Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early renewal of accountable care organization agreements to reset benchmarks in the Medicare Shared Savings Program: a preregistered retrospective cohort study

Samuel Thomas Harrold

samuel.harrold@gmail.com

Independent Researcher

Research Article

Keywords: accountable care organizations, Medicare Shared Savings Program, benchmarks, alternative payment models, value-based payment, preregistration, difference-in-differences, Bayesian analysis

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11565926/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. STH is employed as a data scientist at Indiana University Health. This study was conducted independently of that employment, using only public data, and Indiana University Health had no role in its design, analysis, interpretation or reporting. The views expressed are the author's own and do not represent the official positions of Indiana University Health. The author declares no other competing interests.

Early renewal of accountable care organization agreements to reset benchmarks in the Medicare Shared Savings Program: a preregistered retrospective cohort study

Samuel Thomas Harrold

October 2026

Samuel Thomas Harrold, Independent researcher

ORCID: 0009-0008-4596-6921

Corresponding author: Samuel Thomas Harrold, samuel.harrold@gmail.com

Abstract

Background: Medicare Shared Savings Program (MSSP) benchmarks are reset at each new agreement from an accountable care organization's (ACO's) recent spending, adjusted toward its region. Ending an agreement early reset the benchmark sooner and was a BASIC-track ACO's only way to reach the ENHANCED track's 75% share of savings without leaving the program or waiting out its agreement. No peer-reviewed study has examined why. This study asked which ACOs renewed early, weighing a reset motive (a higher benchmark) against a sharing-rate motive, and whether spending then changed.

Methods: This preregistered retrospective cohort study included the 334 MSSP ACOs whose agreements began July 1, 2019 or January 1, 2020, using public Medicare files (2016-2024). The plan and code were registered before outcome data were read. The reset benchmark was rebuilt from the regulation and validated on 101 other ACOs. 29 ACOs renewing early in January 2022 were compared with 277 stayers on the regional pull (primary), the own-spending component and BASIC risk level, by Bayesian bootstrap, permutation tests and logistic regression. Spending growth in 2022-2024 was compared by difference-in-differences.

Results: The rebuilt benchmark tracked the published one (ρ=0.94). Early renewers' regional pull was 3.4 percentage points higher than stayers' (95% credible interval [CrI], 1.7 to 5.4; *P*<.001), but they also beat their benchmarks by more (own-spending component, −2.1 points; 95% CrI, −3.8 to −0.5), and the total expected gain did not differ (−0.5 points; 95% CrI, −2.0 to 0.8). Post hoc, early renewers' published 2022 benchmarks averaged 1.3% below 2021. Among 206 BASIC-track ACOs, a higher regional pull and BASIC level were associated with early renewal; 26 of all 27 BASIC-track early renewers moved to ENHANCED. Spending growth did not differ detectably (−0.4%; 95% CrI, −4.0% to 2.5%); a preregistered pre-trend condition failed.

Conclusions: Early renewers were already beating their benchmarks; for almost all, renewal was the route to a 75% sharing rate, and the reset did not on average raise benchmarks. This fits the sharing-rate motive better, though the design cannot separate the motives; the rules changed in 2024.

Trial registration: OSF Registries, doi:10.17605/OSF.IO/C53W9. Registered 26 October 2026.

Keywords: accountable care organizations; Medicare Shared Savings Program; benchmarks; alternative payment models; value-based payment; preregistration; difference-in-differences; Bayesian analysis

Background

In the Medicare Shared Savings Program (MSSP), an accountable care organization (ACO) shares in savings when the spending of its assigned beneficiaries falls below a benchmark.[1] As of January 1, 2026, 511 ACOs served 12.6 million assigned Medicare beneficiaries, and ACOs earned $4.1 billion in shared savings for 2024.[2] How benchmarks are set therefore determines

both what Medicare pays and which provider organizations take part. At the start of each agreement period the benchmark is reset from the ACO's own spending in the 3 preceding years and adjusted part of the way toward the spending of its region.[3] Evaluations have found modest savings from the program,[4–6] growing with tenure[5, 6] and larger in physician-group ACOs.[4–6] Whether selection, of clinicians, beneficiaries or whole ACOs, shapes those estimates has been examined with differing results: one analysis found no evidence of favorable risk selection in the original program,[7] while another found that accounting for the nonrandom exit of high-cost clinicians removed the estimated savings,[8] and risk-adjustment rules may lead ACOs to drop high-risk beneficiaries.[9] This study examines a choice the rules added in 2019, ending an agreement early to reset the benchmark sooner, which no peer-reviewed study has examined: under a preregistered plan and with a validated reconstruction of the reset benchmark, it asks which ACOs used the option and what their choice suggests about why.

Benchmarks shape the program through 2 channels. The first is the incentive to save. Rebasing to an ACO's own recent spending weakens its incentive to save, and under earlier weights it rewarded higher spending just before a new agreement;[10] this is the ratchet effect that arises whenever targets are set from past performance.[11] In the MSSP, savings rates were 40% higher in years whose spending did not enter a future benchmark.[12]

The second channel is selection: benchmark rules change which ACOs join, stay and take on risk. ACOs' spending varies widely relative to their regions even after risk adjustment, so moving benchmarks toward regional averages shifts them unevenly.[13] After regional spending entered benchmarks in 2017, participation shifted toward providers whose spending was low relative to their region,[14] ACOs whose effective benchmarks fell because of a separate 2017 change in assignment rules became more likely to leave,[15] ACOs with a positive regional adjustment

were less likely to exit,[12] and the likelihood of earning a bonus rose for ACOs across the efficiency spectrum.[16] Survival in the program has been associated with earning bonuses, higher benchmarks and one-sided risk,[17] and larger ACOs, those with supporting organizations and better performers were more likely to switch to two-sided risk.[18] The 2019 Pathways to Success rules, which shortened the time ACOs could stay in one-sided risk, were followed by more exits than entries,[19] by ACO leaders reassessing potential savings against losses,[20] and by departures with projected costs to Medicare.[21] In a related model, most ACO Investment Model ACOs exited once downside risk was required.[22] This literature describes how benchmark rules sort ACOs into joining, staying, leaving and taking on risk, and how rebasing weakens the incentive to save within an agreement; it has not examined when ACOs choose to reset.

The 2019 rules also allowed an ACO to end its agreement early and begin a new one, which resets the benchmark early.[1] For an ACO in the BASIC track, renewal was also the only way into the ENHANCED track without waiting for its agreement to end or leaving the program, because the final rule published in 2018 bars moving from BASIC to ENHANCED within an agreement period. CMS imposed that bar to guard against selective entry into its highest-reward track by an ACO "very confident of its ability to generate savings" against its current benchmark, and named early renewal, with its rebased benchmark, as the way to make the move.[23] ENHANCED shares 75% of savings, against at most 50% in BASIC.[1] Of the 334 ACOs whose agreements began July 1, 2019 or January 1, 2020, 29 ended them early and began new agreements on January 1, 2022.[24] An industry report found that ACOs rebasing early in 2022 had higher gross savings rates beforehand than ACOs that did not, and that the earlier incumbents among them moved from BASIC to ENHANCED,[25] and a consulting case study

models the choice for a hypothetical ACO.[26] Neither tests why ACOs renewed. To the author's knowledge, no peer-reviewed study has asked whether ACOs used the option to time resets for a higher benchmark or to reach ENHANCED. The question joins the 2 strands above, the incentive created by rebasing and selection by benchmark rules, and it matters more now that the rules in force since 2024 make a reset cheaper for ACOs that have saved.[27]

Two motives predict early renewal. Under a *reset motive*, an ACO renews because its new benchmark would be higher. Under a *sharing-rate motive*, a BASIC-track ACO renews to reach ENHANCED. Whether a reset would raise the benchmark is answered by the *total expected gain*, but its own-spending part is close to a restatement of the savings rate. The registered plan therefore made primary the *regional pull*, how far an ACO's case-mix-adjusted regional spending sat above its current benchmark: it has no term for the ACO's benchmark-year spending and shows whether an ACO had room to gain on the regional side of a reset, which the savings rate does not capture. The own-spending component and the total gain were read beside it. The study asked whether early renewers had more of that room than ACOs that stayed, which characteristics were associated with early renewal, and whether early renewers' spending growth differed afterward, as it might if a new benchmark or a higher sharing rate changed the incentive to save. The analysis plan, hypotheses, decision rules and code were registered before any outcome value was read.[28] These associations bear on the motives but cannot establish which one drove the choice.

Methods

Study Design and Preregistration

This is a preregistered retrospective cohort study of ACO-level public data, reported following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline[29] (checklist in Additional file 1). The analysis plan and the code commit that implements it were registered on OSF (Open Science Framework) Registries on October 26, 2026, 13:22 UTC, with the plan's SHA-256 hash recorded.[28, 30] Before registration only the enrollment files, the Centers for Medicare & Medicaid Services (CMS) data dictionaries, the regulation text and the column names (not the values) of the outcome files were used. The enrollment files include each ACO's track and BASIC level, so the H3 prediction for the BASIC level was made with those data available: it is preregistered but not blind to them. No savings rate, spending value, benchmark or constructed variable was seen. The registry's own timestamp, rather than the repository's commit dates, establishes that order. The code also enforced it: every data file is hashed before it is read, and any file whose hash matches a performance-year results file, a county spending file or the county assignment file could not be opened until the registration record existed. The pipeline was written and tested on synthetic data with planted effects before registration. Analyses ran in stages that each refused to start until the outputs they depend on existed: the power analysis was committed before any confirmatory estimate, the robustness results (which depend only on validation and power) were committed before the confirmatory results and read only together with them, and every results file was committed before its estimates were read.

