δ Firm profits	Δ Total welfare	Δ Gov. spending
Known Preferences	62.71	103.49	166.20	0.02
Unknown Preferences (robust)	-48.56	46.32	-2.24	0.03

Note: This table presents the results of optimal scoring design under the assumption that consumers are policy ignorant. The first row presents the welfare change implied by the optimal linear design, assuming the regulator knows consumers' quality preferences. The second row presents the worst-case total welfare of the optimal linear robust design relative to the worst-case total welfare of the baseline design. Welfare is presented in thousands per member-year. Total welfare ignores government spending.

4. RESULTS UNDER POLICY IGNORANCE

This appendix presents results under an alternative to the informed-choice assumption (main text Assumption 1). In particular, this assumption is replaced with the following, stating that consumers are fully ignorant of design variation.

ASSUMPTION 1—Policy-ignorance: *Consumer's posterior beliefs $\mathcal{E}[\boldsymbol{q}|r,\psi]$ are exogenous, independent of ψ, and bounded in $\mathcal{Q}$*

Under this assumption, Proposition 1 in the main text no longer holds. However, a weaker version persists.

PROPOSITION 4—Quality beliefs and preference identification under ignorance: *Let assumption 1 and Assumption 2 (from the main text) hold, then there is a nontrivial identified lower bound for $\boldsymbol{\gamma}$.*

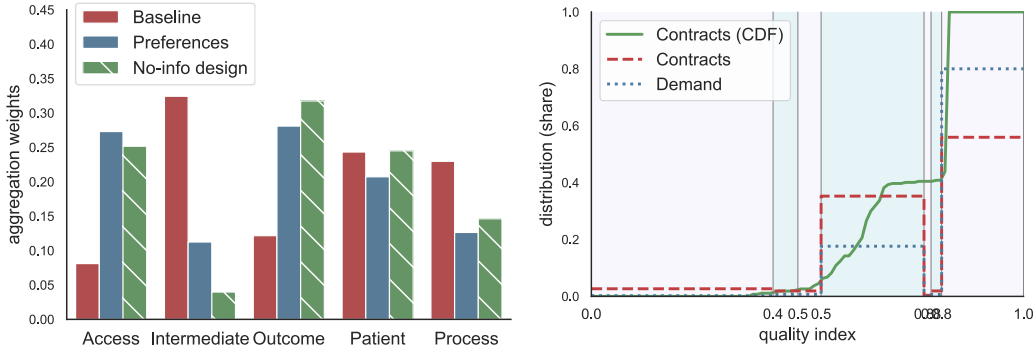
PROOF: As quality is bounded within the unit hypercube, and differences in WTP are identified, the lower bound is obtained by taking the smallest difference in WTP across stars, $\underline{\Delta}\eta$. The lower-bound set is given by $\underline{\Gamma} = \{\boldsymbol{\gamma} \in \mathbb{R}_+ | \boldsymbol{\gamma}'\mathbf{1} = \underline{\Delta}\eta\}$. *Q.E.D.*

As noted in the main text, substituting the assumption of informed choice with policy ignorance affects only the estimates of consumers' preferences for quality and their prior beliefs. Consumers' premium and benefit preferences and all supply-side estimates are invariant to this assumption and thus are used in this appendix without modification.

4.1. *Optimal design under known preferences*

In this problem, the regulator knows consumers' quality preferences (e.g., through external surveys) but cannot change their perspective on what the star ratings mean (as consumers are policy-ignorant). Instead, consumers' WTP for ratings is fixed at its estimated baseline value. The first row of Table XVII and Figure 16 show the results.

The regulator's problem, in this case, can be thought of as labeling products using the Star Rating system, assigning each the "reputation" (i.e., subjective expected quality) associated with a baseline rating. The only constraint on this design is that the labeling must be weakly monotonic in quality. This labeling technology vastly improves the power of pooling qualities at the bottom, as the regulator can now assign all low qualities to the reputation of a baseline 1-star plan regardless of the breadth of the bottom scoring interval. The main downside of eliminating the informational channel is that the regulator cannot offer high-quality plans a reward greater than the perception of a baseline five-star plan. Therefore, relative to a regulator



(a) Aggregator

(b) Cutoffs

FIGURE 16.—Optimal design without informational value

Note: These figures display the optimal design if consumers' interpretation of star ratings is fixed at its baseline value and the regulator knows consumers' quality preferences. Figure (a) shows this design's optimal aggregator for quality dimensions. Figure (b) presents the cutoff placement along the aggregated quality index.

facing sophisticated consumers, one facing naive consumers can improve quality at the bottom of the distribution more and qualities at the top of the distribution less.

