Stars misaligned: Medicare star ratings and the exact optimum of their clustering criterion

Matthew D. Schwartz*
Department of Physics, Harvard University

Draft, September 28, 2026

Abstract

The federal government provides quality scores which are often binned into categories of one to five stars. Six rating systems of the Centers for Medicare & Medicaid Services (CMS) assign stars this way. One rates Medicare Advantage plans, whose stars determine bonus payments, one rates overall hospital quality, and four rate patient experience at hospitals, home health agencies, hospices, and dialysis facilities. All six rest on the same criterion, that the total squared distance of the scores from their bin centers be as small as possible. In practice the bins are found by k-means or Ward's hierarchical clustering, fast algorithms not guaranteed to reach that minimum. The scores being binned are one-dimensional, however, and in one dimension the optimal bins can be computed exactly. We do not judge CMS's intention or interpret the law: we compute the stated minimum, check whether the published numbers attain it, and, for Medicare Advantage, translate the difference into dollars. In Medicare Advantage, CMS's clustering step misses the minimum in 104 of 123 cut-point computations over three rating years. Carried through the rest of the calculation, the optimal bins move 34 published contract-year ratings across the bonus threshold, some up and some down, and roughly a billion dollars in gross bonus payments depend on which bins are used. Rerunning CMS's hospital code reproduces every published star, yet about one hospital in fifteen now displays fewer stars than the optimal bins assign, and in earlier years some displayed more. Two of the four survey programs publish the underlying scores, and on those scores their bins also miss the minimum for most measures examined; the other two publish only stars and cannot be checked.

1 Introduction

A star rating condenses a set of quality scores into a single number, from one star to five, that a government agency computes and millions of people read. The Centers for Medicare & Medicaid Services (CMS) publishes several such ratings, six of which we examine here. Each Medicare Advantage (MA) contract, the unit in which Medicare's private health plans are rated, receives an overall star rating on which its bonus payments depend. A contract rated 4.0 stars or above is paid a bonus under the Quality Bonus Program (QBP), and these bonuses are estimated to total $13.4 billion for 2026 [1]. Each rated hospital receives an Overall Hospital Quality Star Rating, which patients and referring physicians may consult when choosing a hospital. Four

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

patient-experience survey programs assign further stars, shown on Medicare's comparison sites, to hospitals, home health agencies, hospices, and dialysis facilities. The Medicare Advantage rating affects payments directly, whereas the other five inform patients' choices.

All six rating programs state the same mathematical criterion: place the quality scores on a line and group them into bins so that the total squared distance of the scores from the centers of their bins is as small as possible. For Medicare Advantage, the regulations state this criterion and name the algorithm that sets the cut points between star levels [2, 3], and the agency's annual Technical Notes restate the criterion [4–6]. The four survey programs repeat it in their own technical notes. For the hospital ratings, the methodology report prescribes k-means clustering, whose objective is this criterion [7], and CMS publishes the code [8, 9]. That code starts from a fixed initial grouping and moves scores between groups as long as each move improves the grouping. The other five programs use Ward's hierarchical clustering, which begins with every score in its own group and repeatedly merges the two groups whose merger adds least to the total squared distance. In Medicare Advantage, CMS combines the clustering with a resampling step. Because neither algorithm searches over all groupings, either can miss the best one. Indeed, for data of arbitrary dimension, finding the best grouping is NP-hard [10], which means that no known fast algorithm is guaranteed to find it. For scores on a line, however, Fisher's 1958 algorithm provably finds the best grouping [11] and is available in standard statistical software [12].

Earlier work takes the published Medicare Advantage computation as given. One consultancy, Milliman, replicates it and runs policy counterfactuals, reporting 76 contracts affected in one litigated scenario [13]. Academic studies use the published stars as data: Darden and McCarthy on the bonus threshold's effects on payments and enrollment [14], Anderson and Meiselbach on year-to-year rating instability [15], and Vatter on the design of scoring rules [16]. The ratings are already the subject of litigation, in which we take no side. A 2024 ruling in a suit brought by SCAN Health Plan led CMS to recalculate the 2024 ratings nationwide [13]. Alignment Healthcare lost a broader challenge to the rating methodology in June 2025 and on appeal in July 2026, and SCAN and Alignment filed new suits in July 2026 (Nos. 1:26-cv-02391 and 1:26-cv-02441, D.D.C.).