Data Sources

Four CMS public use datasets were used, all aggregate at the ACO or county level:[24, 31–33] MSSP ACO enrollment files for PY2016-PY2024, performance-year (PY) financial and quality results for PY2016-PY2024 (PY2023 and PY2024 only for the spending analysis), county-level expenditure and risk-score data on assignable beneficiaries for 2019-2021, and counts of ACO-assigned beneficiaries by county for 2021. The 40 files downloaded on October 22, 2026, each recorded with its size, SHA-256 hash and URL, also include files kept from the design stage that no analysis reads; an archived copy of the snapshot is on OSF.[34] Values based on few beneficiaries are hidden in the results and county files, consistent with CMS's policy of suppressing counts of 1 to 10.[35]

Population and Groups

The base population was the 334 ACOs whose PY2021 enrollment record showed an agreement start of July 1, 2019 (189 ACOs) or January 1, 2020 (145 ACOs). Both cohorts' agreements end December 31, 2024, so an ACO from either cohort with a January 1, 2022 agreement in the PY2022 enrollment file had renewed early.[1] *Early renewers* (29 ACOs) were compared with *stayers* (277 ACOs), those in the same agreement in PY2022; 42 stayers later renewed early into 2023 or 2024 agreements, so the comparison concerns the 2022 decision. The 28 ACOs absent from the PY2022 file (left after PY2021) were described but entered no registered test (Figure 1). Of the 356 ACOs from these cohorts in the PY2020 enrollment file, 22 were absent from the PY2021 file, and none had renewed into a 2021 agreement.

Variables

Expected Gain From a Reset

CMS does not publish what an early renewer's benchmark would have been had it not renewed, nor a stayer's had it reset. Following 42 CFR 425.601 as in force on January 1, 2022,[3] the registered plan rebuilt it in 6 steps for each of the 4 enrollment types (end-stage renal disease [ESRD], disabled, aged/dual eligible, aged/non-dual eligible): (1) the equally weighted mean of the ACO's per-capita spending in PY2019-PY2021, restated to PY2021 case mix and trended to 2021 dollars with the national-regional growth blend; (2) risk-standardized county spending averaged with weights equal to the ACO's assigned person-years by county in 2021 (the published proxy for the benchmark year's assignment) and adjusted to the ACO's case mix;[36] (3) the regional adjustment, $w \times (\text{regional} - \text{rebased})$, with the phase-in weight *w* and the cap of 425.601; (4) the approximate new benchmark, combining types by the ACO's PY2021 person-year shares; and (5) the expected gain, 100 × (new − current) / current, where the current benchmark is the PY2021 updated per-capita benchmark. The gain splits into an *own-spending component* (rebased spending relative to the current benchmark, which is close to a restatement of the savings rate) and a *regional component* (the adjustment). Writing the adjustment out shows that, before capping,

$$\text{regional component} = w \times (\text{pull} - \text{own}),$$

so the regional component is also tied to past savings. Step (6) is the primary measure, the *regional pull*,

$$\text{regional pull} = 100 \times \frac{\sum_e s_e R_e - B}{B},$$

where R_e is the case-mix-adjusted regional spending for enrollment type e, s_e the ACO's share of person-years of that type, and B its current per-capita benchmark. All 4 are in percentage points of the current benchmark. They use final PY2021 data, published after the 2022 renewal decision was made during 2021, so they are ex post measures of what ACOs could anticipate from CMS's quarterly reports and their own earlier results.

Validation Before Use

Several steps can only be approximated from public files. The plan therefore required the rebuilt benchmark to be validated against the 2022 benchmarks CMS published for regular renewers (ACOs whose scheduled agreements ended in 2021) before any expected-gain variable was estimated, with a pass rule of a Spearman correlation of at least 0.80 and a median absolute percentage error of at most 5%, and a calibration step if the slope of actual on rebuilt fell outside 0.9 to 1.1. Early renewers were excluded from the validation sample and the 2022 results file was filtered as it was loaded, so validation could not reveal their actual reset. A second check compared the own-spending part with CMS's own benchmark-year spending for the same ACOs; its error SD set the measurement-error SD used in the decision model.

Other Variables

From the results files: the savings rate (CMS-reported savings as a share of the benchmark) in PY2020 and PY2021, assigned beneficiaries, and log per-capita spending, PY2020-PY2024. From the enrollment files: cohort, PY2021 track and BASIC level, revenue category, assignment method and agreement number. An ACO's PY2021 BASIC level reflects both the glide path's schedule and its own earlier choices: ACOs choose their entry level, can advance faster, and

could pause the advance for 2021; ACOs experienced with performance-based risk could enter only Level E or ENHANCED.[37]

Suppressed Values

Hidden cells needed for the expected gain were filled by registered rules in order: exact recovery of a single hidden count from totals, the mean over years in which a value was shown, a regional ratio for a type hidden in every year, and, for a hidden risk score, no restatement and the region's mean risk score. ACOs still missing a variable were excluded from that analysis and counted; the plan required reporting as a limitation any exclusion above 10% of either group.

Hypotheses

The registered hypotheses were: *H2b (primary)*, early renewers had a larger regional pull than stayers; *H2a (secondary)*, a larger own-spending component (close to H1 restated, so not counted as separate evidence); *H1 (descriptive)*, a difference in PY2020 savings rate, with no motive attributed to its sign; *H3*, among BASIC-track ACOs, the probability of renewing early rises with the regional pull, the own-spending component and a higher BASIC level; and *H4*, early renewal changed per-capita spending in PY2022-PY2024 (2-sided). H1 to H3 describe who chose to reset and are associations. H4 concerns a bundled package (a new benchmark, for almost all a move to ENHANCED, and new benchmark years) whose parts cannot be separated.

Statistical Analysis

Group Comparisons (H1, H2a, H2b)

The estimate was the difference in means, early renewers minus stayers, within each cohort, averaged with weights proportional to the early renewers in each cohort. Its posterior was a

Bayesian bootstrap[38] (10,000 draws of Dirichlet(1, …, 1) weights within each cohort-by-group cell), reported as the posterior mean, 95% equal-tailed credible interval (CrI) and posterior probability of a positive value. Significance was by stratified permutation: labels were shuffled within cohort 10,000 times, and the 2-sided *P* value was $(1 + \#\{|\hat{\theta}_{\text{perm}}| \geq |\hat{\theta}_{\text{obs}}|\})/(1 + 10{,}000)$.[39] When the *P* value and the CrI disagree, both are reported and the disagreement noted.

Decision Model (H3)

A Bayesian logistic regression of early renewal among PY2021 BASIC-track ACOs on the 3 core predictors and 5 adjustment factors (cohort, log assigned beneficiaries, high revenue, prospective assignment, later agreement) was fit in PyMC[40] with the No-U-Turn sampler[41] (4 chains of 2,000 tuning and 2,000 retained draws, target acceptance 0.95). Continuous predictors were centered and divided by 2 SD[42] and binary predictors centered, so each coefficient is a log-odds difference per 2 SD or per category; coefficients had Normal(0, 1) priors. The own-spending component entered as a latent true value observed with normal error, with the error SD fixed by the validation step. Convergence required, for every coefficient, $\hat{R} \leq 1.01$ and bulk and tail effective sample sizes of at least 400,[43] and no divergent transitions, with one registered rerun; a Firth penalized regression[44, 45] would have decided H3 had the model not converged. Because shuffling labels in a multi-predictor model tests the joint null, each core predictor was tested by direct simulation under the reduced model: the model without that predictor was fit, 10,000 outcome vectors were simulated from its posterior, the full model was refit to each, and the observed coefficient was compared with that null distribution. A fit with Student-t(3, 0, 2.5) priors, the scale Gelman and colleagues recommend for scaled predictors but

with 3 rather than 1 degree of freedom,[46] was the registered sensitivity analysis, with any differing core conclusion to be flagged.

Spending After Renewal (H4)

For each ACO, Δ was the mean log per-capita spending over its included years in PY2022-PY2024 minus the mean in PY2020-PY2021, and H4 applied the group-comparison estimator to Δ, reported as $100 \times (e^{\text{estimate}} - 1)$, the percent difference in spending growth. Log differences are also given in log points (1 log point = 0.01 in log spending, about 1%). The comparison group was ACOs not yet renewed: a stayer's years after it renewed into a 2023 or 2024 agreement were excluded, following the not-yet-treated design of Callaway and Sant'Anna.[47] With one pre-period change (PY2020 to PY2021, both affected by COVID-19) and about 29 treated ACOs, a nonsignificant pre-trend test is weak evidence of parallel trends,[48] so the plan replaced it with an equivalence condition in the spirit of equivalence testing:[49] H4 would be read as an effect only if the larger absolute bound of the pre-trend CrI, scaled by the mean distance between period midpoints, was smaller than the absolute H4 estimate. A trend-adjusted estimate (the pre-period slope carried forward, the exactly linear special case of the restrictions of Rambachan and Roth[50]) and risk-adjusted, balanced-panel, propensity-score-weighted[47, 51] and single-cohort estimates were reported beside it.