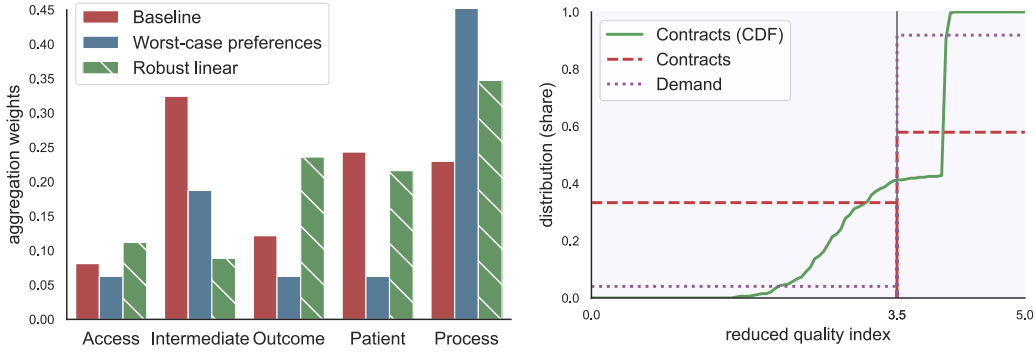
In total, welfare is greater when consumers are naive because the regulator's ability to coordinate demand to offset supply-side distortions is improved. The loss in information value is mitigated because the regulator is concerned with consumers' true valuation for quality and not their naive WTP for ratings. By labeling products correctly, the regulator can ensure consumers with higher WTP for quality are matched with higher-quality products. However, it must be noted that the evidence from the main text, Section IV, rejects the assumption of ignorance. Thus, this exercise provides a bound on the gains from redesigning the system rather than an alternative plausible design. If consumers are partially informed, it stands to reason that the gains from redesign lie between these results and those of the main analysis.

4.2. *Robust Scoring Design*

An important caveat to the prior analysis is that the informed choice assumption was used to identify consumers' quality preferences. Without external data and under the assumption of ignorance, the regulator can only infer a lower bound on consumers' preferences and cannot affect their quality beliefs. However, the regulator has already observed the impact of scores on enrollment and can leverage their assignment to marshal demand and induce quality investments. Knowing only that consumers' preferences are in some set Γ, the robust scoring design objective is to maximize total welfare under worst-case preferences:[18]

$$\max_{\psi\in\Psi}\min_{\gamma\in\Gamma}\int \Big[\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\ \text{surplus}}} + \underbrace{\rho^F\sum_f V_f(\psi,\boldsymbol{q}) - I_f(\boldsymbol{x}_f^*(\psi))}_{\substack{\text{Insurer}\\ \text{profit}}} - \underbrace{\rho^G Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\ \text{spending}}}\Big] dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \quad .$$

[18]Preferences γ are only relevant for consumer surplus because conditional on identified fixed valuations for scores, the demand is independent of consumers' preferences for quality.



(a) Aggregation weights

(b) Quality distribution and cutoffs

FIGURE 17.—Robust certification design

Note: Figure (a) shows the optimal robust and average baseline aggregators and the worst-case preferences under the optimal weights. Preferences are not inversely proportional to weights due to non-uniform identified set bounds. Figure (b) shows the cutoff locations and contract distribution.

This cautious approach matches the decisions of an imperfectly informed regulator that risks significant political or legal losses from implementing a new design that worsens outcomes. Importantly, this objective would remain the same if consumers had heterogeneous preferences for quality. The interior minimization is equivalent to a linear equilibrium constraint, which enables the use of the empirical design methodology.

Figure 17 shows the solution to this problem, and the second row of Table XVII shows its worst-case welfare.[19] The results reveal that the non-linear baseline design slightly outperforms even the best linear robust scoring policy. The fact that a non-linear score would perform better is not surprising in the robust design problem as it consists mainly of labeling qualities without considering the implications of the labeling policy on information or expected quality. A linear aggregator imposes more restrictions than a nonlinear aggregator on the potential labelings the regulator can perform. Nevertheless, it is striking that the baseline design is as effective in solving the robust design problem.

These findings suggest that the regulator might operate under the assumption that consumers are naive and subject to severe penalties for misrepresenting their preferences. This might also rationalize why CMS adopted increasingly more complex designs in MA starting in 2016 and in other regulated markets, such as hospitals. If consumers have naive and fixed preferences for higher-scoring products, the complexity of the mapping between quality and scores is not associated with any loss of information or welfare costs. In fact, it can only benefit the regulator by enhancing its ability to label qualities. However, it is important to note that the underlying assumption of ignorance is easily rejected in the data.

4.3. *Discussion*

Overall, these analyses complement the previous sections in three ways. First, they disentangle the mechanisms by which scores affect the market. In particular, designs in the main anal-

[19]To discipline the worst-case preferences, I restrict γ between half the lowest estimated value and twice the highest among all quality dimensions. This means that the highest quality product can be worth anywhere between \$4,133 and \$44,984 per year in premiums. Otherwise, the worst-case scenario often derives zero utility from the quality dimension with the highest investment, which is unreasonably harsh. This only affects welfare magnitudes, not design choices.

ysis coordinated consumers by changing the assignment of scores to products and consumers' beliefs about the quality represented by scores. This exercise eliminates the second channel, showing that scores can be effective even if consumers are unaware of design changes. Second, it provides a rationale for how the baseline design might have been formulated. This relies on assumptions rejected by the data and a severe aversion to misrepresenting consumers' preferences. Therefore, the welfare value of the proposed alternatives stands. Finally, it provides an alternative solution for the cautious regulator (or reader) unnerved by the assumption of informed choice.