The closest precedent for examining the clustering step itself concerns the hospital ratings. For Hospital Compare in 2017, Bayles showed that CMS's k-means implementation stopped short of convergence, which changed hundreds of hospital ratings [17]. Huang reimplemented that code and estimated the affected share at about a quarter [18]. Barclay, Dixon-Woods, and Lyratzopoulos varied earlier steps of the same computation for the 2021 ratings, the standardization of the measures, their grouping into domains, and the domain weights, holding the clustering step fixed, and found that about half of 3,339 hospitals changed star under reasonable alternatives [19]. That work, on a program without bonus payments, did not compare the resulting grouping with the optimum of the stated criterion. To our knowledge no prior work has tested the published cut points against the criterion that the regulations and Technical Notes state for them; we searched the academic literature, the actuarial consulting publications, the litigation dockets, and public code repositories.

In this paper we compute the minimum of the stated criterion exactly, in each program whose inputs CMS publishes, and compare the grouping that attains it (the exact optimum) with the published stars. For the hospital ratings and Medicare Advantage we also recompute CMS's clustering itself. We rerun the published hospital code, and we rebuild the Medicare Advantage computation stage by stage from the Technical Notes, identifying the rule for its one

randomized stage (the assignment of contracts to resampling groups) by running the documented SAS procedure ourselves. Two inputs to that stage are unpublished, namely the order in which the contracts were read and, in the earlier star years, the random-number generator used. For the hospital and dialysis surveys we compare the published groupings with the exact optimum on the published scores, and rerun Ward's method on those scores solely to determine whether they are the inputs CMS clustered. For the home health and hospice surveys CMS publishes the stars but not the scores behind them, so their groupings cannot be tested against the criterion.

In each program whose inputs are public, we find published groupings that miss the stated minimum. A rerun of the CMS hospital code reproduces every published star. Even so, about one in fifteen hospitals with a displayed star in 2026 has fewer stars than the exact optimum assigns, and in earlier rating years some hospitals instead have more. For Medicare Advantage, on clustering samples reconstructed from the public files, the agency's clustering step misses the minimum in 104 of 123 cut-point computations for star years 2024–2026. When we carry the exact optimum through the remaining published stages, 34 published contract-year ratings change bonus status, some crossing the bonus threshold upward and others downward. We estimate the gross bonus payments that depend on these changes at roughly \$0.9 to \$1.4 billion over payment years 2025 through 2027. Unlike the hospital code, CMS's own Medicare Advantage computation cannot be rerun from the published files, which omit the exact set of scores that entered each computation and the order in which its random step read them. For some contract-years the bonus status it assigns depends on that order, and the published files therefore leave their status undetermined. The published groupings of the hospital and dialysis survey scores also miss the minimum on most of the measures examined.

In Section 2 we state the criterion and the algorithms, then describe how we recompute the agency's clustering and how we convert changes in bonus status into dollars. We present the hospital, Medicare Advantage, and patient-experience results in Sections 3–5 and discuss their regulatory context and limitations in Section 6.

2 Star assignment as one-dimensional clustering

The six programs assign stars by dividing a line of scores into bins and state one common criterion for that division. They differ in what is scored, in the algorithm that places the bins, in how much of the computation is public, and in the payment consequences.

2.1 The six programs and their public files

The Overall Hospital Quality Star Rating collapses each hospital's reported measures into one summary score and divides the summary scores, within each peer group, into five bins [7]. A hospital's peer group (3, 4, or 5) is the number of the method's five measure groups in which it reports enough measures to be scored. The methodology report, the code (SAS for 2021–2025, R for 2026), the input files, and the published stars are public [8, 9, 20].

For Medicare Advantage, CMS bins contract scores into stars separately for each quality measure. At an overall rating of 4.0 stars or above, a contract is paid against a benchmark (its county spending target) raised by five percent per enrollee in the following payment year. A contract at 3.5 gets no increase. Each measure's stars depend on cut points, the thresholds that determine whether a contract's 87% screening rate is worth three stars or four. CMS recomputes the cut points each year from the distribution of contract scores. We call the calculation of one measure's cut points in one star year a computation, which for Part D (drug plan) measures is

also specific to one organization type (MA-PD, with medical coverage, or PDP, a stand-alone Part D plan). Star years 2024–2026 contain 123 computations (40, 40, and 43 by year). A contract-year is one contract's rating in one star year. The agency publishes scores, measure stars, cut points, and ratings in annual data tables with Technical Notes [4–6]. It does not publish which scores entered each computation or the assignment of contracts to resampling groups.

