Residual Confounding in Health Plan Performance Assessments: Evidence From Randomization in Medicaid

Jacob Wallace, PhD; J. Michael McWilliams, MD, PhD; Anthony Lollo, PhD; Janet Eaton, BS; and Chima D. Ndumele, PhD

Background: Risk adjustment is used widely in payment systems and performance assessments, but the extent to which it distinguishes plan or provider effects from confounding due to patient differences is typically unknown.

Objective: To assess the degree to which risk-adjusted measures of health plan performance adequately adjust for the variation across plans that arises because of differences in patient characteristics (residual confounding).

Design: Comparison between plan performance estimates based on enrollees who made plan choices (observational population) and estimates based on enrollees assigned to plans (randomized population).

Setting: Natural experiment in which more than two thirds of a state's Medicaid population in 1 region was randomly assigned to 1 of 5 plans.

Participants: 137 933 enrollees in 2013 to 2014, of whom 31.1% selected a plan and 68.9% were randomly assigned to 1 of the same 5 plans.

Measurements: Annual total spending (that is, payments to providers), primary care use, dental care use, and avoidable emergency department visits, all scored as plan-specific deviations from the "average" plan performance within each population.

Results: Enrollee characteristics were appreciably imbalanced across plans in the observational population, as expected, but were not in the randomized population. Annual total spending varied across plans more in the observational population (SD, $147 per enrollee) than in the randomized population (SD, $70 per enrollee) after accounting for baseline differences in the observational and randomized populations and for differences across plans. On average, a plan's spending score (its deviation from the "average" performance) in the observational population differed from its score in the randomized population by $67 per enrollee in absolute value (95% CI, $38 to $123), or 4.2% of mean spending per enrollee (P = 0.009, rejecting the null hypothesis that this difference would be expected from sampling error). The difference was reduced modestly by risk adjustment to $62 per enrollee ($P$ = 0.012). Residual confounding was similarly substantial for most other performance measures. Further adjustment for social factors did not materially change estimates.

Limitation: Potential heterogeneity in plan effects between the 2 populations.

Conclusion: Residual confounding in risk-adjusted performance assessments can be substantial and should caution policymakers against assuming that risk adjustment isolates real differences in plan performance.

Primary Funding Source: Arnold Ventures.

Ann Intern Med. 2022;175:314-324. doi:10.7326/M21-0881 **Annals.org**
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 4 January 2022.

During the past 15 years, payers in the United States have increasingly shifted accountability for health care spending and quality to health plans and providers. Commercial health plans have long borne risk, and more than half of the combined Medicare and Medicaid population is now enrolled in managed care plans that receive a prospective per-enrollee (capitation) payment to cover total spending (1, 2). It is common for private plans to pass along some or all of this risk to providers, and risk contracting between public payers and providers (for example, accountable care organizations and episode-based payments) has also grown. As part of these contracts, payers link payments to performance on spending and quality metrics.

See also:

Editorial comment. 448

Web-Only
Supplement

A critical element of such "value-based" payment arrangements is accounting for underlying differences in the patient populations served by different plans or providers when measuring performance and setting capitation rates (or spending targets). Inadequate adjustment for patient risk has several consequences. First, plans or providers with higher-risk patients may be insufficiently compensated or unfairly penalized, limiting resources for the care of some groups (3–9). Second, incomplete risk adjustment can make it profitable for plans and providers to engage in risk-selection strategies that are wasteful and can undermine quality of care (10). Third, when differences in performance reflect patient differences, as opposed to differences in actual quality or efficiency, public reporting can misinform patients (11–14).

Prior studies have shown that adjusting for additional clinical or social predictors of an outcome meaningfully alters conclusions about plan or provider performance. However, even with adjustment for additional observable predictors, residual unmeasured confounding may remain or even grow if the additional predictors are inversely correlated with unmeasured predictors (14–17). The extent of unmeasured confounding is unknown because true differences in performance are not typically observed.

This study leverages a natural experiment in which the Louisiana Medicaid program randomly assigned a subset of enrollees to 1 of 5 managed care plans, whereas others chose a plan from the same set of options. To characterize the effectiveness of standard risk adjustment in removing confounding from observational estimates of health plan effects on health care spending and use, we compared plan estimates derived from the population of enrollees who chose a plan (observational population) with estimates derived from the population of enrollees assigned to a plan (randomized population), before and after adjustment for patient characteristics.

METHODS

Study Design and Population

In 2012, Louisiana Medicaid shifted from traditional fee for service to a Medicaid managed care (MMC) model in which the state contracted with capitated private plans. In primary analyses, we analyzed data from 2013 to 2014 for Medicaid enrollees who resided in Louisiana's first region to implement MMC in February 2012 ($n = 137\,933$) (**Appendix Figure 1**, available at Annals.org). Those who did not select a plan within 30 days were randomly assigned to 1 of 5 plans. We omit 2012 as a transition year.

To preserve enrollee–provider relationships from the pre-MMC primary care case management program, randomization was done within primary care case management providers ($n = 166$), ensuring that enrollees were assigned to a plan that contracted with their provider (**Supplement Table 1**, available at Annals.org). Providers ranged from primary care practices to large organizations; the median number of enrollees per provider was 1318 (interquartile range, 567 to 2659), and most providers participated in all plans (**Supplement Table 2**, available at Annals.org). Because assignment probabilities to plans varied across providers, it was important to control for the provider in our analysis (18–21). The **Supplement** (available at Annals.org) provides additional details on the state's auto-assignment process.

The objective of our analysis was to compare a plan's risk-adjusted performance (relative to the "average" plan performance) for patients who select a plan on their own (observational population) with its performance for those randomly assigned to the plan (the gold standard). The estimated difference will reflect the degree to which current methods of risk adjustment for comparisons across health plans fail to adequately adjust for variation across plans in patient characteristics that affect outcomes of interest (that is, residual confounding by patient case mix).

We expected plans to have effects on spending (or use), even for patients of the same primary care provider, because of differences in specialist and hospital networks; drug formularies; or utilization management tools, such as claim denials and prior authorization (**Table 1**). For example, 2 of the 5 plans used the traditional fee-for-service specialist network (rather than their own) and had weaker incentives to restrict use (partial rather than full-risk contracts with the state). We also expected the characteristics of enrollees who chose a plan to differ across plans because choices reflect their preferences and health care needs.

Estimating residual confounding required us to address 3 challenges. First, to separate plan effects from provider effects, we controlled for enrollees' prior provider (within which enrollees were randomized) (22, 23) and implemented additional approaches to estimate and properly account for within-provider plan differences that would have been observed had the plan assignment probabilities been equal across providers. Second, to separate the quantity of interest (residual confounding due to incomplete adjustment for enrollee characteristics) from heterogeneous effects of plans on different types of enrollees (because those who chose a plan differed from those who were assigned a plan), we weighted the randomized population to more closely resemble the observational population. The weighting better approximates the ideal experiment in which enrollees would be randomized to being either randomly assigned or left to choose a plan. Third, to address bias toward the null from plan switching in the randomized population after plan assignment (akin to crossover between treatment groups in a randomized controlled trial), we used instrumental variables (IV) methods to rescale the intention-to-treat estimates according to rates of adherence to plan assignment (24)—that is, to arrive at the plan's average treatment effect for the enrollees that were assigned to that plan and adhered to their assignment (that is, stayed in the plan).

For our primary study population and analysis, we excluded enrollees in Medicaid eligibility categories exempt from the shift to MMC (for example, dual-eligible beneficiaries and persons with disabilities), auto-assigned enrollees whose family members chose a plan (because those assignments were not random), and enrollees not continuously enrolled in Medicaid during the study period (**Appendix Figure 1**).

Study Variables

Plan Exposure

From Medicaid administrative data, we determined each enrollee's plan in each study year and whether it was assigned or chosen. For enrollees assigned to a plan, we used initial plan assignments in intention-to-treat and IV analyses because as-treated (that is, per protocol) analyses would introduce selection bias because of the 15.1% of randomly assigned enrollees who switched plans (25).

Primary and Secondary Outcomes

Total (per-enrollee annual) health care spending was prespecified as the primary study outcome because it relates to capitation payments to plans (or analogous spending benchmarks for providers in population-based payment models) (26, 27). We truncated spending at the 99.9th percentile to reduce the influence of outliers. For sensitivity analyses, we price-standardized spending to remove variation due to any differences in provider prices (**Supplement**). We found no evidence of capitated payments to providers in the form of zero or missing payment amounts, suggesting capitated plans paid providers on a fee-for-service basis during our study period.