Power

Minimum detectable effects (MDEs) at 80% power and 2-sided $\alpha = .05$ for 29 vs 277 ACOs were computed after registration and committed before any confirmatory estimate: from the cross-ACO SD of the savings rate in PY2016-PY2018 (H1), the SD of each constructed variable

among stayers, computed after validation (H2a, H2b), and the SD of the pre-period spending change (H4). For H3, detection was simulated at a log-odds difference of 0.5.

Decision Rules

Fixed in the plan before any result. For H2b, $P<.05$ with a positive difference was to be reported as "early renewers' regions spent more relative to their current benchmarks, evidence for the reset motive beyond past savings," $P<.05$ with a negative difference as evidence against it, and otherwise as no detectable difference. H2a used the same wording; H3 predictors were supported, contradicted or undetermined by their direct-simulation *P* values; and H4 was reported as no detectable difference, grew less or grew more, read as an effect only if the equivalence condition held. No test here could rule the sharing-rate motive in or out. As registered, there is one primary test, and the 3 H3 tests and the H4 test are not adjusted for multiplicity.

Covariate-Adjusted Estimates, Conventional Methods, and Robustness

Reported beside the registered methods and not used for decisions: covariate-adjusted estimates for H1, H2a and H2b from robust Bayesian linear regressions (Student-t errors on the standardized variable, Normal(0, 1) priors) on early renewal and cohort, or cohort, track, level and size; Welch *t* tests in SciPy;[52, 53] ordinary least squares with HC2 errors[54] and maximum-likelihood logistic regression in statsmodels;[55] 2-way fixed effects with an event study and errors clustered by ACO for H4; and the registered robustness checks: differences in medians, the January 1, 2020 cohort alone, PY2021 savings rates, the Firth model, benchmark years 2-3 only, no measurement-error model, complete cases and unadjusted county spending. Variables were built in dbt[56] on DuckDB;[57] the environment was locked with uv;[58]

Markov chain Monte Carlo diagnostics used ArviZ;[59] every refit had its own seed from a fixed NumPy seed sequence.[60]

Deviations

All 7 analysis stages ran with code identical to the registered commit. After the last stage the lockfile took security updates to 2 packages (filelock and virtualenv); no result was produced with it, and it is logged as the one deviation. One sentence of the plan's background is imprecise: it describes ENHANCED as sharing 75% of savings and losses, whereas it shares 75% of savings and up to 75% of losses.

Post Hoc Checks (Not Registered)

After all registered results had been read, and in response to critical review of the draft, 3 checks were added (eTable 13 in Additional file 2). They reuse the registered pipeline's tables and estimator, change no registered file, and decide nothing. First, the validation was repeated on the change scale, the rebuilt against the actual change from the PY2021 benchmark, for the same 101 regular renewers. Second, early renewers' realized reset, their 2022 historical benchmark as CMS published it relative to their PY2021 benchmark, was compared with the rebuilt gain, and their PY2021-to-PY2022 change in updated benchmark was compared with stayers' with the group-comparison estimator above. Third, the rank correlation of the regional pull with the PY2020 and PY2021 savings rates and the own-spending component was computed, to ask whether the pull restates past savings.

Use of Large Language Models

Claude Code (Anthropic; model Claude Opus 5.5), a large language model tool, was used from October 2026. It proposed the topic and drafted the analysis plan, offering alternatives from which the author chose; the author revised the plan and registered it. Under the author's direction it drafted the analysis pipeline and its tests, first against synthetic data with planted effects; the scripts that generate every table and figure from the committed results and that check every number in the text against them; and, from the author's input, first drafts of most of the text, which the author revised and approved. It ran the literature searches the author specified and retrieved the metadata of references with a DOI from Crossref; the author checked the cited claims against the sources. Separate instances of the same tool critically reviewed the plan, the code and the manuscript, and the author decided which comments to act on. The author chose the research question, hypotheses, decision rules and statistical approach, led the interpretation and conclusions, ran the analysis, verified every reported number against the results files and takes full responsibility for the content. No image-generation tools were used; the figures are drawn by plotting code from the results files.

Results

Validation and Sample

The rebuilt benchmark passed validation (eTable 1 in Additional file 2). For 101 of the 111 eligible regular renewers, it had a Spearman correlation of 0.94 with the benchmark CMS published and a median absolute percentage error of 1.6%; the slope of actual on rebuilt was 0.944, inside the registered range, so no calibration was applied. The own-spending part agreed with CMS's benchmark-year spending (Spearman 0.95; median error 1.4%), and its error SD set

the measurement-error SD at 4.38 percentage points. That SD exceeds the own-spending component's spread among stayers (SD 3.3). Under a classical error model it could not, so the validation errors are probably heavy-tailed (a normal error with that SD would have a median absolute value near 3, against the 1.4% observed) or do not transfer exactly to the study sample; a scale estimate from the median error ($1.4826 \times 1.4 \approx 2.1$) would be below the stayers' SD. By the registered error model, the component is mostly noise at the level of one ACO; the savings rates, which CMS reports directly, are the firmer evidence of how far ACOs were beating their benchmarks.

Early renewers were more often in the January 1, 2020 cohort and at BASIC's two-sided levels C to E in PY2021 than stayers (Table 1); 16 of the 27 BASIC-track early renewers were at the highest level, E. Of those 27, 26 moved to ENHANCED in their new agreements; stayers could not, because the rules bar that move within an agreement period. The expected-gain variables could not be built for 3 early renewers (10.3%) and 36 stayers (13.0%), both above the plan's 10% line. A diagnostic query run after the results, outside the registered pipeline, found that in every one of these 39 ACOs CMS had suppressed the count of assigned ESRD beneficiaries in every county of the 2021 assignment file, so the ESRD part of regional spending had no county weights (for 15, the aged/dual counts were hidden too). The plan's fill rules cover hidden cells in the results files, not fully hidden county weights, so these ACOs were excluded and counted, as registered. The same query showed them to be smaller ACOs (median 6,702 and 8,578 assigned beneficiaries for excluded early renewers and stayers, against 10,757 and 15,649 for those included). One early renewer and 3 stayers had hidden values filled by the registered rules.

Table 1. Characteristics of the 334 ACOs in the 2019 and 2020 agreement cohorts, by group

Characteristic	Early renewers	Stayers	Left after PY2021
ACOs, No.	29	277	28
Agreement start, No. (%)			
July 1, 2019	11 (38)	159 (57)	19 (68)
January 1, 2020	18 (62)	118 (43)	9 (32)
Track in PY2021, No. (%)			
BASIC levels A-B (one-sided)	7 (24)	133 (48)	23 (82)
BASIC levels C-E (two-sided)	20 (69)	76 (27)	4 (14)
ENHANCED	2 (7)	68 (25)	1 (4)
Assigned beneficiaries, PY2020, mean (SD), %			
Aged 85 y or older	10.9 (3.7)	12.4 (4.1)	13.3 (5.8)
Female	56.8 (1.9)	57.1 (2.3)	57.3 (2.5)
Savings rate, PY2020, pp			
Mean (SD)	6.4 (3.6)	4.6 (3.9)	2.7 (5.5)
Median	5.8	4.3	1.5
Regional pull, pp of benchmark			
ACOs with the variable, No.	26	241	20
Mean (SD)	5.0 (5.0)	1.8 (4.6)	4.8 (6.7)
Median	3.5	1.4	2.0
Own-spending component, pp of benchmark			
ACOs with the variable, No.	26	241	20
Mean (SD)	−5.3 (4.4)	−3.2 (3.3)	−1.8 (4.6)
Median	−4.2	−2.9	−1.0
Regional component, pp of benchmark			
ACOs with the variable, No.	26	241	20
Mean (SD)	3.7 (1.6)	2.2 (1.8)	2.0 (2.3)
Median	3.8	2.1	1.3

Abbreviations: ACO, accountable care organization; pp, percentage points; PY, performance year. Early renewers ended their agreements and began new ones on January 1, 2022; stayers remained in their agreements; leavers left the program after PY2021. The savings rate is CMS's reported savings as a share of the benchmark (positive when spending was below the benchmark).

The expected-gain variables could not be built for every ACO; a diagnostic query run after the results, outside the registered pipeline, traced each exclusion to county counts of assigned beneficiaries with end-stage renal disease that were all suppressed (Results). Leavers were not compared in any registered test.