The four patient-experience programs, all built on surveys of the CAHPS (Consumer Assessment of Healthcare Providers and Systems) family, cluster survey scores by measure. CMS publishes those scores for the hospital (HCAHPS) and dialysis (ICH CAHPS) surveys, but only stars for the home health and hospice surveys.

2.2 The stated criterion

Each program's rules prescribe an algorithm and also state a criterion for the bins: the sum of squared distances of the scores from their bin centers should be as small as possible. The regulations call this sum the within-cluster sum of squares (SSQ). For scores $x_1 \le \cdots \le x_n$ grouped into bins $B_1, \ldots, B_k$ with bin means $\bar{x}_1, \ldots, \bar{x}_k$,

$$\mathrm{SSQ} \;=\; \sum_{b=1}^{k} \sum_{i \in B_b} \left(x_i - \bar{x}_b\right)^2. \tag{1}$$

The total in Eq. (1) has a minimum over all groupings of a given set of scores, and we test whether each published grouping attains that minimum.

For Medicare Advantage two provisions of the regulations bear on the criterion. The definition of clustering, at 42 CFR 422.162(a) [2], states the objective:

> "The Star Ratings levels are assigned to the clusters that minimize the within-cluster sum of squares."

The calculation rule, at 42 CFR 422.166(a)(2)(i) [3], prescribes the method:

> "The method maximizes differences across the star categories and minimizes the differences within star categories using mean resampling with the hierarchal clustering of the current year's data."

The Technical Notes [4–6] repeat the objective sentence verbatim in all three years. The two provisions point in different directions, since hierarchical clustering with mean resampling does not in general produce the minimum that §422.162(a) names. Which provision controls, or how they are to be reconciled, is a question of regulatory interpretation for the courts, and we do not address it.

For the hospital ratings the criterion is stated through the algorithm: the methodology report and code prescribe k-means within each peer group [7], and the k-means objective is Eq. (1). The technical notes of the four patient-experience programs prescribe Ward's hierarchical clustering and describe it as minimizing the same total.

2.3 The prescribed algorithms and the exact optimum

The hospital code uses k-means, in the SAS procedure FASTCLUS [21] for 2021–2025 and in R for 2026. From initial bin centers it moves a bin boundary whenever a single move lowers the

total, and it stops when no such move remains. Thus it ends at a local optimum (a grouping no single move improves), and which one it reaches depends on the starting centers. The CMS code fixes these by a deterministic rule, and the hospital computation involves no randomness.

For Medicare Advantage and the four patient-experience programs, CMS uses Ward's hierarchical clustering, which starts from single-score clusters and repeatedly joins the two clusters whose merger increases the total least. Merging clusters a and b, with n_a and n_b scores and means $\bar{x}_a$ and $\bar{x}_b$, increases Eq. (1) by the amount [22]

$$\Delta(a,b) \;=\; \frac{n_a\, n_b}{n_a + n_b}\,\big(\bar{x}_a - \bar{x}_b\big)^2 \tag{2}$$

and merging stops when k clusters remain.

Neither Ward's method nor k-means searches over all groupings. Ward's method never undoes a merge, even when the optimal grouping separates the scores that the merge joined. The k-means iteration can stop at a local optimum that is not global, since moving several scores at once could still lower the total. For example, on fifteen synthetic scores the hospital procedure returns a grouping with total 1.498, against 0.132 for the optimal grouping. In particular, nine of the fifteen scores sit one bin lower than in the optimal grouping, and when the sign of every score is reversed the procedure places the same nine too high instead.