REFERENCES

ABALUCK, JASON AND JONATHAN GRUBER (2011): "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program," *American Economic Review*, 101 (4), 1180–1210. [28]

BAKER, ANDREW, DAVID F. LARCKER, AND CHARLES C. Y. WANG (2021): "How Much Should We Trust Staggered Difference-In-Differences Estimates?" *SSRN Electronic Journal*, (March). [17]

BROWN, JASON, MARK DUGGAN, ILYANA KUZIEMKO, AND WILLIAM WOOLSTON (2014): "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program," *American Economic Review*, 104 (10), 3335–3364. [29]

CALLAWAY, BRANTLY AND PEDRO H.C. SANT'ANNA (2020): "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 1–45. [17]

CENTERS FOR MEDICARE AND MEDICAID SERVICES (2016): "Trends in Part C & D Star Rating Measure Cut Points," 2018. [13]

CHEN, TIANQI AND CARLOS GUESTRIN (2016): "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, 785–794. [31]

CURTO, VILSA, LIRAN EINAV, JONATHAN LEVIN, AND JAY BHATTACHARYA (2021): "Can health insurance competition work? Evidence from medicare advantage," *Journal of Political Economy*, 129 (2), 570–606. [37, 38]

DAFNY, LEEMORE AND DAVID DRANOVE (2008): "Do report cards tell consumers anything they don't already know? The case of Medicare HMOs," *RAND Journal of Economics*, 39 (3), 790–821. [13]

DALMASSO, NICCOLÒ, TAYLOR POSPISIL, ANN B LEE, RAFAEL IZBICKI, PETER E FREEMAN, AND ALEX I MALZ (2020): "Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference," *Astronomy and Computing*, 30, 100362. [31]

ENGLANDER, FRED, THOMAS J HODSON, AND RALPH A TERREGROSSA (1996): "Economic dimensions of slip and fall injuries," *Journal of Forensic Science*, 41 (5), 733–746. [15]

FIORETTI, MICHELE AND HONGMING WANG (2021): "Performance Pay in Insurance Markets: Evidence from Medicare," *The Review of Economics and Statistics*, 1–45. [29]

GOODMAN-BACON, ANDREW (2021): "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*. [17]

HANDEL, BEN, IGAL HENDEL, AND MICHAEL D. WHINSTON (2015): "Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk," *Econometrica*, 83 (4), 1261–1313. [28]

HOROWITZ, JOEL L. (2019): "Bootstrap Methods in Econometrics," *Annual Review of Economics*, 11 (Volume 11, 2019), 193–224. [35]

HOROWITZ, JOEL L. AND MARIANTHI MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63 (1), 145–168. [31]

HOWCROFT, JENNIFER, JONATHAN KOFMAN, AND EDWARD D LEMAIRE (2013): "Review of fall risk assessment in geriatric populations using inertial sensors," *Journal of neuroengineering and rehabilitation*, 10 (1), 1–12. [15]

IZBICKI, RAFAEL AND ANN B. LEE (2017): "Converting high-dimensional regression to high-dimensional conditional density estimation," *Electronic Journal of Statistics*, 11 (2), 2800–2831. [31]

KEENAN, PATRICIA S, SHARON-LISE T NORMAND, ZHENQIU LIN, ELIZABETH E DRYE, KANCHANA R BHAT, JOSEPH S ROSS, JEREMIAH D SCHUUR, BRETT D STAUFFER, SUSANNAH M BERNHEIM, ANDREW J EPSTEIN, ET AL. (2008): "An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure," *Circulation: Cardiovascular Quality and Outcomes*, 1 (1), 29–37. [16]

MALHERBE, CÉDRIC AND NICOLAS VAYATIS (2017): "Global optimization of Lipschitz functions," *34th International Conference on Machine Learning, ICML 2017*, 5 (1972), 3592–3601. [37]

MASUD, TAHIR AND ROBERT O. MORRIS (2001): "Epidemiology of falls," *Age and Ageing*, 30, 3–7. [15]

MURPHY, KEVIN M AND ROBERT H TOPEL (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business & Economic Statistics*, 20 (1), 88–97. [5]

NEWHOUSE, JOSEPH P, MARY PRICE, J MICHAEL MCWILLIAMS, JOHN HSU, AND THOMAS G MCGUIRE (2015): "How much favorable selection is left in Medicare Advantage?" *American journal of health economics*, 1 (1), 1–26. [38]

SCHENNACH, SUSANNE M (2016): "Recent advances in the measurement error literature," *Annual Review of Economics*, 8 (1), 341–377. [30, 31]

SPENCE, A MICHAEL (1975): "Monopoly, Quality, and Regulation," *The Bell Journal Of Economics*, 6 (2), 417–429. [21]

STEVENS, JUDY A, PHAEDRA S CORSO, ERIC A FINKELSTEIN, AND TED R MILLER (2006): "The costs of fatal and non-fatal falls among older adults," *Injury prevention*, 12 (5), 290–295. [15]