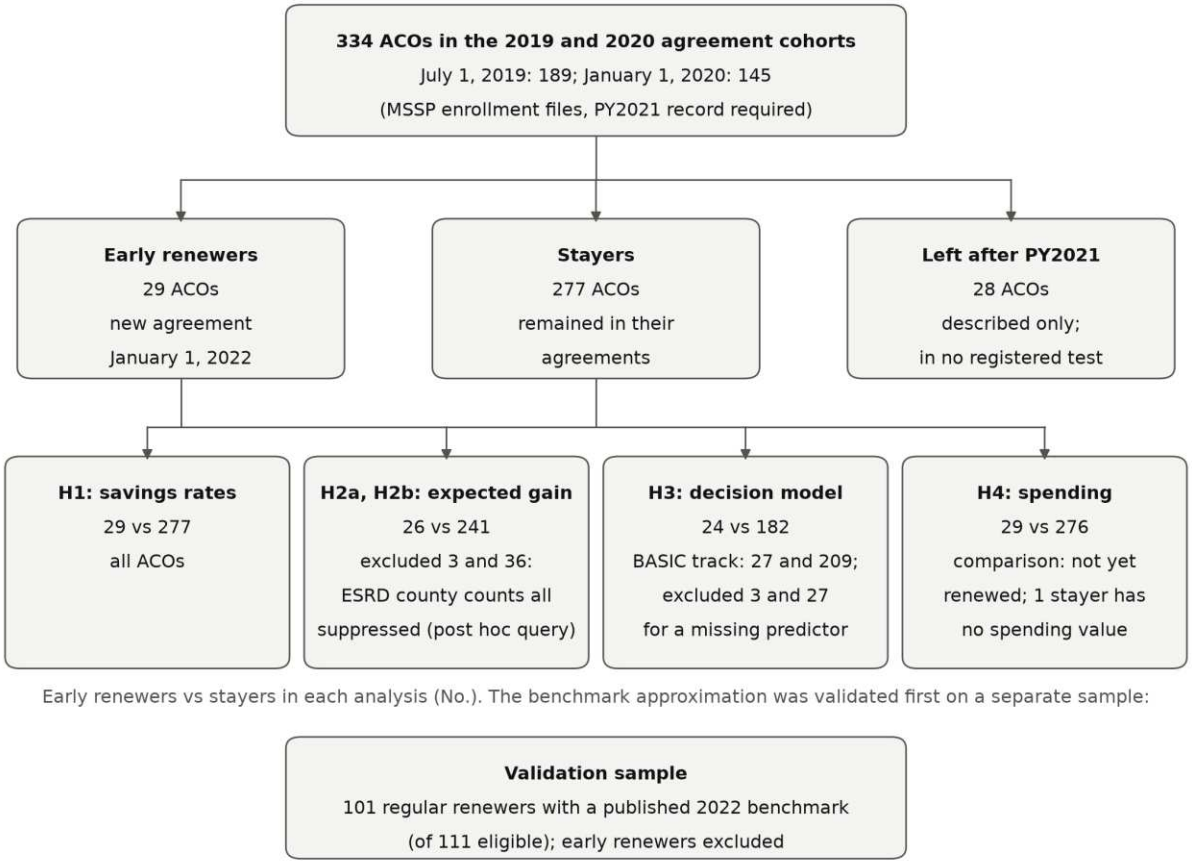


Figure 1. Flow of ACOs into each analysis. Counts are early renewers vs stayers. The reason for the expected-gain exclusions comes from a diagnostic query run after the results, outside the registered pipeline. ACO, accountable care organization; ESRD, end-stage renal disease; PY, performance year. Drawn from the results files by code drafted with Claude Code (see Use of Large Language Models).

Who Renewed Early

Relative to their current benchmarks, early renewers' case-mix-adjusted regional spending was 3.4 percentage points higher than stayers' (95% CrI, 1.7 to 5.4; *P*<.001), above the MDE of 2.5 computed for 29 vs 277 ACOs, which is slightly understated for the 26 vs 241 compared (Table 2 and Figure 2). By the registered rule, H2b was supported, which the plan worded as "evidence for the reset motive beyond past savings." The decomposition shows that this wording overstates the result. Weighted by the phase-in weight, the pull difference was 1.6 points (eTable 10), and the regional component differed by the same 1.6 points partly because a lower own-spending component also raises it; the total expected gain did not differ. The pull difference therefore indicates that early renewers were low-cost relative to their regions more than that they stood to gain a higher benchmark. The pull is related to past savings without restating it: post hoc, its rank correlation with the PY2020 savings rate was 0.45 among the 267 compared ACOs (eTable 13 in Additional file 2). At the median PY2021 benchmark of the study's ACOs, $11,312 per beneficiary per year, the difference of 3.35 points is about $379 per beneficiary-year of regional spending relative to the benchmark; the difference in the regional component itself, what a reset would add, was 1.6 points, about $176.

The own-spending component pointed the other way (−2.1 points; 95% CrI, −3.8 to −0.5; *P*=.003): early renewers' recent spending was lower relative to their benchmarks, contrary to the reset motive by the rule's wording. Early renewers also had higher savings rates in PY2020 (1.9 points; 95% CrI, 0.7 to 3.3) and PY2021 (1.6 points; 95% CrI, 0.2 to 3.1). The 2 parts of the gain offset: the regional component was larger for early renewers (1.6 points; 95% CrI, 0.9 to 2.2), the own-spending component smaller, and the total expected gain did not differ (−0.5 points; 95% CrI, −2.0 to 0.8; *P*=.33). In unweighted means the own-spending and regional parts summed to

−5.3 + 3.7 = −1.6 points for early renewers and −3.2 + 2.2 = −1.0 for stayers: by the approximation, a reset would on average have lowered both groups' benchmarks. Post hoc, CMS's published 2022 benchmarks bear this out for early renewers: their reset historical benchmarks averaged 1.3% below their PY2021 benchmarks (median −1.5%; 66% of the 29 were lower), the rebuilt gain had predicted −1.6 points on average against −1.4 realized for the 26 with both values, and their updated benchmarks rose from PY2021 to PY2022 no more than stayers' (difference 0.0 points; 95% CrI, −2.0 to 2.2) (eTable 13 in Additional file 2).

Table 2. Early renewers compared with stayers on the hypothesis tests and the expected gain

Measure (role)	Early / stayers, No.	Difference (95% CrI)	Pr(>0)	*P*	MDE	Verdict
H2b, regional pull (primary)	**26 / 241**	**+3.4 (1.7 to 5.4)**	**>0.999**	**<.001**	**2.5**	**Supported**
H2a, own-spending component (secondary)	26 / 241	−2.1 (−3.8 to −0.5)	0.006	.003	1.8	Contradicted
H1, savings rate, PY2020 (descriptive)	29 / 277	+1.9 (0.7 to 3.3)	>0.999	.01	2.7	
H1, savings rate, PY2021 (descriptive)	29 / 277	+1.6 (0.2 to 3.1)	0.988	.03	2.7	
Total expected gain (descriptive)	26 / 241	−0.5 (−2.0 to 0.8)	0.233	.33		
Regional component (descriptive)	26 / 241	+1.6 (0.9 to 2.2)	>0.999	<.001		

Abbreviations: CrI, credible interval; MDE, minimum detectable effect at 80% power, two-sided $\alpha = .05$, computed before the tests for 29 vs 277 ACOs (for the 26 vs 241 compared on the expected-gain variables it is 2.6 for H2b and 1.9 for H2a); Pr(>0), posterior probability that the difference is positive; pp, percentage points; PY, performance year. Differences are early renewers minus stayers within agreement cohort, averaged with weights proportional to early renewers per cohort. Posterior means and 95% equal-tailed CrIs are from a Bayesian bootstrap (10,000 draws); P values are from stratified permutation (10,000 shuffles within cohort).

Expected-gain variables are in percentage points of the current benchmark; savings rates in percentage points. Verdicts follow the registered decision rules; H1 and the two descriptive rows carry none. Bold: the primary hypothesis.

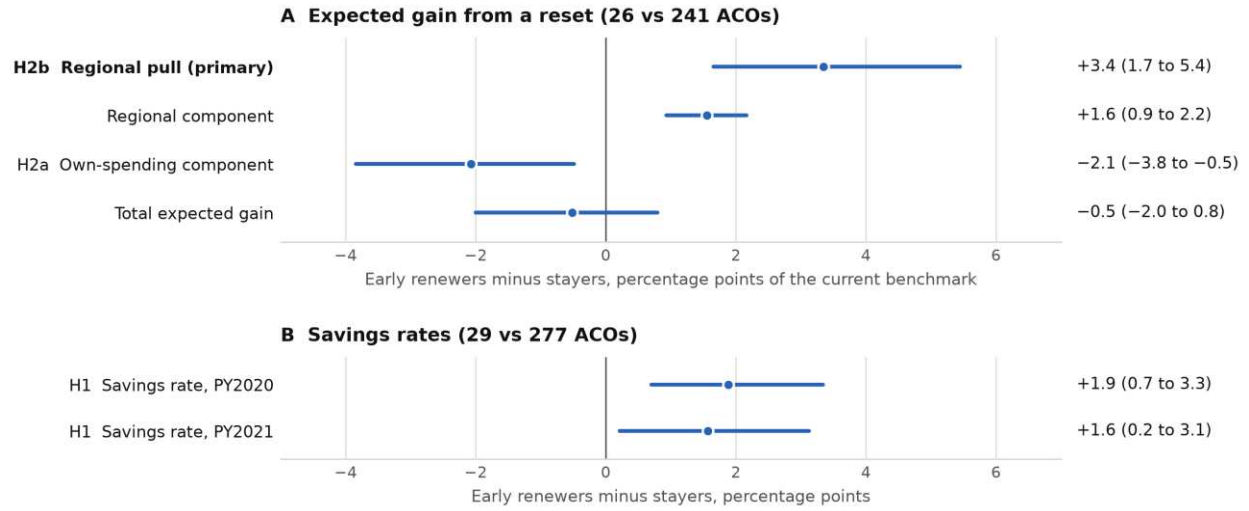


Figure 2. Early renewers minus stayers. A, Expected-gain variables, in percentage points of the current benchmark. B, Savings rates, in percentage points. Points are posterior means and lines 95% credible intervals from the Bayesian bootstrap. Drawn from the results files by code drafted with Claude Code (see Use of Large Language Models).