In general dimension, minimizing Eq. (1) is NP-hard [10]. On a line, however, the scores can be sorted, and an optimal grouping consists of consecutive runs of sorted scores. Fisher [11] showed in 1958 how to search the choices of $k-1$ cut points among the $n-1$ gaps between adjacent scores without repeating work, by dynamic programming. With $D_m(i)$ the smallest total for the first i sorted scores in m bins and $\mathrm{SSQ}(x_j,\dots,x_i)$ that of a single bin,

$$\begin{aligned} D_m(i) \;&=\; \min_{m \le j \le i} \big\{\, D_{m-1}(j-1) + \mathrm{SSQ}(x_j,\dots,x_i) \,\big\}, \\ D_1(i) \;&=\; \mathrm{SSQ}(x_1,\dots,x_i), \qquad \min_{\text{groupings}} \mathrm{SSQ} \;=\; D_k(n). \end{aligned} \tag{3}$$

An enumeration variant of the algorithm lists every grouping attaining the minimum, which settles whether the optimum is unique.

We run the dynamic program, and every comparison in the paper, in exact rational arithmetic: each decimal CMS prints becomes a ratio of integers, so that no conclusion depends on floating-point behavior. Where the agency's computation used an internal precision that the public files do not show, we propagate the interval of values consistent with the printed decimals and state a result only if it holds throughout the interval.

Both k-means and Ward's method are deterministic, and a rerun on the same inputs returns the same grouping. Reproducing a published computation therefore establishes only that the same program was run on the same inputs, not that the grouping it returns attains the stated minimum.

2.4 Recomputing CMS's clustering

We compare the exact optimum with the grouping that CMS's procedure produced. That grouping must be recomputed step for step from the public files, because an approximate reimplementation cannot separate a real discrepancy from its own approximation error. For hospitals, rerunning the CMS code reproduces every published star (Section 3). Hence any difference between a published star and the star the optimum assigns comes from the published grouping itself.

For each Medicare Advantage computation the Notes prescribe the following steps [4–6]. Scores beyond the Tukey outer fence (a quartile-based outlier rule) are trimmed, and the SAS procedure `SURVEYSELECT` assigns the rest at random to ten equal-sized groups. Ten times, one group is set aside and the rest are clustered by Ward's method. The ten values of each cut point are then averaged; this "mean resampling" of the regulation is a delete-one-group jackknife. CMS then applies a guardrail (the cap that limits how far a cut point may move from its prior-year value) and rounds the cut points to display precision. Measure stars are assigned and aggregated to Part C, Part D, and overall ratings.

For the random assignment to resampling groups, the Notes print the SAS call and explain, in all three star years, that its `seed=8675309` option "allows for future replication of the randomization process." They also state that "It is not possible to replicate CMS's calculations exactly," citing third-party rounding and contracts excluded from the public files. We identified the assignment rule, which neither the SAS manual nor the Notes spell out, by running the step in SAS (Appendix C). The assignment still depends on two unpublished inputs: the order in which the procedure reads the contracts (the input order) and, for 2024 and 2025, which of SAS's two random-number generators ran.

In our recomputation we take every score from the agency's files and reconstruct the clustering sample, the set of scores that enters each computation. The Notes exclude scores that are "missing, flagged, excluded by disaster rules or voluntary," which we read to include the elective Cost contracts under Social Security Act section 1876. In addition, CMS suppresses the scores of some small-enrollment contracts in the public files. The Notes print the two bounds of the Tukey outer fence for every computation. Because these bounds are functions of the sample's quartiles [23], any candidate sample must reproduce them. Under the documented rules alone, with no fitted quantity, the reconstructed sample reproduces both bounds for 82 of the 123 computations. For the rest we fit score profiles for unpublished contracts, adding 56 contracts and removing 10 over the three years, after which every computation matches. A fitted profile is not an identified contract, and we make no claim that any particular real contract carried these values. We use this documented-rules sample and two variants of it, and report where a result differs among the three. One variant also excludes employer-only contracts (contracts that offer only employer-group plans); the other differs only in how one removal forced by the bounds is carried out.

On each reconstructed sample we run the clustering step both by CMS's Ward method, in the double precision of SAS and with its documented tie rule [22], and by dynamic programming in exact arithmetic. We treat the agency's own resampled computation separately: wherever a result depends on an unpublished input, we report its range over sampled values.