Table 1. Differences in Enrollee Characteristics Between Plans in the Randomized and Observational Populations

Characteristic	All Plans	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	SD of Plan Means
Plan characteristics*							
Financial risk	–	Full risk	Full risk	Full risk	Shared savings	Shared savings	–
Plan selectively contracts with specialty network	–	Yes	Yes	Yes	No	No	–
Primary care providers in plan, *n*	–	1497	1530	1573	471	930	–
Enrollee characteristics							
Randomized population, *n*	–	18 384	18 871	18 358	20 507	18 852	–
Age, %							
≤5 y	30.6	31.5	30.5	30.4	30.3	30.5	0.51
6–17 y	63.5	62.5	63.6	63.6	64.1	63.4	0.61
18–64 y	5.9	6.0	5.9	6.0	5.6	6.1	0.17
Female, %	52.9	52.6	53.3	52.9	52.5	53.3	0.41
Clinical risk group, %†							
Nonusers	9.3	9.0	9.4	9.1	9.4	9.6	0.24
No acute or chronic diagnoses	63.4	63.5	63.3	63.3	63.5	63.2	0.13
Acute disease	9.6	9.9	9.4	9.5	9.6	9.6	0.18
Single chronic disease	15.6	15.5	15.8	15.8	15.7	15.5	0.16
Multiple chronic diseases	2.1	2.2	2.1	2.3	1.9	2.1	0.15
Predicted spending based on enrollee characteristics, $‡	1623	1638	1628	1637	1593	1624	20
Observational population, *n*	–	4472	8815	2838	12 041	14 795	–
Age, %							
≤5 y	35.9	33.3	37.8	34.9	36.7	35.1	1.76
6–17 y	59.6	61.5	58.3	54.7	59.4	60.8	2.69
18–64 y	4.5	5.2	3.9	10.4	3.9	4.1	2.79
Female, %	52.5	53.5	52.9	55.2	51.7	52.2	1.39
Clinical risk group, %†							
Nonusers	4.1	4.3	4.2	4.7	3.6	4.3	0.38
No acute or chronic diagnoses	60.5	59.9	61.5	62.0	59.3	60.9	1.10
Acute disease	12.6	12.1	12.4	11.3	14.2	12.0	1.07
Single chronic disease	20.1	21.0	19.5	18.6	20.4	20.1	0.93
Multiple chronic diseases	2.6	2.7	2.4	3.5	2.5	2.7	0.41
Predicted spending based on enrollee characteristics, $	1810	1845	1770	1831	1802	1826	50

* In our setting, 2 of the 5 plans used the traditional fee-for-service specialist network (rather than selectively contract with their own providers) and had weaker incentives to restrict use due to shared savings rather than full-risk contracts with the state.

† Clinical risk group (i.e., health status) is calculated using the 3M Clinical Risk Groups model with claims from February 2011 through January 2012. Enrollees who did not have any claims during this period were assigned the status of "nonusers." Clinical risk groups were aggregated into the categories above to summarize. Risk adjustment for our primary analyses used clinical risk groups derived from the claims in prior (rather than baseline) years (Supplement, available at Annals.org).

‡ The predicted spending means reflect the weighting of the randomized population to balance its distribution of characteristics with that seen in the observational population (Supplement). The weighted means do not match exactly because the weighting did not reflect all 170 risk indicators. The unweighted mean for the randomized population was $1502 (Appendix Table 1, available at Annals.org). All other estimates in Table 1 are unweighted.

As secondary outcomes, we analyzed the following 4 measures of annual use: completion of an age-specific well-child visit, use of 1 or more primary care office visits, use of any dental care, and number of avoidable emergency department visits (28, 29). We selected these measures because they could be reliably constructed from administrative claims data and were applicable to the entire study population of children and adults or its pediatric majority. The first 3 measures were selected from Healthcare Effectiveness Data and Information Set measures used to assess MMC plan performance (see Supplement Methods and Supplement Figures 2 and 3 for additional details on measures, available at Annals.org).

Enrollee Characteristics

From administrative data, we obtained enrollee age, sex, and Medicaid eligibility category. For each study year, we used claims from the prior year to categorize enrollees' clinical risk, a common approach to risk adjustment. We used 3M Clinical Risk Grouping software, developed for use in Medicaid (30, 31), to construct 170 mutually exclusive clinical risk groups on the basis of diagnosis codes. As a summary measure of enrollee risk, we also report predicted spending as a function of these observed characteristics. In a sensitivity analysis, we used claims from the preassignment period, rather than the prior year, to assess clinical risk, thereby removing any bias introduced by the effect of MMC plans on claims, diagnosis, or coding (32).

Statistical Analysis

First, we examined the distribution of observed enrollee characteristics across plans to assess balance in the randomized and observational populations. Second, we used linear regression to estimate plan effects on outcomes separately in the randomized and observational populations. To estimate plan effects in the observational population with and without risk adjustment, we fit a model of each annual outcome as a function of the plan chosen by the enrollee that year, with and without adding the enrollee characteristics as covariates (age, sex, eligibility category, and clinical risk groups). In the randomized population, we estimated an analogous model but with use of IV to address nonadherence to plan assignment (plan switching) to arrive at the estimate of the plan's average treatment effect for the enrollees who were assigned to that plan and adhered to their assignment (that is, stayed in the plan) (24).

From these models, for each outcome in both of the populations, we calculated a plan score for each plan equal to that plan's deviation from average plan performance in the population (that is, how much better or worse a plan did compared with "average" performance). We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.

Third, we summarized within-plan differences in scores as derived from the observational versus randomized populations—that is, how much observational, risk-adjusted estimates of plan scores differed from gold-standard estimates —in several ways. First, we calculated the absolute value of the score difference for each plan as derived from the 2 populations and averaged those differences across the 5 plans. We prespecified this "mean absolute difference" in plan scores as our primary summary measure of confounding. Second, we also reported the absolute value of the largest within-plan difference to describe the plan most affected by confounding. Third, we took the square root of the mean squared score difference between the observational and randomized populations (to place greater weight on outlier score differences and account for negative differences). Fourth, we estimated the linear relationship between the plan scores in the 2 populations, expressed as the change in the randomized score associated with a 1-unit increase in the observational score (33).

To test whether score differences were greater than expected from sampling error, we resampled from plans' randomly assigned enrollees to create a distribution of score differences expected from sampling error alone and obtained a 2-sided *P* value. To control the false discovery rate within families of independent hypotheses, we used the Benjamini–Hochberg procedure to adjust *P* values (**Supplement**).

In sensitivity analyses, we adjusted for additional enrollee characteristics (race, ethnicity, and ZIP code of residence) and applied various weighting approaches to equalize proportions of enrollees assigned to each plan within each baseline provider and to address potential treatment effect heterogeneity (that is, differing plan effects on the 2 populations owing to differences in the populations overall and within each plan).

The study was approved by the Yale University Institutional Review Board. Analyses were done using Python, version 3.8.7 (Python Software Foundation), and Stata, version 14 (StataCorp). The **Supplement** provides more detail on the analytic approaches, including annotated code.

Role of the Funding Source

The funders had no role in the study design, data collection and analysis, preparation of the manuscript, or decision to publish.

RESULTS

Study Population

After inclusion criteria (**Appendix Figure 1**), the study population included 137 933 enrollees, of whom 94 972 (68.9%) were randomly assigned and 42 961 (31.1%) chose a plan. The observational and randomized populations did not differ markedly in their demographic or clinical characteristics (**Table 1**; **Appendix Table 1**, available at Annals.org; and **Supplement Table 4**, available at Annals.org) and were similarly distributed across primary care providers of record at baseline (**Supplement Figure 4**, available at Annals.org). Adults accounted for 9.3% of the enrollee-year observations and 16.0% of total spending. Among enrollees who were randomly assigned, 84.9% remained in their assigned plans in 2013 to 2014 (**Appendix Figure 2**, available at Annals.org), suggesting plan assignment was a strong instrument for plan exposure ($P < 0.001$) (34).

Differences in Mean Enrollee Characteristics and Outcomes Among Plans

Enrollees' age, sex, clinical risk, and predicted spending differed appreciably across plans in the observational population but not in the randomized population (**Table 1**). For example, plan 3 attracted older enrollees, and plan 4 attracted more enrollees with a history of acute disease than other plans in the observational population.