Decision Model

Among 206 BASIC-track ACOs (24 early renewers), early renewal was associated with a higher regional pull (log-odds difference per 2 SD, 1.34; 95% CrI, 0.44 to 2.30; direct-simulation *P*=.002) and a higher BASIC level (log-odds difference for levels C-E vs A-B, 1.34; 95% CrI, 0.38 to 2.31; *P*=.003); both predictions were supported (Table 3 and Figure 3). A log-odds difference of 1.34 multiplies the odds of renewing early by about 3.8. The own-spending component was undetermined (−0.05; 95% CrI, −2.01 to 1.92; *P*=.19); its interval is the prior's 95% range, because its measurement error exceeds its spread and the model could learn almost

nothing about it. Without the measurement-error layer it was −0.64 (95% CrI, −1.51 to 0.23). The model converged on its first fit (largest $\hat{R}$ 1.001, no divergences), and the observed 24 early renewers fell inside the posterior predictive 95% range of 14 to 37 (eTable 5 in Additional file 2). Under the Student-t prior, which converged on its registered rerun, the pull (1.65; 95% CrI, 0.53 to 2.87) and BASIC-level (1.63; 95% CrI, 0.53 to 2.85) intervals also excluded zero, so no core conclusion differs between priors. The power analysis gave detection rates of only 17% (pull) and 14% (BASIC level) at a log-odds difference of 0.5, so estimates that pass the test are likely to overstate the true effects;[61] the directions are the result, and coefficients near 1.3 are probably too large.

Table 3. Decision model of early renewal among BASIC-track ACOs (H3)

Predictor	Posterior mean (95% CrI)	Pr(>0)	*P*	Verdict	Student-t prior (95% CrI)	Conventional OR (95% CI)
Regional pull	1.34 (0.44 to 2.30)	0.999	.002	Supported	1.65 (0.53 to 2.87)	4.01 (1.20 to 13.40)
Own-spending component	−0.05 (−2.01 to 1.92)	0.480	.19	Undetermined	−0.56 (−9.98 to 7.36)	0.53 (0.19 to 1.44)
Higher BASIC level (C-E)	1.34 (0.38 to 2.31)	0.998	.003	Supported	1.63 (0.53 to 2.85)	4.79 (1.47 to 15.59)
2020-01-01 cohort	0.63 (−0.24 to 1.51)	0.920		Adjustment	0.73 (−0.26 to 1.76)	1.97 (0.71 to 5.49)
Log assigned beneficiaries	0.45 (−0.49 to 1.36)	0.827		Adjustment	0.66 (−0.43 to 1.77)	2.00 (0.65 to 6.15)
High revenue	−0.53 (−1.69 to 0.59)	0.182		Adjustment	−0.55 (−2.10 to 0.85)	0.63 (0.14 to 2.89)
Prospective assignment	−0.77 (−1.79 to 0.19)	0.058		Adjustment	−1.01 (−2.27 to 0.15)	0.37 (0.11 to 1.31)
Later agreement	−0.67 (−1.62 to 0.31)	0.089		Adjustment	−0.91 (−2.04 to 0.22)	0.38 (0.12 to 1.19)

Abbreviations: CrI, credible interval; OR, odds ratio; Pr(>0), posterior probability that the coefficient is positive. Bayesian logistic regression of early renewal among 206 BASIC-track ACOs (24 early renewers; 3 early renewers and 27 stayers excluded for a missing predictor).

Continuous predictors are centered and divided by 2 SD, so their coefficients are log-odds differences per 2 SD; binary predictors are centered, so theirs are per category (the BASIC level compares levels C-E with A-B). Core predictors (bold) are tested by direct simulation under the reduced model (10,000 simulated datasets); adjustment factors are estimates only. The own-spending component enters with a measurement-error model (σ_v = 4.38 pp). The Student-t column is the registered prior sensitivity analysis; the maximum-likelihood ORs are the conventional counterpart, reported only.

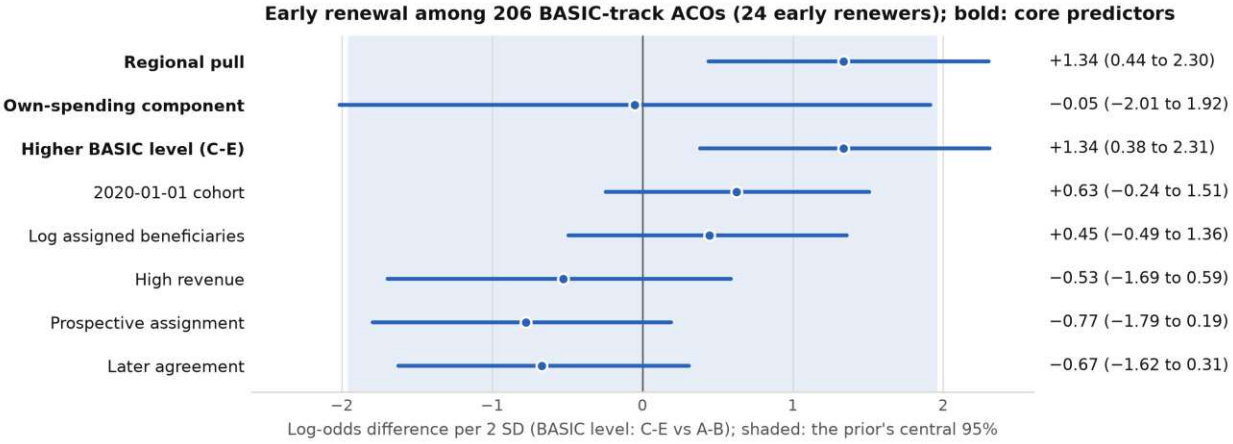


Figure 3. Decision model of early renewal among BASIC-track ACOs (H3). Points are posterior means and lines 95% credible intervals; the shaded band is the Normal(0, 1) prior's central 95%. Bold labels mark the core predictors. Drawn from the results files by code drafted with Claude Code (see Use of Large Language Models).

Spending After Renewal

Spending growth from PY2020-PY2021 to PY2022-PY2024 did not differ detectably between 29 early renewers and 276 comparison ACOs not yet renewed (−0.4%; 95% CrI, −4.0% to 2.5%; *P*=.78; MDE 3.8 log points) (Figure 4; eTable 8 in Additional file 2). The pre-trend estimate was 0.7% (95% CrI, −2.3% to 4.9%; *P*=.58), not evidence against parallel trends, but the equivalence

condition did not hold: the scaled pre-trend bound, 2.41 years times 4.8 log points, was 11.5 log points (about 12%), far above the absolute H4 estimate of 0.4 log points. The condition could have held only for an estimate above about 12% in absolute value, 3 times the MDE, so with one pre-period it could not support reading an estimate of plausible size as an effect, and a trend difference of the size the pre-period allows could hide an effect in either direction. H4 is therefore reported as a difference in spending changes that cannot be told apart from a continuation of a pre-period difference in trends. Under parallel trends, the interval rules out a drop in spending growth of more than about 4% or a rise of more than about 2.5%; the trend-adjusted interval (−12.0% to 2.7%) is too wide to inform.

Attrition was uneven: by PY2024, 115 of 277 stayers (42%) had left the spending panel (72 left the program, 42 renewed and 1 had no spending value), against 4 of 29 early renewers (14%) (eTable 9 in Additional file 2). The balanced panel of ACOs with all 5 years gave −1.6% (95% CrI, −5.6% to 1.6%; *P*=.30).

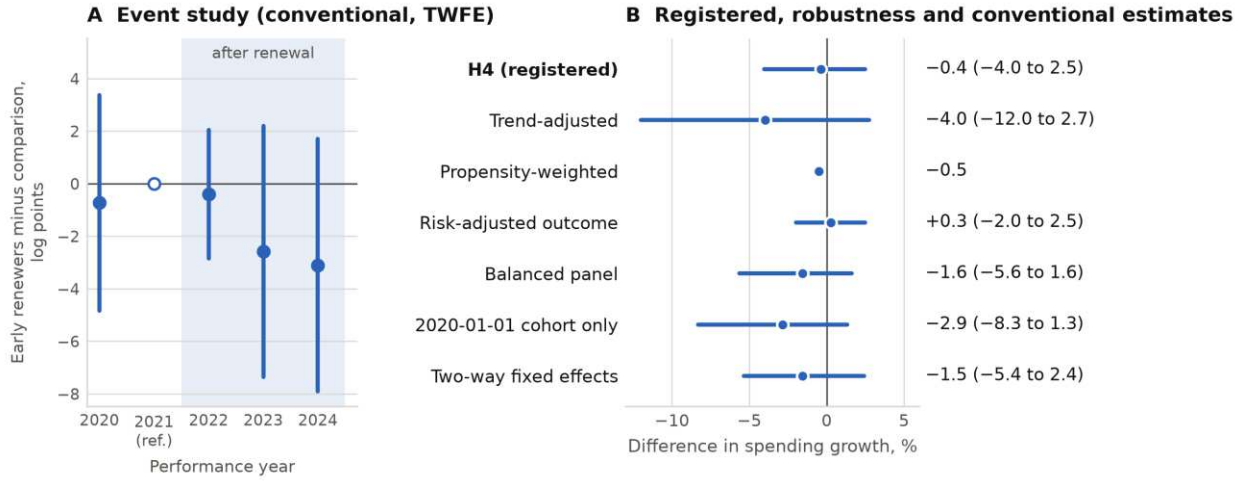


Figure 4. Spending after early renewal. A, Event study from the conventional 2-way fixed-effects model, in log points relative to PY2021, with 95% CIs clustered by ACO. B, The registered H4 estimate and the robustness and conventional estimates, as the percent difference in spending

growth with 95% intervals (the propensity-weighted estimate has no interval); sample sizes are in eTable 8 in Additional file 2. TWFE, 2-way fixed effects. Drawn from the results files by code drafted with Claude Code (see Use of Large Language Models).