2.5 Conventions for bonus status and dollar estimates

The rulebooks state an objective and prescribe a procedure, and where the two disagree we do not attempt to judge CMS's intention or interpret the law. We compute the optimum of the stated objective, check whether the published numbers attain it, and translate the difference into dollars. If one reads the procedure as the rule then this paper measures how far it lands from the objective; if one reads the objective as the rule then we measure the payments that may have been misdirected. We provide the tools and the numbers, but not the interpretation.

To determine bonus status we take the Ward cut points and the optimal cut points to the overall rating through the same published stages that follow clustering (resampling, the guardrail, rounding, and aggregation). The one exception is on the criterion side, where the

resampling is not run and the stated minimum is computed once on the full trimmed sample. The translation into bonus status and dollars uses seven conventions. (1) Improvement measures and the CAHPS survey measures inside the Medicare Advantage rating keep their published stars, because improvement scores are unpublished and the stars of these CAHPS measures are assigned without clustering. The counts of changed overall ratings are therefore conditional on those published stars: a different improvement or CAHPS star could remove a change we count as well as add one we do not, and Appendix C.6 bounds the improvement part in both directions. (2) We apply the guardrail against the published prior-year cut points, matched by measure name and part. (3) Each prior-year cut point, public only to display precision, is treated as the interval of values that round to its displayed value. (4) The clustering sample is the documented-rules reconstruction (adding the employer-only exclusion yields the same bonus-status changes). (5) We score each star year at one data vintage (2024 recalculated, 2025 December-updated, 2026 current, with the July 22, 2026 summary file as the baseline for payment year 2027 bonuses). That file, posted August 17, 2026, contains the per-contract results of the agency's June 17, 2026 upward-only recalculation. In it, 50 overall ratings differ from those originally published: 48 are higher, and the other two changed between a numeric rating and "Not enough data available." (6) On the criterion side we apply the agency's extreme-and-uncontrollable-circumstances hold-harmless, which keeps an affected contract's measure star at or above its prior-year published star. Its separate rule for contracts affected in two consecutive years, evaluated from the earlier public files, changes none of the 34 valued contract-years of Section 4.2 (Appendix C.6). (7) We value a contract-year in dollars only when it is bonus-relevant (not a stand-alone Part D contract), has a published overall rating, is not a Cost contract, and changes bonus status against both baselines. The two baselines are the published overall rating and the rating recomputed from published measure stars under the documented rules (Appendix C).

We call a change in a contract-year's bonus status, in either direction, a flip. All dollar figures below and in Appendix B are estimates, a per-enrollee value applied to the set of flips, with enrollment from the agency's monthly enrollment report. Bonus status is valued at \$400–600 per enrollee-year [1, 24], midpoint \$500, and each dollar total is a gross band (upward and downward changes are not netted). The band uses no model and equals the valued contract-years' enrollment times that range. The \$400–600 band is a reference value, not a measurement. Neither of its two sources states a value per bonus-plan enrollee, and the ratios they imply are roughly \$560 and \$575. The published averages by plan segment, \$318 to \$466, are diluted over all enrollees rather than bonus-plan enrollees. And \$600 is not a ceiling (the MedPAC figures imply roughly \$620) nor \$400 a floor (benchmark caps under the Affordable Care Act can pull individual contracts lower).

No error is attributed to anyone; each divergence is between a published grouping and the grouping that minimizes the program's stated criterion on the same scores. We count affected hospitals and identify contracts only by CMS contract number, and no plan or hospital is named. No dollar figure is attached to a hospital or patient-experience star.

3 Overall Hospital Quality Star Ratings

For the Overall Hospital Quality Star Rating, CMS publishes the code, the input files, and, on its Care Compare site, the rating itself. Because no step of the computation is randomized, we can recompute the published stars in full and compare them directly with the optimum of the

Rating year	Reproduced / mismatched	Hospitals rated	Differ from optimum	Up / down
2021	3,339 / 0	3,355	99 (3.0%)	40 / 59
2022	3,094 / 0	3,121	153 (4.9%)	107 / 46
2023	3,061 / 0	3,076	370 (12.0%)	136 / 234
2024	2,834 / 0	2,847	273 (9.6%)	15 / 258
2025	2,872 / 0	2,891	38 (1.3%)	33 / 5
2026, pre-cap	3,182 / 0	3,203	213 (6.7%)	213 / 0
2026, post-cap		3,203	208 (6.5%)	208 / 0
2026, display-visible		3,182	205 (6.4%)	205 / 0
Six years		18,493	1,146 pre-cap; 1,141 post-cap	