In the observational population, plan scores for total spending per enrollee and rates of primary care, dental care, and avoidable emergency department use differed meaningfully across plans in unadjusted analyses (**Table 2**; **Supplement Table 5**, available at Annals.org). For example, unadjusted plan spending scores in the observational population ranged from −$126 for plan 2 (that is, spending in plan 2 was $126 per enrollee lower than the average across all plans) to $172 for plan 5. By comparison, IV estimates (based on the randomized population) of plan effects on these outcomes varied less, ranging from −$56 per enrollee for the lowest-spending plan (plan 1) to $79 for the highest-spending plan (plan 5)—suggesting that MMC plans do exert influence on health care spending but to a lesser degree than suggested by estimates derived from the observational population. The difference between plan spending scores (as measured by the SD of the plan scores squared) was 4 times greater in the observational population than in the randomized population; this difference was not due to the smaller observational sample (**Supplement Figures 6** and **7**, available at Annals.org) and persisted after risk adjustment (**Appendix Table 2**, available at Annals.org).

Residual Confounding After Risk Adjustment

On average, the unadjusted annual spending scores derived from the observational population differed from those derived from the randomized population by $67 per enrollee in absolute value ($P = 0.009$, rejecting the null hypothesis that this difference would be expected from sampling error alone) (**Table 3**). To put this average amount of confounding in perspective, $67 per enrollee was 4.2% of mean spending in the study population and is similar in

Table 2. Differences in Enrollee Outcomes Between Plans in the Randomized and Observational Populations, 2013 to 2014

Outcome	All Plan Mean	Plan Performance Scores (Plan Deviations From Average Plan Performance in the Population)*					SD of Plan Scores	*P* Value for Test of Difference in Plan Scores†
		Plan 1	Plan 2	Plan 3	Plan 4	Plan 5		
Score* derived from randomized population‡								
Total health care spending, *$*	1612	−56 (−95 to −18)	−54 (−92 to −17)	−43 (−93 to 8)	75 (34 to 115)	79 (45 to 113)	70	<0.001 (<0.001)
HEDIS annual use of primary and dental care, *percentage points*§								
Age-specific well-child visit	51.40	−2.22 (−3.86 to −0.58)	1.37 (0.17 to 2.57)	0.72 (−0.51 to 1.94)	1.61 (−0.67 to 3.90)	−1.48 (−2.64 to −0.32)	1.74	<0.001 (<0.001)
≥1 primary care office visit (children and adults)	80.07	1.22 (−1.05 to 3.48)	−2.75 (−4.61 to −0.89)	0.06 (−1.56 to 1.68)	0.60 (−0.27 to 1.47)	0.88 (−0.02 to 1.78)	1.60	0.004 (0.005)
Any dental care (children and adults)	60.30	1.21 (0.10 to 2.31)	−0.37 (−1.25 to 0.51)	−0.18 (−1.44 to 1.08)	−0.33 (−1.50 to 0.84)	−0.32 (−1.42 to 0.77)	0.68	0.32 (0.32)
Annual avoidable emergency department visits per 100 enrollees, *n*‖	11.06	1.23 (0.48 to 1.97)	−0.27 (−0.84 to 0.30)	1.14 (0.39 to 1.89)	−1.32 (−1.99 to −0.66)	−0.78 (−1.31 to −0.25)	1.14	<0.001 (<0.001)
Score* derived from observational population without risk adjustment								
Total health care spending, *$*	1835	−95 (−183 to −7)	−126 (−190 to −61)	−100 (−238 to 37)	148 (82 to 214)	172 (100 to 245)	147	<0.001 (<0.001)
HEDIS annual use of primary and dental care, *percentage points*§								
Age specific well-child visit	61.38	−5.30 (−7.31 to −3.30)	2.79 (1.15 to 4.44)	−0.37 (−2.82 to 2.08)	4.76 (1.95 to 7.57)	−1.89 (−3.00 to −0.77)	3.95	<0.001 (<0.001)
≥1 primary care office visit (children and adults)	88.78	1.09 (−0.66 to 2.83)	−2.51 (−5.08 to 0.05)	−1.54 (−3.03 to −0.04)	1.21 (0.17 to 2.25)	1.76 (0.69 to 2.82)	1.90	0.023 (0.023)
Any dental care (children and adults)	72.64	−0.44 (−2.13 to 1.25)	−0.04 (−1.38 to 1.30)	−2.92 (−4.94 to −0.90)	2.97 (1.80 to 4.14)	0.43 (−0.49 to 1.35)	2.10	<0.001 (<0.001)
Annual avoidable emergency department visits per 100 enrollees, *n*‖	9.92	−0.29 (−1.31 to 0.73)	0.48 (−0.43 to 1.39)	3.03 (1.45 to 4.62)	−1.87 (−2.58 to −1.16)	−1.36 (−2.12 to −0.59)	1.93	<0.001 (<0.001)

HEDIS = Healthcare Effectiveness Data and Information Set.

* We calculated a plan score for each plan equal to the plan's deviation from the average plan performance in the population (i.e., how much better or worse a plan performed compared with "average" performance). We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.

† *P* values in parentheses are adjusted for multiple inference using the Benjamini–Hochberg procedure (Supplement, available at Annals.org).

‡ We weighted the randomized population to balance its distribution of enrollee characteristics with the observational population. For precision, the analyses in the randomly assigned population are adjusted for age; sex; Medicaid eligibility category; and 170 clinical risk groups, defined using the 3M Clinical Risk Grouping software (Supplement).

§ Based on HEDIS specifications for administrative claims measures of primary care and dental care use.

‖ Avoidable emergency department visit measure based on the Statewide Collaborative Quality Improvement Project of the California Department of Health Care Services (34).

magnitude to the largest effect of a plan on spending (relative to the average plan performance) noted earlier. The largest absolute score difference between populations for a single plan was $94 per enrollee, or 5.8% of mean spending (**Table 3**); this was for plan 5 (score: $172 vs. $79 in the observational vs. randomized population, respectively) (**Table 2**). The spread between the most positive and negative score differences was $170 per enrollee, or 10.5% of overall mean spending (**Figure 1**).

Risk adjustment altered plan spending scores derived from the observational population somewhat but moved them statistically significantly closer to the scores derived from the randomized population for only 2 of 5 plans (**Figure 2**; **Appendix Figure 3**, available at Annals.org). On average, risk adjustment reduced the mean absolute difference between the observational and randomized scores from $67 per enrollee ($P = 0.009$) to $62 per enrollee ($P = 0.012$), which was 3.8% of mean spending (**Table 3**). This reduction in confounding was small and not statistically significant even though the risk-adjustment model explained a substantially larger proportion of the variance in enrollee-level spending, with an increase in model R^2 from 2% to 29% (35). Further adjustment for enrollee race, ethnicity, and ZIP code of residence increased enrollee-level R^2 slightly and did not meaningfully reduce confounding (**Supplement Tables 7** to **10**, available at Annals.org).

Results for use of primary care, dental care, and avoidable emergency department visits were qualitatively similar to those for spending (**Table 2** and **Table 3**; **Figure 1**; and **Supplement Figures 9** and **10** and **Supplement Table 11**, available at Annals.org). Risk adjustment reduced the mean absolute difference between scores derived from the 2 populations by 12.5% to 36.2% (**Table 3**). For the measure of child and adult use of primary care office visits, residual confounding after risk adjustment was small and no longer statistically distinguishable from differences expected from sampling error. The average residual confounding in plan scores for avoidable emergency department visits also was no longer statistically significant after risk adjustment but remained substantial (7.1% of the study population mean).

Plan rankings in spending scores were roughly consistent in the randomized and observational populations because higher-cost enrollees selected plans that allowed for higher spending (**Figure 1**) (36). The consistency of rankings in the randomized and observational populations varied across measures (**Figure 2**).

For spending, a $1 higher observational score was only associated on average with a $0.47 higher randomized score. This relationship was minimally altered by risk adjustment (**Supplement Figures 9**, **11**, and **12**, available at Annals.org; **Supplement Table 11**).

Additional Analyses

Intention-to-treat and IV estimates of plan scores were similar, as expected from the limited plan switching (**Supplement Figure 13**, available at Annals.org). Our results were also robust to price standardization (for spending) (**Supplement Figure 14**, available at Annals.org), alternate weighting approaches, and alternative risk-adjustment methods (**Supplement Tables 8** to **10** and **12** and **Supplement Figures 15** and **16**, available at Annals.org). Additional sensitivities supported our main conclusions (**Supplement Tables 13** to **16**, available at Annals.org).

Discussion

In this study of a state's Medicaid plans, estimated plan performance on spending and use measures based on a population of enrollees that chose plans differed substantially from estimates of plan performance based on a population randomly assigned to plans. Differences in performance estimated on the basis of the observational population were only modestly reduced by risk adjustment, suggesting considerable residual confounding from variation in unmeasured enrollee characteristics. We estimated that the $62 per enrollee average residual confounding after risk adjustment represented 3.8% of mean spending—a large difference relative to health plan margins in Medicaid (37)—and approximated the largest "true" plan effects on spending, as seen in the randomized population. Despite this residual confounding, plan rankings based on spending estimated from the observational population were largely consistent with those derived from the randomized population.