Robustness and Conventional Methods

Every registered verdict agreed with its conventional counterpart (eTable 7 in Additional file 2). For H2b, ordinary least squares with HC2 errors gave 3.3 points (95% CI, 1.3 to 5.4; *P*=.001), and for H3 the maximum-likelihood odds ratios were 4.0 (95% CI, 1.2 to 13.4) for the regional pull and 4.8 (95% CI, 1.5 to 15.6) for the BASIC level; 2-way fixed effects gave −1.5% (95% CI, −5.4% to 2.4%) for H4. H2b kept its direction in every robustness check (eTable 3 in Additional file 2), but its size depended on the estimator: the difference in medians was 2.8 points, and the robust (Student-t) regressions gave 2.1 (95% CrI, 0.5 to 3.8) with cohort and 2.2 (95% CrI, 0.7 to 3.8) with track, level and size added, below the MDE of 2.5, so a few large values among early renewers carry part of the difference in means; without the case-mix adjustment of regional spending the difference was 6.9. H2a and H1 kept their direction but were weaker by medians (H2a, *P*=.09; H1, *P*=.06) and in the adjusted regressions, whose intervals reached zero (H2a, −2.6 to 0.02; H1, −0.1 to 2.6). In H3 the BASIC-level interval stayed above zero in every check, while the pull interval reached zero with the 2020 cohort alone (1.15; 95% CrI, −0.02 to 2.39) and with unadjusted county spending (0.75; 95% CrI, −0.02 to 1.56) (eTable 4 in Additional file 2).

Among the exploratory analyses (eTables 10-12 in Additional file 2), the 28 ACOs that left after PY2021 had the lowest PY2020 savings rates (mean 2.7 points, against 6.4 for early renewers and 4.6 for stayers) and, for the 20 with the variable, a mean regional pull similar to early

477 renewers' (4.8 vs 5.0; medians 2.0 vs 3.5) (Table 1). The 2 early renewers already in
478 ENHANCED, whose renewal could not raise the sharing rate, fell between the 26th and 68th
479 percentiles of stayers on every variable.

480 # Discussion

481 ## Principal Findings

482 In this preregistered study of the 2019 and 2020 MSSP cohorts, the ACOs that ended their
483 agreements early were already beating their benchmarks by more than ACOs that stayed, had
484 more room under their regions' spending, and, in the BASIC track, were more often already
485 bearing two-sided risk. For 26 of the 27 BASIC-track early renewers, renewal was the move to
486 ENHANCED's 75% sharing rate. By the study's approximation, a reset would not on average
487 have raised their benchmarks, because the regional gain and the own-spending loss roughly
488 offset, and the benchmarks CMS published for 2022, examined post hoc, agree. Spending growth
489 in the 3 years after renewal did not differ detectably from that of ACOs not yet renewed.

490 To the author's knowledge, this is the first peer-reviewed study of which ACOs used the early-
491 renewal option. It adds 3 things. First, the intuitive account that ACOs renewed early to raise
492 their benchmarks does not hold on average. This rests on a secondary decomposition, built from
493 data the ACOs did not yet have when they decided, and on the benchmarks CMS later published,
494 examined post hoc. Second, early renewers were sorted on the positions found for joining and
495 staying, room under their regions' spending and greater risk, which extends that literature to
496 when ACOs reset; the size of the regional difference depended on the estimator, and in the
497 decision model its interval reached zero in 2 robustness checks. Third, the study provides a
498 public, validated reconstruction of the reset benchmark and a preregistered, code-enforced

analysis that can be rerun for later cohorts. Spending growth after renewal did not differ detectably; under parallel trends, which the single pre-period could not establish, the interval rules out a drop of more than about 4% or a rise of more than about 2.5%.

Comparison With Prior Work

The primary result extends to a new decision the selection described in the Background: providers with low spending for their region joining,[14] ACOs whose effective benchmarks fell leaving,[15] and ACOs with a positive regional adjustment about 70% less likely to exit.[12] Here the choice is when to reset rather than whether to join or leave, and early renewers had, on average, more room under their regions' spending, the position associated with joining and staying. In a Medicare bundled-payment program made voluntary for half of its hospitals after 2 mandatory years, hospitals that could gain without changing their behavior were more likely to opt in, which the authors called selection on levels;[62] early renewal shows selection on a similar position, though the gain that most early renewers took appears to have been the sharing rate rather than a higher benchmark. The industry report on the same cohort[25] is consistent with the higher prior savings rates here; it places the earlier incumbents among the early renewers at BASIC levels C to E before they renewed but does not compare levels with stayers. Beyond that report, this study rebuilt the reset benchmark and validated it against CMS's, separated its own-spending and regional parts, modeled the decision within the BASIC track and followed spending afterward, under a preregistered plan. The finding that early renewers were savers also fits reports that ACOs earning bonuses survived longer in the program[17] and, with borderline precision, that ACOs that had achieved savings were more likely to stay after the 2019 rules.[19] The same success, however, meant that the own-spending part of a reset would have cut their benchmarks. That the size of the regional difference doubled without case-mix

adjustment is consistent with evidence that differences between ACOs and their regions are sensitive to risk adjustment.[13] The concern that rebasing rewards higher spending before a new agreement[10] could be neither confirmed nor excluded here, and the costs to Medicare of early renewal, which earlier work estimated for departures,[21] were not estimated.

The association with a higher BASIC level bears on the sharing-rate motive at least as much as on the reset motive. In these data renewing early and moving to ENHANCED were one choice: 26 of the 27 BASIC-track early renewers moved to ENHANCED, and the rules let no stayer do so. ACOs already bearing two-sided risk at Level E faced a short step to ENHANCED's 75% sharing rate. Because ACOs partly choose their BASIC level, and ACOs experienced with performance-based risk could enter only Level E or ENHANCED,[37] the association may also reflect earlier risk experience or appetite rather than only proximity to ENHANCED. A high regional pull is also a plausible reason to want the higher rate, because an ACO with room under its region's spending may expect larger savings to share. The 2 early renewers already in ENHANCED, for whom only the reset could change, looked like typical stayers on every variable. Taken together, with renewal and the move to ENHANCED one choice, the higher BASIC levels, a total expected gain near zero and reset benchmarks that fell, the pattern fits the sharing-rate motive better than the reset motive, and it is the trade CMS built into the rules: the move to ENHANCED required a reset.[23] That is an interpretation, not a test; the design cannot separate the motives, and 2 ACOs already in ENHANCED are too few to weigh.

Implications

For program design, the results suggest that under the 2022 rules the use of the early-renewal option is consistent with how CMS described it: a route to the ENHANCED track for ACOs

already beating their benchmarks, at the price of a reset that did not on average raise those benchmarks. They do not suggest that ACOs used the option to time resets for a higher benchmark. The rules have since changed. For agreement periods beginning January 1, 2024 or later, an ACO that saved in its benchmark years can receive a prior savings adjustment, and its benchmark gets the higher of that adjustment and the regional adjustment; a regional adjustment that is negative overall is not applied, and negative amounts within a positive one are reduced for ACOs with more dually eligible or higher-risk beneficiaries; and the annual update blends in a fixed projected trend.[27] A reset now costs a saving ACO less, and saving ACOs are the ones that renewed early, so the option may become more attractive to them and its cost to Medicare may change; its use is worth monitoring, and the 34 stayers that renewed early into 2024 agreements are a first sample. Two questions remain open: what early renewal cost Medicare, and how much of the choice reflected the sharing rate rather than the benchmark, which a period in which ACOs could move to ENHANCED without resetting would separate.

Strengths and Limitations

The study's strengths are in its design. The plan, decision rules and code were registered before any outcome data were read, and the code enforced that order; the benchmark CMS does not publish was rebuilt from the regulation and validated before use; the measurement error found in validation was carried into the decision model; and every registered verdict agreed with its conventional frequentist counterpart. The code, data snapshot and results are public, so every estimate can be rerun.

It also has limitations. First, the motives are bundled, and the regional pull, though it has no term for benchmark-year spending, is not free of the ACO's own data: its benchmark was set from

earlier spending, and regional spending is adjusted with the ACO's own PY2021 risk scores, so a high pull partly measures how low-cost the ACO had been relative to its region and may carry differences in coding intensity. The case-mix adjustment has no published counterpart to validate it. Second, the expected gain is an approximation built from final PY2021 data that ACOs did not have when they decided. Validation was on benchmark levels, which vary widely across ACOs, while the hypotheses concern differences of a few points of the benchmark, so it says less about the accuracy of the gain than its correlation suggests: repeated post hoc on the change scale, it gave a Spearman correlation of 0.61 and a median absolute error of 1.7 points with little mean bias (−0.3), so group means are better supported than any one ACO's value; the own-spending component's estimated error exceeds its spread, so the own-spending comparison and the total gain are weak evidence for any single ACO, and the regional part could not be checked. Third, with 29 early renewers, 24 in the decision model, the decision model was underpowered and its supported coefficients are probably overestimates. ACOs under common ownership may decide together, and ACOs in the same market share regional spending, so the effective sample is smaller than the ACO count, and the permutation tests, which treat ACOs as exchangeable within cohort, may overstate precision. Fourth, exclusions exceeded the plan's 10% line in both groups, and the excluded ACOs were smaller, so the expected-gain comparisons and the decision model describe the larger ACOs. Fifth, the spending comparison had one COVID-19-affected pre-period; uneven attrition left the comparison group's later years increasingly selected; the decision was made during 2021, the last pre-period year and a benchmark year of the new agreement, while trade groups were asking CMS to allow pre-pandemic benchmark years,[63] so anticipation and timing effects are possible; and spending can move through changes in participant lists or assignment method (which changed at PY2022 for 2 of 29 early renewers and

14 of 277 stayers) rather than through behavior. Sixth, the base population required a PY2021 enrollment record (22 cohort ACOs left before then); stayers include 42 ACOs that renewed early in 2023 or 2024, so the comparisons and the decision model describe the 2022 decision rather than early renewal at any time; and leavers, a third choice with a mean pull similar to early renewers', were set aside by comparing early renewers only with stayers. Seventh, the results describe the rules for agreements beginning in 2022.