Table 1: The CMS code reproduces every published hospital star in all six years, yet each year the exact optimum of the stated objective assigns some hospitals a different star, in a direction that varies by year. Reproduced / mismatched: jointly rated hospitals whose published star the code does or does not return, against that year's Care Compare release file. Hospitals rated: those the code assigns a star. Differ from optimum: those the exact optimum rates differently; up, the optimum lies above the published star. The three 2026 rows count the differences before the safety cap, after it, and among hospitals with a displayed star. Code and inputs from public mirrors of CMS's releases (see text).

k-means objective.

3.1 Reproducing the published stars

The code that implements the methodology report [7] consists of SAS packages for the 2021–2025 rating years [8] and an R package for 2026 [9]. We ran the R package without modification. Its output matches the published 2026 Care Compare star [20] for all 3,182 jointly rated hospitals (those with a star in both the output and the published file). The three peer groups of that year contain 177, 749, and 2,277 hospitals, the sizes given in the methodology report. The SAS packages call one procedure with no public implementation (the k-means routine FASTCLUS), which we reimplemented from its documentation [21] (Appendix C). With that reimplementation the packages reproduce every published star of each year from 2021 through 2025 (Table 1). In each comparison we use the Care Compare file that contains that year's release. Compared instead with the April 2025 file, which still carried the previous release, the 2025 recomputation differs on 1,245 hospitals for that reason alone.

CMS's hosting site for the code and its input files returned an access error to our requests. Instead, we took both from two public GitHub repositories that mirror the agency's releases. One (klocey/stars-data-builder) covers all six years, and the other (Kentucky-Open-Science/CMS-Star-Rankings) provides a second copy of the 2025 package and inputs. The two copies of the 2025 inputs are identical, and for the other years one mirror is the only source. This qualification, that the code and its inputs come from the mirrors and not from CMS directly, applies to all results in this section that rest on them. The published stars in every comparison come directly from the Care Compare files posted by CMS.

3.2 Comparison with the exact optimum

We then compare the reproduced grouping, hospital by hospital, with the grouping of the same scores that minimizes Eq. (1) in each peer group. In 2026 the exact optimum assigns a different

star to 205 of 3,182 displayed hospitals (6.4%), in every case a higher one. The count depends on the stage of the computation at which it is taken: at the clustering step the optimum differs from the published assignment for 213 of the 3,203 hospitals the code rates (6.7%, all of them upward). Each peer group's optimum is unique, but the criterion barely distinguishes it: most of the 213 regain their published star under some grouping closer to the optimum than the published one (Appendix C.6).

After the clustering step the methodology applies a safety cap [7]. A hospital that falls in the lowest quartile of the safety-of-care measure group and reports at least three safety measures receives at most four stars. The cap returns five of the 213 to their published star and creates no new difference, which leaves 208 (6.5%). All five are hospitals in peer group 5 that the optimum places at five stars and the cap holds at four. Conversely, none of the 15 hospitals capped in the published computation is among the 213. Three of the 208 have no star in the published 2026 file (two suppressed by footnote, one absent), and the remaining 205 are displayed hospitals.

Figure 1 shows the 2026 comparison for peer group 3 on its score line; the other two peer groups, where the remaining differences occur, are in Figure A1. In both figures the hospitals that the two groupings rate differently are those between a published cut point and its optimal counterpart. Over the six rating years the optimum differs from the published grouping in 1,146 of the 18,493 stars issued (Table 1). The six-year total includes the 2026 differences as they stand before the safety cap and becomes 1,141 if they are counted after it. Across all six years, 12 of the differences belong to hospitals whose star was not shown in that year's file.

We find a unique optimum, with no score tied at a cut point, in each of the eighteen peer-group computations of the six years. The set of hospitals the code rates is consistent with the published file in every year. For 2026 the 3,203 rated hospitals comprise 3,182 with a published star, 14 whose star CMS suppresses with a footnote, and 7 absent from the published file. Suppression, which affects 1 to 14 hospitals a year, is applied for display after the computation and does not enter it.