Our results are consistent with prior research showing that the addition of otherwise omitted patient variables to risk-adjustment models would alter payments or performance scores for plans or providers (3, 4, 9, 15, 17, 38–46). Likewise, we show that enrollees differ in observable ways across health plans, that enrollee characteristics are strongly predictive of spending, and that adjustment for observed characteristics alters assessments of plan performance. However, our study goes beyond prior work to quantify the extent to which adjustment of observational estimates better approximates true plan effects on care and the amount of confounding that may remain. The limited effect of adjustment on confounding is consistent with our finding that plan variation in observed

Table 3. Differences in Plan Performance Scores Between the Randomized and Observational Populations, 2013 to 2014

Outcome	Strength of Patient-Level Prediction With Risk Adjustment, R^2*	Mean Absolute Difference in Plan Scores Derived From the Randomized and Observational Populations†				Largest Absolute Difference in Plan Scores	
		Difference Without Risk Adjustment (95% CI)‡	*P* Value§	Difference With Risk Adjustment (95% CI)‡	*P* Value§	Difference Without Risk Adjustment (95% CI)‡	Difference With Risk Adjustment (95% CI)‡
Total health care spending, *$*	0.29	67 (38–123)	0.009 (0.011)	62 (39–106)	0.012 (0.020)	94 (67–215)	106 (64–212)
HEDIS annual use of primary and dental care, *percentage points*							
Age-specific well-child visit‖	0.10	1.83 (1.07–3.00)	<0.001 (<0.001)	1.45¶ (0.76–2.61)	0.005 (0.013)	3.15 (2.07–5.71)	2.52¶ (1.47–4.88)
≥1 primary care office visit (children and adults)‖	0.19	0.69 (0.41–1.59)	0.051 (0.051)	0.44 (0.27–1.29)	0.38 (0.38)	1.60 (0.74–3.39)	0.98 (0.48–2.66)
Any dental care (children and adults)‖	0.03	1.76 (1.00–2.94)	<0.001 (<0.001)	1.54¶ (0.86–2.67)	0.003 (0.013)	3.30 (1.94–5.55)	3.00¶ (1.72–5.07)
Annual avoidable emergency department visits per 100 enrollees, *n*	0.06	1.06 (0.52–1.67)	0.005 (0.008)	0.74 (0.39–1.35)	0.080 (0.100)	1.89 (0.95–3.58)	1.29 (0.76–2.81)

HEDIS = Healthcare Effectiveness Data and Information Set.

* The adjusters in the risk-adjustment model include age; sex; Medicaid eligibility category; and 170 clinical risk groups, defined using the 3M Clinical Risk Grouping software. The R^2 represents the share of the variation in the outcome for the observational population that is explained by the risk-adjustment model.

† We calculated a plan score for each plan equal to the plan's deviation from the average plan performance in the population (i.e., how much better or worse a plan performed compared with "average" performance). We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.

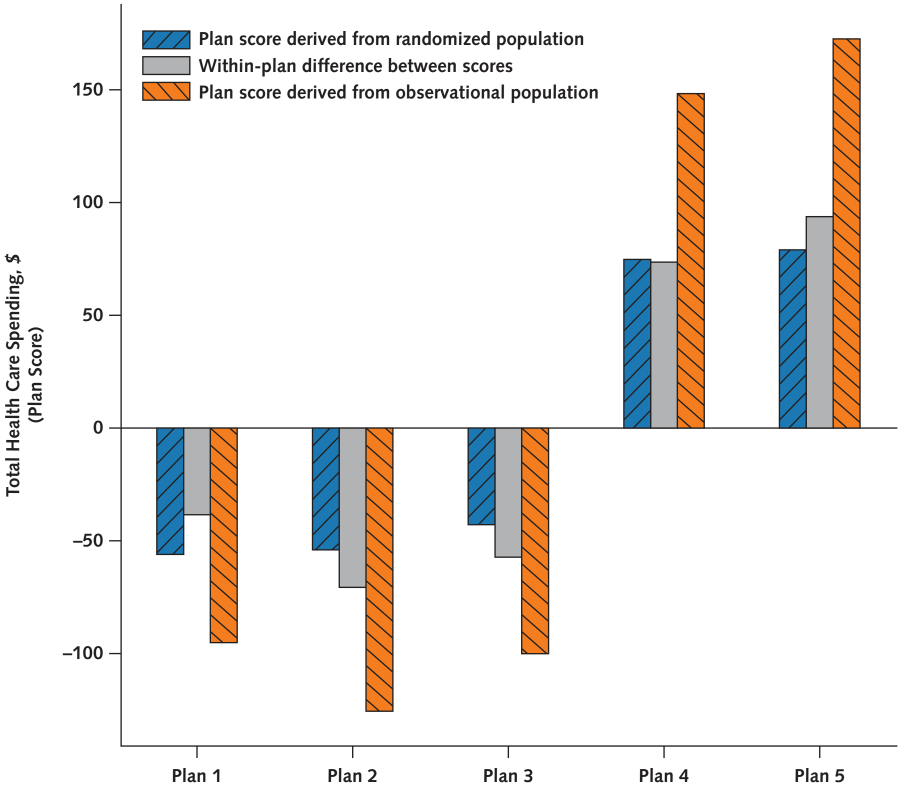
‡ The 95% CIs are constructed based on a clustered bootstrap procedure (Supplement, available at Annals.org).

§ *P* values are from a test of how extreme the observed difference in plan scores between randomized and observational populations is relative to their expected differences due to sampling error; the null hypothesis is that the observed difference is only due to sampling error (Supplement Figure 10, available at Annals.org). *P* values in parentheses are adjusted for multiple inference using the Benjamini–Hochberg procedure (Supplement).

‖ Based on HEDIS specifications for administrative claims measures of primary care and dental care use.

¶ Risk-adjusted difference differs statistically significantly from difference without risk adjustment ($P < 0.050$) based on a clustered bootstrap procedure (Supplement).

Figure 1. Differences in plan total health care spending scores derived from the observational and randomized populations.



Each bar corresponds to 1 of the 5 plans. The blue area of the bar corresponds to a plan's randomized spending score (relative to the "average" plan mean) based on the randomly assigned population. The orange bar corresponds to a plan's spending score based on the observational population before risk adjustment. The grey unhatched portion indicates the difference between the 2 scores, or the extent of residual confounding in the observational scores. For these 5 Medicaid plans, higher-cost enrollees selected plans that control spending to a lesser extent. We calculated a plan score for each plan equal to the plan's deviation from the population-specific plan mean. We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.

enrollee characteristics caused by nonrandom sorting in the observational population was weakly related to the variation in unobserved predictors of spending.

Our results should caution policymakers against assuming that current risk-adjustment models adequately isolate effects directly attributable to plans (or providers). Additional strategies to promote the goals of risk adjustment merit consideration. First, risk adjustment may be improved by adding predictors or using advanced statistical techniques (for example, matching or machine learning) (47). Improving prediction, however, can weaken incentives to reduce use or improve care. For example, adjustment for a history of stroke increases the costs to plans or providers of preventing strokes (they are penalized with lower payments). In addition, predictors that can be manipulated (for example, diagnoses) create incentives for wasteful practices such as upcoding (32, 48, 49). Although further adjustment for social risk factors has been shown to affect performance assessments (9, 14, 17, 41, 50, 51), our results suggest substantial confounding may remain. As efforts to improve risk adjustment proceed, our findings underscore the pitfalls of focusing on patient-level prediction (52, 53). Despite a high patient-level R^2 of 29% for health care spending, indicating that the enrollee variables included in our risk-adjustment approach captured more than a quarter of the variation in the outcome, risk adjustment did not meaningfully reduce confounding at the plan level for spending in our study.