Conclusions

In this cohort, early renewal looks less like timing a reset for a higher benchmark than like the price of reaching ENHANCED's 75% sharing rate: the reset did not on average raise early renewers' benchmarks, and almost all moved to ENHANCED. Spending growth afterward showed no detectable difference. These are associations, measured with public data under a preregistered plan, and the design cannot separate the motives. They extend the evidence that benchmark rules sort ACOs to the timing of resets and give a baseline for the rules in force since 2024.

List of abbreviations

ACO: accountable care organization

CFR: Code of Federal Regulations

CMS: Centers for Medicare & Medicaid Services

CrI: credible interval

ESRD: end-stage renal disease

MDE: minimum detectable effect

MSSP: Medicare Shared Savings Program

OSF: Open Science Framework

PY: performance year

Declarations

Ethics approval and consent to participate

The study used only publicly available, aggregate data on organizations and counties published by CMS. It involved no intervention or interaction with individuals and no identifiable private information, so it was not human subjects research as defined by the US Federal Policy for the Protection of Human Subjects (45 CFR 46.102(e)(1)).[64] Ethics review was therefore not sought, and informed consent, privacy protections for participants and compensation did not apply.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are available in the Zenodo repository (https://doi.org/10.5281/zenodo.23737045) and the OSF project.[34] All data are public CMS files.[24, 31–33] The list of the 40 files downloaded, with sizes, SHA-256 hashes and URLs, is in the study repository at the registered commit,[30] and a script that downloads and verifies them is in the repository at a later commit (https://github.com/stharrold/stharrold.github.io/tree/a0ef82f5c63bf1c0f26e88597355e8d7afaa90

95/analysis/mssp-benchmark-ratchet); an archived copy of the snapshot and the results notebook exports are in the OSF project.[34] The committed results files, from which every table and figure here is generated, are in the study repository and in this manuscript's repository, archived on Zenodo with the post hoc scripts. The registered analysis code (commit a4064f5) and its tests are in the study repository,[30] and step-by-step instructions to rerun and adapt the analysis are in its README at the later commit. A longer account of the study for a general audience was published online (https://stharrold.github.io/20261027-preregistered-mssp-early-renewal.html).

Competing interests

STH is employed as a data scientist at Indiana University Health. This study was conducted independently of that employment, using only public data, and Indiana University Health had no role in its design, analysis, interpretation or reporting. The views expressed are the author's own and do not represent the official positions of Indiana University Health. The author declares no other competing interests.

Funding

This work received no funding.

Authors' contributions

STH: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing (original draft) and writing (review and editing). STH read and approved the final manuscript and is accountable for all aspects of the work.

Acknowledgements

Not applicable.

Additional files

Additional file 1 (.docx). STROBE checklist. Where each item of the STROBE statement for cohort studies is reported.

Additional file 2 (.docx). eTables and eMethods. Registration record and deviations, validation, power, robustness, conventional methods, exploratory analyses and post hoc checks (eTables 1-13).

References

1. Centers for Medicare & Medicaid Services. 42 CFR part 425: Medicare Shared Savings Program. 2022. https://www.ecfr.gov/on/2022-01-01/title-42/chapter-IV/subchapter-B/part-425. Accessed 28 Oct 2026.

2. Centers for Medicare & Medicaid Services. Shared Savings Program fast facts: As of January 1, 2026. 2026. https://www.cms.gov/files/document/2026-shared-savings-program-fast-facts.pdf. Accessed 28 Oct 2026.

3. Centers for Medicare & Medicaid Services. 42 CFR 425.601: Establishing, adjusting, and updating the benchmark for agreement periods beginning on July 1, 2019, and in subsequent years. 2022. https://www.ecfr.gov/on/2022-01-01/title-42/chapter-IV/subchapter-B/part-425/subpart-G/section-425.601. Accessed 28 Oct 2026.

4. McWilliams JM, Hatfield LA, Chernew ME, Landon BE, Schwartz AL. Early Performance of Accountable Care Organizations in Medicare. New England Journal of Medicine. 2016;374:2357–66. https://doi.org/10.1056/nejmsa1600142.

5. McWilliams JM, Hatfield LA, Landon BE, Hamed P, Chernew ME. Medicare Spending after 3 Years of the Medicare Shared Savings Program. New England Journal of Medicine. 2018;379:1139–49. https://doi.org/10.1056/nejmsa1803388.

6. Bond AM, Civelek Y, Schpero WL, Casalino LP, Zhang M, Pierre R, et al. Long-Term Spending of Accountable Care Organizations in the Medicare Shared Savings Program. JAMA. 2025;333:1897–905. https://doi.org/10.1001/jama.2025.3870.

7. McWilliams JM, Hatfield LA, Landon BE, Chernew ME. Savings or Selection? Initial Spending Reductions in the Medicare Shared Savings Program and Considerations for Reform. The Milbank Quarterly. 2020;98:847–907. https://doi.org/10.1111/1468-0009.12468.

8. Markovitz AA, Hollingsworth JM, Ayanian JZ, Norton EC, Yan PL, Ryan AM. Performance in the Medicare Shared Savings Program After Accounting for Nonrandom Exit: An Instrumental Variable Analysis. Annals of Internal Medicine. 2019;171:27–36. https://doi.org/10.7326/m18-2539.

9. Markovitz AA, Hollingsworth JM, Ayanian JZ, Norton EC, Moloci NM, Yan PL, et al. Risk Adjustment In Medicare ACO Program Deters Coding Increases But May Lead ACOs To Drop High-Risk Beneficiaries. Health Affairs. 2019;38:253–61. https://doi.org/10.1377/hlthaff.2018.05407.

10. Douven R, McGuire TG, McWilliams JM. Avoiding Unintended Incentives In ACO Payment Models. Health Affairs. 2015;34:143–9. https://doi.org/10.1377/hlthaff.2014.0444.

11. Weitzman ML. The “Ratchet Principle” and Performance Incentives. The Bell Journal of Economics. 1980;11:302. https://doi.org/10.2307/3003414.

12. Cappello A. Essays in Industrial Organization and Health Economics. PhD dissertation. Boston College; 2026.

https://www.proquest.com/openview/92b6e8f86c7ba39e2ba879d91063898b/1. Accessed 28 Oct 2026.

13. Rose S, Zaslavsky AM, McWilliams JM. Variation In Accountable Care Organization Spending And Sensitivity To Risk Adjustment: Implications For Benchmarking. Health Affairs. 2016;35:440–8. https://doi.org/10.1377/hlthaff.2015.1026.

14. Lyu PF, Chernew ME, McWilliams JM. Benchmarking Changes And Selective Participation In The Medicare Shared Savings Program. Health Affairs. 2023;42:622–31. https://doi.org/10.1377/hlthaff.2022.01061.

15. Chen AJ, McWilliams JM. How Benchmark Changes Affect Participation in Accountable Care Organizations: Prospects for Voluntary Payment Models. American Journal of Health Economics. 2025;11:38–62. https://doi.org/10.1086/726748.

16. Srivastava A, Shay A, Kaufman SR, Liu X, Maganty A, Oerline MK, et al. Accountable Care Organization Efficiency on Entry and Shared Savings Bonuses. JAMA Network Open. 2026;9:e260166. https://doi.org/10.1001/jamanetworkopen.2026.0166.

17. Bleser WK, Saunders RS, Muhlestein DB, McClellan M. Why Do Accountable Care Organizations Leave The Medicare Shared Savings Program? Health Affairs. 2019;38:794–803. https://doi.org/10.1377/hlthaff.2018.05097.

18. Zhu X, Huang H, MacKinney AC, Ullrich F, Mueller K. Medicare accountable care organization characteristics associated with participation in 2-sided risk. The Journal of Rural Health. 2023;39:302–8. https://doi.org/10.1111/jrh.12672.

19. Ying M, Hirth RA, Yan P, Shay A, Murali S, Ryan A, et al. Changes in Shared Savings Program Participation After Launch of Pathways to Success. Journal of General Internal Medicine. 2023;38:1780–2. https://doi.org/10.1007/s11606-022-07947-x.

20. Ying M, Forman JH, Murali S, Gauntlett LE, Krein SL, Hollenbeck BK, et al. Factors affecting Accountable Care Organizations' decisions to remain in or exit the Medicare Shared Savings Program following Pathways to Success. Health Affairs Scholar. 2024;2:qxad093. https://doi.org/10.1093/haschl/qxad093.

21. Moloci NM, Si Y, Norton EC, Ryan AM, Hollingsworth JM. Predicting Losses from Medicare Shared Savings Program Departures. Journal of General Internal Medicine. 2021;36:2490–1. https://doi.org/10.1007/s11606-020-06424-7.