3.3 Mechanism and direction of the differences

The star differences arise as in the fifteen-score example of Section 2: the published assignment is the local optimum of Eq. (1) that the k-means iteration reaches from the code's deterministic starting centers. The 2026 iteration stops after two steps in each peer group, and the code's second k-means pass then returns the same centers and changes no assignment. Thus the computation converges (as does the FASTCLUS iteration of the earlier packages), and the SSQ of the converged grouping still exceeds the minimum. In the 2017 ratings examined by Bayles and Huang [17, 18], by contrast, the iteration stopped before it converged. When we start the 2026 iteration from the optimal centers in place of the code's, it remains at the optimum. Moreover, all twelve published cut points of 2026, four on each of the three peer-group score lines, lie above the optimal ones.

To examine what determines the direction of the differences, we reverse the sign of every score and repeat the 2026 comparison; the 55 and 68 differences in peer groups 3 and 4 then all become downward. Among 40 synthetic data sets that we drew with skew of either sign, the k-means result was already optimal in 7. In each of the other 33, every difference was upward when the set was left-skewed and downward when it was right-skewed.

In the real data the differences within one peer group and year (a group-year) mostly share one sign. When we count each year once, with 2026 taken before the safety cap, every difference has the same sign in 16 of 18 group-years (19 of 21 with 2026 taken both before and after the