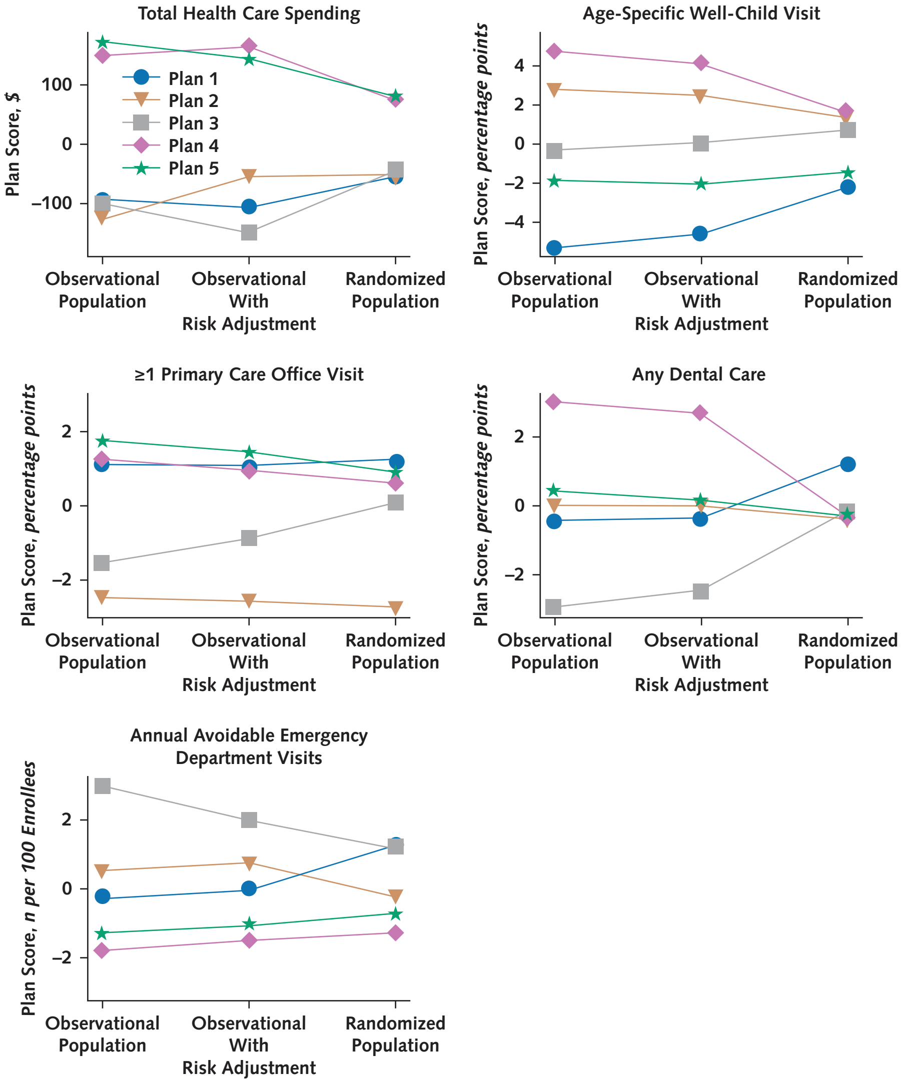
Second, rather than rely entirely on predictive models to correct under- or overpredictions (54), reinsurance mechanisms—that is, policies that limit an organization's liability for individual enrollees by covering costs that exceed a certain threshold—can be used to improve model fit and mitigate incentives to avoid high-risk patients or attract favorable risks (48, 55–57). This approach could also foster a shift in focus from narrow high-risk case management to systemic changes in care delivery (58–61).

Third, public payers could reconsider pay-for-performance strategies that penalize providers who neither perform well nor improve to maintain budget neutrality. Instead, pay-for-improvement programs that offer rewards for improvement, perhaps targeted to poor performers at baseline (those with poor quality or unobservably high-risk patients), could direct resources for improvement where they are needed most and obviate the need for risk adjustment.

Fourth, our analysis focused on how completely risk adjustment predicted differences in plan spending due to differences in plan enrollees, but prediction should not be the only goal of risk adjustment. For example, risk adjustment that accurately predicts lower spending for underserved groups would result in lower risk-adjusted prospective payments for those groups, thereby entrenching underspending for patients with unmet needs (62). Instead, it may be socially desirable to purposefully adjust payments to plans or providers for underserved groups to levels above current spending for those groups; such a system would be less predictive but more equitable (63, 64).

This study has limitations. First, it is based on a single Medicaid program, a predominantly pediatric population, and a limited set of spending and use measures. Therefore, the findings may not generalize to all Medicaid programs or other settings where risk adjustment is used (for example, Medicare Advantage). Second, health plan scores based on observational and randomized populations may differ because of heterogeneity in plan effects on different populations; hence, we may be under- or overstating problems with risk adjustment. However, weighting to balance characteristics between randomized and observational populations did not meaningfully affect our estimates (65). Third, although commonly used (66), IV methods rely on assumptions that cannot be fully verified—for example, the plan effect

Figure 2. Plan performance scores in the observational versus randomized populations.



These figures plot plan scores based on the observational and randomly assigned populations for each outcome. For each outcome, we plot the unadjusted observational plan score, the risk-adjusted observational plan score, and the plan score in the randomized population. Each set of connected points corresponds to 1 of the plans in our sample. We calculated a plan score for each plan equal to the plan's deviation from the population-specific plan mean. We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.

on enrollees who comply with assignment are similar to those who do not (24). Fourth, we rely on linear models to measure plan performance—including for binary or skewed outcomes—a common approach when using IV (67) and broadly consistent with how risk adjustment is implemented in practice (32). However, potential model misspecification may lead us to misstate the extent of residual confounding.

In conclusion, we found evidence of substantial confounding in observational estimates of Medicaid plan performance that was reduced only modestly by standard risk adjustment methods. Our results should caution policymakers against presuming that current risk-adjustment approaches achieve the goal of isolating plan effects on care and encourage the development of improved methods and additional policies to improve payment system performance.

From Yale School of Public Health, New Haven, Connecticut (J.W., A.L., C.D.N.); Department of Health Care Policy, Harvard Medical School, and Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts (J.M.M.); and Yale School of Public Health, and Tobin Center for Economic Policy, Yale University, New Haven, Connecticut (J.E.).

Disclaimer: The content is solely the responsibility of the authors and does not necessarily reflect the official views of the Tobin Center for Economic Policy or Arnold Ventures.

Financial Support: By the Tobin Center for Economic Policy and a grant from Arnold Ventures.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M21-0881.

Reproducible Research Statement: *Study protocol and data set:* Not available. *Statistical code:* See the **Supplement** (available at Annals.org).

Corresponding Author: Jacob Wallace, PhD, Department of Health Policy and Management, Yale School of Public Health, 60 College Street, New Haven, CT 06510; e-mail, jacob.wallace@yale.edu.

Author contributions are available at Annals.org.