22. Trombley MJ, McWilliams JM, Fout B, Morefield B. ACO Investment Model Produced Savings, But The Majority Of Participants Exited When Faced With Downside Risk. Health Affairs. 2022;41:138–46. https://doi.org/10.1377/hlthaff.2020.01819.

23. Centers for Medicare & Medicaid Services. Medicare program; Medicare Shared Savings Program; accountable care organizations: Pathways to Success and extreme and uncontrollable circumstances policies for performance year 2017. Final rule. 2018. https://www.federalregister.gov/documents/2018/12/31/2018-27981/medicare-program-medicare-shared-savings-program-accountable-care-organizations-pathways-to-success. Accessed 28 Oct 2026.

24. Centers for Medicare & Medicaid Services. Accountable care organizations: Shared Savings Program enrollment, performance years 2014-2026. 2026. https://data.cms.gov/medicare-shared-savings-program/accountable-care-organizations. Accessed 22 Oct 2026.

25. Smith M, Jensen B, Anderson J, Lockett K. Medicare Shared Savings Program: ACO financial results for 2022. Milliman research report. Milliman; 2024. https://www.milliman.com/-/media/milliman/pdfs/2024-articles/4-2-

24_medicare_shared_savings_program_aco_financial_results_for_2022.ashx. Accessed 28 Oct 2026.

26. Milliman. Annual renewal considerations in the Medicare Shared Savings Program (MSSP). Case study. Milliman; 2025. https://edge.sitecorecloud.io/millimaninc5660-milliman6442-prod27d5-0001/media/Milliman/PDFs/2025-Articles/4-7-25_MSSP-annual-renewal-considerations.pdf. Accessed 28 Oct 2026.

27. Centers for Medicare & Medicaid Services. 42 CFR 425.652, 425.656 and 425.658: Benchmarks, regional adjustment and prior savings adjustment for agreement periods beginning on January 1, 2024, and in subsequent years. 2024. https://www.ecfr.gov/on/2024-01-01/title-42/chapter-IV/subchapter-B/part-425/subpart-G/section-425.652. Accessed 28 Oct 2026.

28. Harrold ST. Why did Medicare ACOs reset their benchmarks early? A pre-registered analysis of early renewal in the Medicare Shared Savings Program. 2026. https://doi.org/10.17605/OSF.IO/C53W9.

29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Medicine. 2007;4:e296. https://doi.org/10.1371/journal.pmed.0040296.

30. Harrold ST. Registered analysis code for the MSSP early-renewal study (analysis/mssp-benchmark-ratchet), commit a4064f5. 2026. https://github.com/stharrold/stharrold.github.io/tree/a4064f592a492d971ce3c27e95afb3050e6bc01d/analysis/mssp-benchmark-ratchet. Accessed 28 Oct 2026.

31. Centers for Medicare & Medicaid Services. Performance year financial and quality results, performance years 2013-2024. 2026. https://data.cms.gov/medicare-shared-savings-program/performance-year-financial-and-quality-results. Accessed 22 Oct 2026.

32. Centers for Medicare & Medicaid Services. County-level aggregate expenditure and risk score data on assignable beneficiaries. 2026. https://data.cms.gov/medicare-shared-savings-program/county-level-aggregate-expenditure-and-risk-score-data-on-assignable-beneficiaries. Accessed 22 Oct 2026.

33. Centers for Medicare & Medicaid Services. Number of accountable care organization assigned beneficiaries by county. 2026. https://data.cms.gov/medicare-shared-savings-program/number-of-accountable-care-organization-assigned-beneficiaries-by-county. Accessed 22 Oct 2026.

34. Harrold ST. Why did Medicare ACOs reset their benchmarks early? A pre-registered analysis of early renewal in the Medicare Shared Savings Program: Project files. 2026. https://doi.org/10.17605/OSF.IO/Z4AN5.

35. Centers for Medicare & Medicaid Services. CMS cell suppression policy. 2020. https://web.archive.org/web/20250516214605/https://www.hhs.gov/guidance/document/cms-cell-suppression-policy. Accessed 28 Oct 2026.

36. Centers for Medicare & Medicaid Services. Medicare Shared Savings Program: Shared savings and losses and assignment methodology specifications, version 10. 2022. https://www.cms.gov/files/document/medicare-shared-savings-program-shared-savings-and-losses-and-assignment-methodology-specifications.pdf-1. Accessed 28 Oct 2026.

37. Centers for Medicare & Medicaid Services. 42 CFR 425.600: Selection of risk model. 2022. https://www.ecfr.gov/on/2022-01-01/title-42/chapter-IV/subchapter-B/part-425/subpart-G/section-425.600. Accessed 30 Oct 2026.

38. Rubin DB. The Bayesian Bootstrap. The Annals of Statistics. 1981;9:130–4. https://doi.org/10.1214/aos/1176345338.

39. Phipson B, Smyth GK. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology. 2010;9:39. https://doi.org/10.2202/1544-6115.1585.

40. Abril-Pla O, Andreani V, Carroll C, Dong L, Fonnesbeck CJ, Kochurov M, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. PeerJ Computer Science. 2023;9:e1516. https://doi.org/10.7717/peerj-cs.1516.

41. Hoffman MD, Gelman A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014;15:1593–623. https://jmlr.org/papers/v15/hoffman14a.html. Accessed 28 Oct 2026.

42. Gelman A. Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine. 2008;27:2865–73. https://doi.org/10.1002/sim.3107.

43. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-normalization, folding, and localization: an improved R-hat for assessing convergence of MCMC. Bayesian Analysis. 2021;16:667–718. https://doi.org/10.1214/20-ba1221.

44. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80:27–38. https://doi.org/10.1093/biomet/80.1.27.

45. Luo JZ. firthmodels 0.8.2: Firth-penalized models in Python. 2026. https://doi.org/10.5281/zenodo.17863280.

804 46. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for
805 logistic and other regression models. The Annals of Applied Statistics. 2008;2:1360–83.
806 https://doi.org/10.1214/08-aoas191.
807 47. Callaway B, Sant'Anna PHC. Difference-in-Differences with multiple time periods. Journal
808 of Econometrics. 2021;225:200–30. https://doi.org/10.1016/j.jeconom.2020.12.001.
809 48. Roth J. Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends.
810 American Economic Review: Insights. 2022;4:305–22. https://doi.org/10.1257/aeri.20210236.
811 49. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses.
812 Social Psychological and Personality Science. 2017;8:355–62.
813 https://doi.org/10.1177/1948550617697177.
814 50. Rambachan A, Roth J. A More Credible Approach to Parallel Trends. Review of Economic
815 Studies. 2023;90:2555–91. https://doi.org/10.1093/restud/rdad018.
816 51. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies
817 for causal effects. Biometrika. 1983;70:41–55. https://doi.org/10.1093/biomet/70.1.41.
818 52. Welch BL. The generalization of "Student's" problem when several different population
819 variances are involved. Biometrika. 1947;34:28–35. https://doi.org/10.1093/biomet/34.1-2.28.
820 53. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
821 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–
822 72. https://doi.org/10.1038/s41592-019-0686-2.
823 54. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators
824 with improved finite sample properties. Journal of Econometrics. 1985;29:305–25.
825 https://doi.org/10.1016/0304-4076(85)90158-7.

826 55. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python.
827 Proceedings of the Python in Science Conference. 2010;92–6. https://doi.org/10.25080/majora-
828 92bf1922-011.
829 56. dbt Labs, Wills J. dbt-core 1.12.6 and dbt-duckdb 1.11.0. 2026. https://github.com/dbt-
830 labs/dbt/tree/v1.12.6. Accessed 30 Oct 2026.
831 57. Raasveldt M, Mühleisen H. DuckDB: an Embeddable Analytical Database. Proceedings of
832 the 2019 International Conference on Management of Data. 2019;1981–4.
833 https://doi.org/10.1145/3299869.3320212.
834 58. Astral Software. uv: An extremely fast Python package and project manager. 2026.
835 https://github.com/astral-sh/uv. Accessed 28 Oct 2026.
836 59. Kumar R, Carroll C, Hartikainen A, Martin O. ArviZ a unified library for exploratory
837 analysis of Bayesian models in Python. Journal of Open Source Software. 2019;4:1143.
838 https://doi.org/10.21105/joss.01143.
839 60. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al.
840 Array programming with NumPy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-
841 020-2649-2.
842 61. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M
843 (Magnitude) Errors. Perspectives on Psychological Science. 2014;9:641–51.
844 https://doi.org/10.1177/1745691614551642.
845 62. Einav L, Finkelstein A, Ji Y, Mahoney N. Voluntary Regulation: Evidence from Medicare
846 Payment Reform. The Quarterly Journal of Economics. 2021;137:565–618.
847 https://doi.org/10.1093/qje/qjab035.

63. American Hospital Association, National Association of ACOs, et al. Letter to CMS: Request to provide MSSP ACOs an option to use pre-pandemic years to set benchmarks. 2021. https://www.aha.org/lettercomment/2021-10-14-aha-others-cms-re-request-provide-mssp-acos-option-use-pre-pandemic-years. Accessed 28 Oct 2026.

64. US Department of Health and Human Services. 45 CFR 46.102: Definitions for purposes of this policy. 2026. https://www.ecfr.gov/on/2026-10-01/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102#p-46.102%28e%29%281%29. Accessed 28 Oct 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2.docx