References

1. **Kaiser Family Foundation.** State health facts (2017). Data source: KFF analysis of the Centers for Medicare and Medicaid Services' Medicaid Managed Care Enrollment Reports, 2019. Accessed atwww.kff.org/ecdc6ad/ on 25 May 2020.
2. **Jacobson G, Freed M, Damico A, et al.** A dozen facts about Medicare Advantage in 2019. Accessed at http://files.kff.org/attachment/Data-Note-A-Dozen-Facts-About-Medicare-Advantage-in-2019 on 26 May 2021.
3. **Joynt KE, Jha AK.** A path forward on Medicare readmissions. N Engl J Med. 2013;368:1175-7. [PMID: 23465069] doi:10.1056/NEJMp1300122
4. **Barnett ML, Hsu J, McWilliams JM.** Patient characteristics and differences in hospital readmission rates. JAMA Intern Med. 2015;175:1803-12. [PMID: 26368317] doi:10.1001/jamainternmed.2015.4660
5. **DuGoff E, Bishop S, Rawal P.** Hospital readmission reduction program reignites debate over risk adjusting quality measures. Accessed at https://www.healthaffairs.org/do/10.1377/hblog20140814.040725/full/ on 1 December 2021.
6. **Gu Q, Koenig L, Faerberg J, et al.** The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. Health Serv Res. 2014;49:818-37. [PMID: 24417309] doi:10.1111/1475-6773.12150
7. **Ryan AM.** Will value-based purchasing increase disparities in care. N Engl J Med. 2013;369:2472-4. [PMID: 24369072] doi:10.1056/NEJMp1312654
8. **Lipstein SH, Dunagan WC.** The risks of not adjusting performance measures for sociodemographic factors. Ann Intern Med. 2014;161:594-6. [PMID: 25048401] doi:10.7326/M14-1601
9. **Chen LM, Epstein AM, Orav EJ, et al.** Association of practice-level social and medical risk with performance in the Medicare physician value-based payment modifier program. JAMA. 2017;318:453-461. [PMID: 28763549] doi:10.1001/jama.2017.9643
10. **Luft HS, Miller RH.** Patient selection in a competitive health care system. Health Aff (Millwood). 1988;7:97-119. [PMID: 3145918]
11. **Buntin MB, Ayanian JZ.** Social risk factors and equity in Medicare payment. N Engl J Med. 2017;376:507-510. [PMID: 28177864] doi:10.1056/NEJMp1700081
12. **National Academies of Sciences, Engineering, and Medicine.** Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. National Academies Pr; 2016.
13. **Jha AK, Zaslavsky AM.** Quality reporting that addresses disparities in health care. JAMA. 2014;312:225-6. [PMID: 25027134] doi:10.1001/jama.2014.7204
14. **Roberts ET, Zaslavsky AM, Barnett ML, et al.** Assessment of the effect of adjustment for patient characteristics on hospital readmission rates: implications for pay for performance. JAMA Intern Med. 2018;178:1498-1507. [PMID: 30242346] doi:10.1001/jamainternmed.2018.4481
15. **Johnston KJ, Wen H, Hockenberry JM, et al.** Association between patient cognitive and functional status and Medicare total annual cost of care: implications for value-based payment. JAMA Intern Med. 2018;178:1489-1497. [PMID: 30242381] doi:10.1001/jamainternmed.2018.4143
16. **Rose S, Zaslavsky AM, McWilliams JM.** Variation in accountable care organization spending and sensitivity to risk adjustment: implications for benchmarking. Health Aff (Millwood). 2016;35:440-8. [PMID: 26953298] doi:10.1377/hlthaff.2015.1026
17. **Roberts ET, Zaslavsky AM, McWilliams JM.** The value-based payment modifier: program outcomes and implications for disparities. Ann Intern Med. 2018;168:255-265. [PMID: 29181511] doi:10.7326/M17-1740
18. **Finkelstein A, Ji Y, Mahoney N, et al.** Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: interim analysis of the first year of a 5-year randomized trial. JAMA. 2018;320:892-900. [PMID: 30193277] doi:10.1001/jama.2018.12346
19. **Dumville JC, Hahn S, Miles JN, et al.** The use of unequal randomisation ratios in clinical trials: a review. Contemp Clin Trials. 2006;27:1-12. [PMID: 16236557]
20. **Chan A, Tetzlaff JM, Gøtzsche PC, et al.** SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013;346:e7586. [PMID: 23303884] doi:10.1136/bmj.e7586
21. **Barnett ML, Wilcock A, McWilliams JM, et al.** Two-year evaluation of mandatory bundled payments for joint replacement. N Engl J Med. 2019;380:252-262. [PMID: 30601709] doi:10.1056/NEJMsa1809010
22. **Krueger AB.** Experimental estimates of education production functions. Q J Econ. 1999;114:497-532.
23. **Manning WG, Newhouse JP, Duan N, et al.** Health insurance and the demand for medical care: evidence from a randomized experiment. Am Econ Rev. 1987;77:251-77. [PMID: 10284091]
24. **Sussman JB, Hayward RA.** An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised

controlled trials. BMJ. 2010;340:c2073. [PMID: 20442226] doi:10.1136/bmj.c2073

25. **Geruso M, Layton TJ, Wallace J.** What difference does a health plan make? Evidence from random plan assignment in Medicaid. NBER working paper 27762. Accessed at www.nber.org/papers/w27762.pdf on 1 December 2021.

26. **Centers for Medicare & Medicaid Services.** Hospital Readmissions Reduction Program (HRRP). Accessed at www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program on 9 July 2020.

27. **Agency for Healthcare Research and Quality.** Prevention Quality Indicators overview. Accessed at www.qualityindicators.ahrq.gov/modules/pqi_overview.aspx on 9 July 2020.

28. **Centers for Medicare & Medicaid Services.** Medicaid Adult Core Set measures. Accessed at www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-and-child-health-care-quality-measures/adult-core-set-reporting-resources/index.html on 26 May 2021.

29. **Medi-Cal Managed Care Division, California Departmen of Health Care Services.** Statewide Collaborative Quality Improvement Project: reducing avoidable emergency room visits. Final Remeasurement Report: January 1, 2010–December 31, 2010. Accessed at www.dhcs.ca.gov/dataandstats/reports/Documents/MMCD_Qual_Rpts/EQRO_QIPs/CA2011-12_QIP_Coll_ER_Remeasure_Report.pdf on 1 December 2021.

30. **Hughes JS, Averill RF, Eisenhandler J, et al.** Clinical risk groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. Med Care. 2004;42:81-90. [PMID: 14713742]

31. **Lee I, Monahan S, Serban N, et al.** Estimating the cost savings of preventive dental services delivered to Medicaid-enrolled children in six southeastern states. Health Serv Res. 2018;53:3592-3616. [PMID: 29194610] doi:10.1111/1475-6773.12811

32. **Geruso M, Layton T.** Upcoding: evidence from Medicare on squishy risk adjustment. J Polit Econ. 2020;12:984-1026. [PMID: 32719571] doi:10.1086/704756

33. **Chetty R, Friedman JN, Rockoff JE.** Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Am Econ Rev. 2014;104:2593–2632. doi:10.1257/aer.104.9.2593

34. **Angrist JD, Pischke J.** Mostly Harmless Econometrics: An Empiricist's Companion. Princeton Univ Pr; 2009.

35. **Rose S, McGuire TG.** Limitations of P-values and R-squared for stepwise regression building: a fairness demonstration in health policy risk adjustment. Am Stat. 2019;73:152-156. [PMID: 31263291] doi:10.1080/00031305.2018.1518269

36. **Cutler DM, Reber SJ.** Paying for health insurance: the trade-off between competition and adverse selection. Q J Econ. 1998;113:433-466. doi:10.1162/003355398555649

37. **Palmer JD, Pettit CT, McCulla IM.** Medicaid managed care financial results for 2017. Accessed at www.milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/2018/medicaid-managed-care-financial-results-2017.ashx on 26 May 2021.

38. **Gilman M, Adams EK, Hockenberry JM, et al.** Safety-net hospitals more likely than other hospitals to fare poorly under Medicare's value-based purchasing. Health Aff (Millwood). 2015;34:398-405. [PMID: 25732489] doi:10.1377/hlthaff.2014.1059

39. **Gilman M, Hockenberry JM, Adams EK, et al.** The financial effect of value-based purchasing and the hospital readmissions reduction program on safety-net hospitals in 2014. A cohort study. Ann Intern Med. 2015;163:427-36. [PMID: 26343790] doi:10.7326/M14-2813

40. **Joynt Maddox KE.** Financial incentives and vulnerable populations–will alternative payment models help or hurt. N Engl J Med. 2018;378:977-979. [PMID: 29539282] doi:10.1056/NEJMp1715455

41. **Joynt Maddox KE, Reidhead M, Hu J, et al.** Adjusting for social risk factors impacts performance and penalties in the hospital readmissions reduction program. Health Serv Res. 2019;54:327-336. [PMID: 30848491] doi:10.1111/1475-6773.13133

42. **Johnston KJ, Bynum JPW, Joynt Maddox KE.** The need to incorporate additional patient information into risk adjustment for medicare beneficiaries. JAMA. 2020;323:925-926. [PMID: 31999298] doi:10.1001/jama.2019.22370

43. **Werner RM, Asch DA.** The unintended consequences of publicly reporting quality information. JAMA. 2005;293:1239-44. [PMID: 15755946]

44. **Chatterjee P, Werner RM.** The hospital readmission reduction program and social risk. Health Serv Res. 2019;54:324-326. [PMID: 30848490] doi:10.1111/1475-6773.13131

45. **Werner RM, Goldman LE, Dudley RA.** Comparison of change in quality of care between safety-net and non-safety-net hospitals. JAMA. 2008;299:2180-7. [PMID: 18477785] doi:10.1001/jama.299.18.2180

46. **Joynt KE, Jha AK.** Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. 2013;309:342-3. [PMID: 23340629] doi:10.1001/jama.2012.94856

47. **Rose S.** Robust machine learning variable importance analyses of medical conditions for health care spending. Health Serv Res. 2018;53:3836-3854. [PMID: 29527659] doi:10.1111/1475-6773.12848

48. **Geruso M, McGuire TG.** Tradeoffs in the design of health plan payment systems: fit, power and balance. J Health Econ. 2016;47:1-19. [PMID: 26922122] doi:10.1016/j.jhealeco.2016.01.007

49. **Ellis RP, Martins B, Rose S.** Risk adjustment for health plan payment. In: McGuire T, Van Kleef R, eds. Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice. Elsevier; 2018:55-104.

50. **Montz E, Layton T, Busch AB, et al.** Risk-adjustment simulation: plans may have incentives to distort mental health and substance use coverage. Health Aff. 2016;35:1022–8. doi:10.1377/hlthaff.2015.1668

51. **Kind AJ, Jencks S, Brock J, et al.** Neighborhood socioeconomic disadvantage and 30-day rehospitalization. A retrospective cohort study. Ann Intern Med. 2014;161:765-74. [PMID: 25437404] doi:10.7326/M13-2946

52. **O'Malley AJ, Zaslavsky AM, Elliott MN, et al.** Case-mix adjustment of the CAHPS Hospital Survey. Health Serv Res. 2005;40:2162-81. [PMID: 16316443] doi:10.1111/j.1475-6773.2005.00470.x

53. **Wagner TH, Upadhyay A, Cowgill E, et al.** Risk adjustment tools for learning health systems: a comparison of DxCG and CMS-HCC V21. Health Serv Res. 2016;51:2002-19. [PMID: 26839976] doi:10.1111/1475-6773.12454

54. **MedPAC.** Improving risk adjustment in the Medicare program. In: MedPAC, ed. Report to the Congress: Medicare and the Health Care Delivery System. MedPAC; 2014:21–32.

55. **McWILLIAMS JM, Hatfield LA, Landon BE, et al.** Savings or selection? Initial spending reductions in the Medicare shared savings program and considerations for reform. Milbank Q. 2020;98:847-907. [PMID: 32697004] doi:10.1111/1468-0009.12468

56. **McWilliams JM, Chen AJ.** Understanding the latest ACO "savings": curb your enthusiasm and sharpen your pencils–part 2. Health Affairs Blog. 13 November 2020. Accessed at www.healthaffairs.org/do/10.1377/hblog20201106.1578/full/ on 1 December 2021. doi:10.1377/hblog20201106.1578

57. **McGuire TG, Schillo S, van Kleef RC.** Reinsurance, repayments, and risk adjustment in individual health insurance: Germany, the Netherlands, and the US marketplaces. Am J Health Econ. 2020;6:139-168. doi:10.1086/706796

58. **McWilliams JM, Schwartz AL.** Focusing on high-cost patients–the key to addressing high costs. N Engl J Med. 2017;376:807-809. [PMID: 28249127] doi:10.1056/NEJMp1612779

59. **Glied S.** How policymakers can foster organizational innovation in health care. Health Affairs Blog. 15 July 2016. Accessed at www.healthaffairs.org/do/10.1377/hblog20160715.055867/full/ on 1 December 2021. doi:10.1377/hblog20160715.055867

60. **McWilliams JM, Chernew ME, Landon BE.** Medicare ACO program savings not tied to preventable hospitalizations or concentrated

among high-risk patients. Health Aff (Millwood). 2017;36:2085-2093. [PMID: 29200328] doi:10.1377/hlthaff.2017.0814

61. **Finkelstein A, Zhou A, Taubman S, et al.** Health care hotspotting—a randomized, controlled trial. N Engl J Med. 2020;382:152-162. [PMID: 31914242] doi:10.1056/NEJMsa1906848

62. **Obermeyer Z, Powers B, Vogeli C, et al.** Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447-453. [PMID: 31649194] doi:10.1126/science.aax2342

63. **Bergquist SL, Layton TJ, McGuire TG, et al.** Data transformations to improve the performance of health plan payment methods. J Health Econ. 2019;66:195-207. [PMID: 31255968] doi:10.1016/j.jhealeco.2019.05.005

64. **van Kleef RC, McGuire TG, van Vliet RCJA, et al.** Improving risk equalization with constrained regression. Eur J Health Econ. 2017;18:1137-1156. [PMID: 27942966] doi:10.1007/s10198-016-0859-1

65. **Mansournia MA, Altman DG.** Inverse probability weighting. BMJ. 2016;352:i189. [PMID: 26773001] doi:10.1136/bmj.i189

66. **Finkelstein A, Taubman S, Wright B, et al; Oregon Health Study Group.** The Oregon Health Insurance Experiment: evidence from the first year. Q J Econ. 2012;127:1057-1106. [PMID: 23293397]

67. **Angrist JD, Pischke J.** The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect. 2010;24:3–30. doi:10.1257/jep.24.2.3

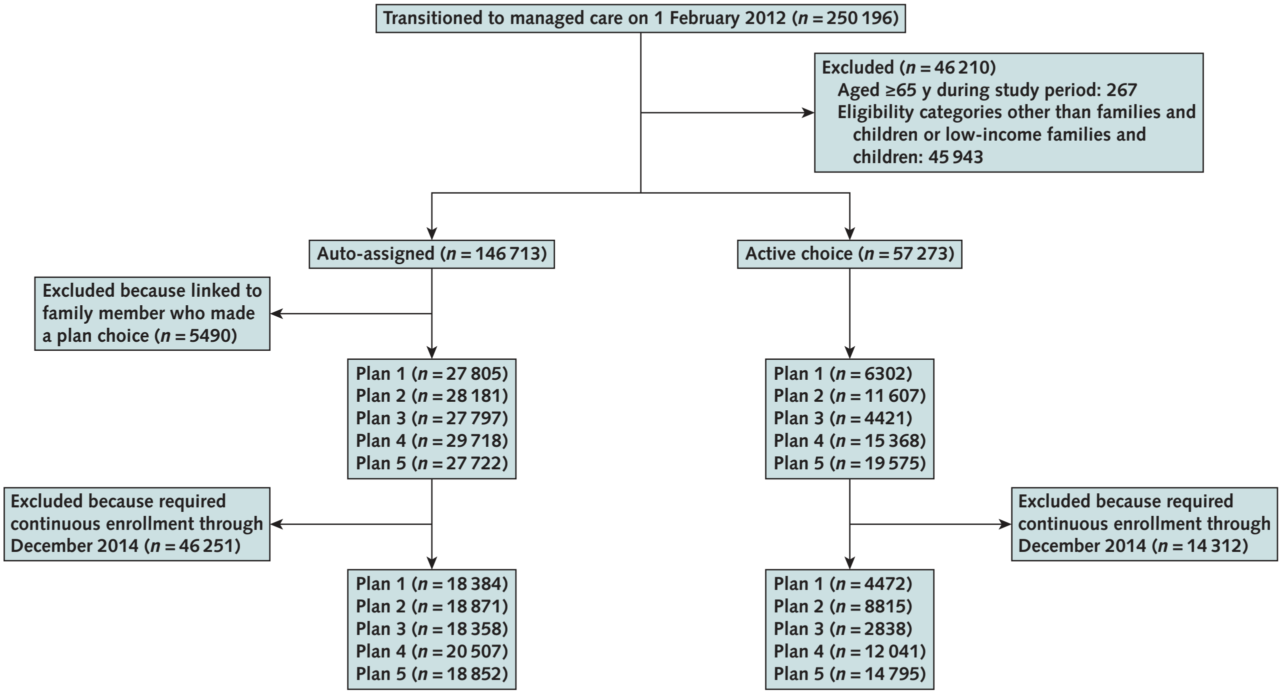
Author Contributions: Conception and design: J. Wallace, J.M. McWilliams, A. Lollo, C.D. Ndumele.
Analysis and interpretation of the data: J. Wallace, J.M. McWilliams, A. Lollo, J. Eaton, C.D. Ndumele.
Drafting of the article: J. Wallace, A. Lollo, J. Eaton.
Critical revision of the article for important intellectual content: J. Wallace, J.M. McWilliams, A. Lollo, C.D. Ndumele.
Final approval of the article: J. Wallace, J.M. McWilliams, A. Lollo, J. Eaton, C.D. Ndumele.
Statistical expertise: J. Wallace, J.M. McWilliams, A. Lollo.
Obtaining of funding: J. Wallace, J.M. McWilliams, C.D. Ndumele.
Administrative, technical, or logistic support: C.D. Ndumele.
Collection and assembly of data: A. Lollo.

Appendix Table 1. Study Population Characteristics, Including Weighted Randomized Population

Characteristic	All Enrollees (*n* = 137 933)	Randomly Assigned Enrollees (*n* = 94 972)	Randomly Assigned Enrollees Weighted by Health Status (*n* = 94 972)	Enrollees Who Made Active Plan Choices (*n* = 42 961)
Age, *percentage points*				
≤5 y	32.3	30.6	30.7	35.9
6–17 y	62.2	63.5	62.8	59.6
18–64 y	5.5	5.9	6.5	4.5
Female, *percentage points*	52.8	52.9	52.6	52.5
Clinical risk groups based on prior claims, *percentage points**				
Nonusers†	7.7	9.3	8.3	4.1
No acute/chronic diagnoses†	62.5	63.4	56.4	60.5
Acute disease	10.5	9.6	12.6	12.6
Single minor chronic disease	8.5	7.9	9.8	9.8
Single dominant or moderate chronic disease	8.6	7.8	10.3	10.3
Predicted health care spending, *$*	1598	1502	1623	1810

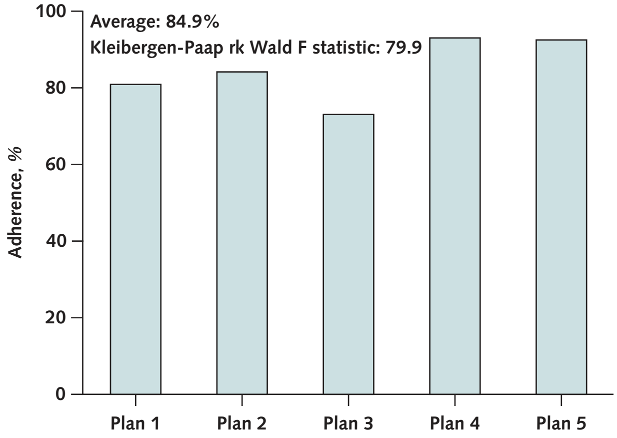
* We only display rows for clinical risk groups that contain >1% of each population's enrollees (5 clinical risk groups are suppressed).
† For purposes of weighting, these categories are combined into a single category: "healthy/nonuser."

Appendix Figure 1. Flow diagram illustrating sample exclusions to arrive at the study populations.


Transitioned to managed care on 1 February 2012 (n = 250 196)
Excluded (n = 46 210)
Aged ≥65 y during study period: 267
Eligibility categories other than families and children or low-income families and children: 45 943
Auto-assigned (n = 146 713)
Active choice (n = 57 273)
Excluded because linked to family member who made a plan choice (n = 5490)
Plan 1 (n = 27 805)
Plan 2 (n = 28 181)
Plan 3 (n = 27 797)
Plan 4 (n = 29 718)
Plan 5 (n = 27 722)
Plan 1 (n = 6302)
Plan 2 (n = 11 607)
Plan 3 (n = 4421)
Plan 4 (n = 15 368)
Plan 5 (n = 19 575)
Excluded because required continuous enrollment through December 2014 (n = 46 251)
Excluded because required continuous enrollment through December 2014 (n = 14 312)
Plan 1 (n = 18 384)
Plan 2 (n = 18 871)
Plan 3 (n = 18 358)
Plan 4 (n = 20 507)
Plan 5 (n = 18 852)
Plan 1 (n = 4472)
Plan 2 (n = 8815)
Plan 3 (n = 2838)
Plan 4 (n = 12 041)
Plan 5 (n = 14 795)

This figure details the sample restrictions we made to arrive at our final analysis population. We obtained claims and enrollment data for 100% of the Medicaid enrollees in our sample state for the period from 2010 to 2016. From these data, we identified 250 196 enrollees who resided in the state's first region to adopt Medicaid managed care and then made the additional exclusions detailed above.

Appendix Figure 2. The effect of plan assignment on plan enrollment: the share of years enrollees remain in their assigned plans, 2013 to 2014.



Adherence refers to the share of enrollees who remain in their randomly assigned plan each calendar year for the plurality of months in 2013 and 2014.

Appendix Table 2. Plan Variation in Performance for the Randomized Versus Observational Population of Enrollees, With Adjustment for Observed Enrollee Characteristics, 2013 to 2014

Outcome	Population Mean	Enrollees Randomly Assigned to Different Plans (*n* = 94 972)*		Enrollees Who Made Active Plan Choices (*n* = 42 961)		Risk-Adjusted Enrollees Who Made Active Plan Choices (*n* = 42 961)	
		Variation in Plan Means (SD)†	*P* Value for Test of Plan Variation†	Variation in Plan Means (SD)†	*P* Value for Test of Plan Variation†	Variation in Plan Means (SD)†	*P* Value for Test of Plan Variation†
Total health care spending, *$*	1682	70.09	<0.001	146.98	<0.001	144.89	<0.001
Annual use of primary and dental care, *percentage points*							
Age-specific well-child visit‡	54	1.74	<0.001	3.95	<0.001	3.49	<0.001
≥1 primary care office visit (children and adults)‡	83	1.60	0.004	1.90	0.023	1.71	0.101
Any dental care‡	64	0.68	0.32	2.10	<0.001	1.83	<0.001
Annual avoidable emergency department visits per 100 enrollees, *n*§	10.70	1.14	<0.001	1.93	<0.001	1.42	<0.001

HEDIS = Healthcare Effectiveness Data and Information Set.

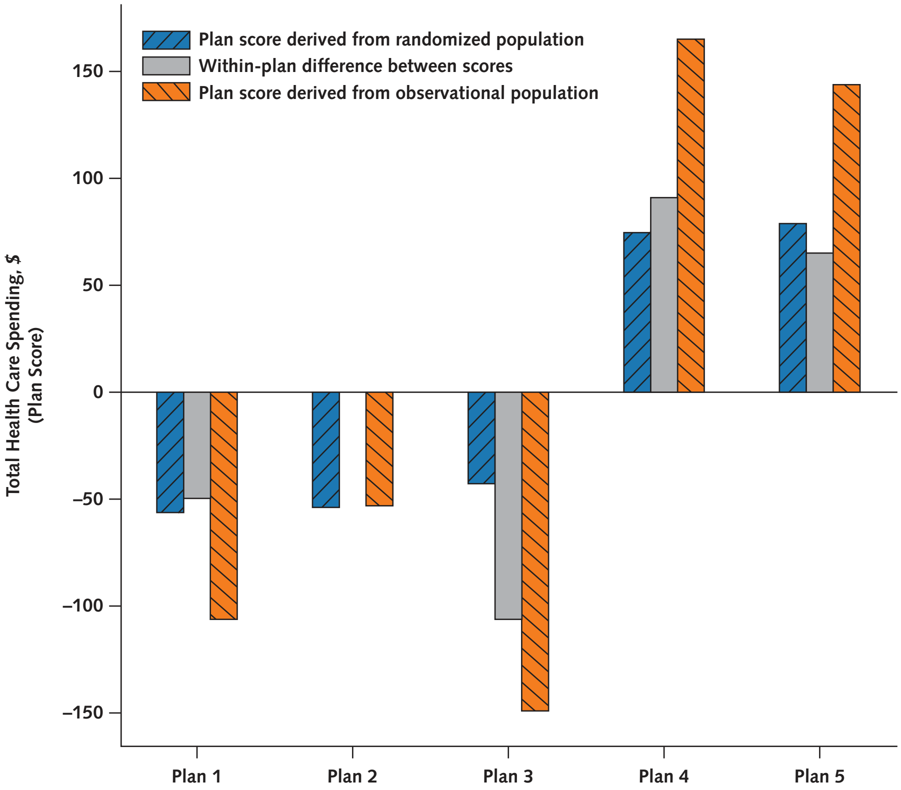
* We weighted the randomly assigned population to balance its distribution of enrollee characteristics with the observational population. For precision, the analyses in the randomized population are adjusted for age; sex; Medicaid eligibility category; and 170 clinical risk groups, defined using the 3M Clinical Risk Grouping software (Supplement Methods V, available at Annals.org).

† The SDs are not corrected for sampling error. The *P* values reflect *F* tests of the joint significance of the plan indicators in a regression model for each outcome within each population (Supplement Methods IX, available at Annals.org). When equalizing the population sizes of the randomized and observational populations, SDs for the observational population remained similarly larger and thus would not be expected from greater sampling error due to the smaller population (Supplement Figure 7, available at Annals.org).

‡ Based on HEDIS specifications for administrative claims measures of primary care and dental care use.

§ Avoidable emergency department visit measure based on the Statewide Collaborative Quality Improvement Project of the California Department of Health Care Services (10).

Appendix Figure 3. Comparison of risk-adjusted observational plan spending scores with those based on the randomly assigned population.



Each bar corresponds to 1 of the 5 plans. The blue area of the bar corresponds to a plan's randomized spending score (relative to the plan mean) based on the randomly assigned population. The orange bar corresponds to a plan's spending score based on the observational population with risk adjustment. The grey unhatched portion indicates the difference between the 2 scores or the extent of residual confounding in the observational scores. For these 5 Medicaid plans, higher-cost enrollees selected plans that control spending to a lesser extent. We calculated a plan score for each plan equal to the plan's deviation from the population-specific plan mean. We compared plan scores between the 2 populations, instead of raw plan means, because population means differed somewhat. Thus, we compared how a plan performed relative to other plans in 1 population with its relative performance in the other population.