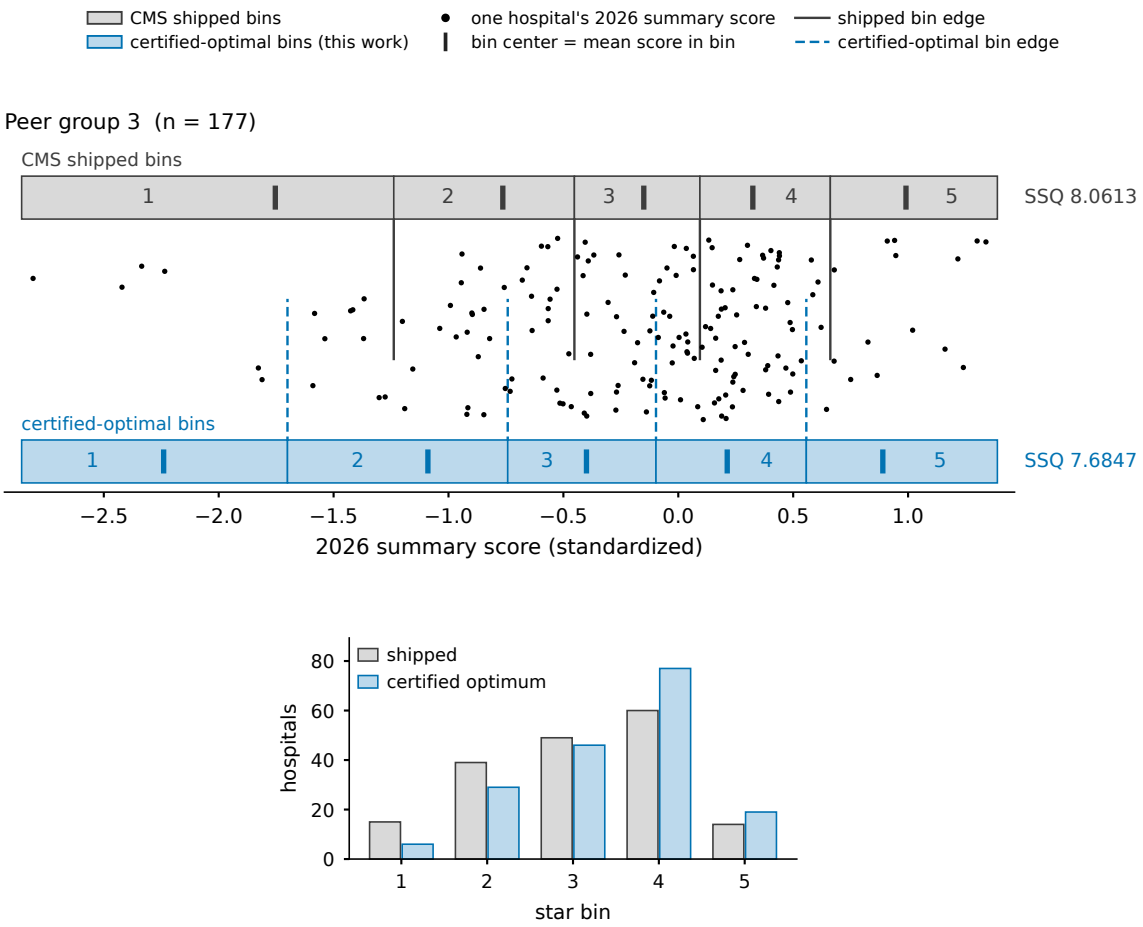

Figure 1: In 2026 all twelve published cut points lie above their optimal counterparts, and the 213 hospitals whose summary scores fall between a published cut point and its optimal counterpart receive one star fewer than the exact optimum assigns (208 of 3,203 after the safety cap; 205 of 3,182 among displayed hospitals). Peer group 3 is shown (177 hospitals, 55 between paired cut points); peer groups 4 and 5 are in Figure A1. Dots, hospitals at their summary scores (vertical jitter has no meaning); upper and lower rulers, the published and optimal star bins, with solid and dashed lines marking their edges; thick ticks, bin centers (mean score in the bin); right, the SSQ of each grouping; beneath, bin populations under both.

cap). The two mixed group-years are 2021 peer group 3 (8 up and 59 down) and 2023 peer group 4 (15 up and 234 down). The latter also shows the largest divergence of the six years: the optimum and the published grouping disagree on 249 of the 462 hospitals in that group (54%).

All eighteen group-year score distributions are left-skewed, and in the synthetic data sets with left skew every difference was upward. In three group-years of the real data, however, every difference is downward, and so the skew alone does not determine the direction and no cross-year sign rule is claimed. That all 2026 differences are upward is a measured 2026-specific fact, not a general claim. No payment is attached to the overall hospital star, but patients and referring physicians may rely on the grade shown on Care Compare when they choose a hospital. For every displayed hospital that the two groupings rate differently in 2026, Care Compare shows

one star fewer than the optimum of the stated criterion assigns.

4 Medicare Advantage star ratings

In Medicare Advantage a contract's overall star rating determines its bonus payments [1, 3], and less of the computation is public than for the hospital ratings. We first compare CMS's clustering step with the exact optimum and recompute each contract's bonus status from the optimal cut points. Neither analysis uses the agency's resampling, whose result depends on the unpublished order of the input file. We then rerun the resampled computation itself at a sample of input orders.

4.1 Optimality gaps at the clustering step

For each of the 123 cut-point computations of star years 2024–2026 we assemble the clustering sample and apply the documented exclusions and the Tukey outer-fence trim [4–6]. On the identical trimmed scores we then compute the Ward grouping that the published method prescribes, with SAS's documented tie rule [22], and the grouping that attains the minimum of Eq. (1). The comparison is made before resampling, guardrails, and rounding, so it isolates the single question of whether the clustering step as CMS runs it attains the minimum the regulations and Technical Notes state.

On the documented-rules reconstruction of the clustering samples, we find that the Ward grouping's sum of squares is strictly above the stated minimum in 104 of the 123 computations and equals it in 16 (Figure 2 and Table 2). The other 3 computations are degenerate. In the two 2026 computations for Part D measure D07 every trimmed score is identical. In the 2025 D07 computation for stand-alone plans the 24 trimmed scores fill only four bins, too few for a five-cluster grouping.

On the two alternative reconstructions of the clustering sample (Section 2), the number of computations that miss the minimum is 105 when employer-only contracts are also excluded (Appendix C) and 106 when one forced removal is carried out differently. Every computation that misses the minimum on the first reconstruction also does so on the other two. When instead every fitted profile and removal is dropped, 103 computations miss the minimum (Appendix C.6). Moreover, in all 120 non-degenerate computations the optimal grouping is unique, and hence wherever the Ward step misses the minimum, one specific grouping attains it. When candidate merges tie, Ward's method itself depends slightly on the input order, whereas the exact optimum is the same for every order. For eight input orders of the same samples, the number of computations in which the Ward grouping is above the minimum ranges from 103 to 106, with four small stand-alone Part D computations changing status.

Across the 104 computations that miss the minimum, the gap in the objective spans roughly seven orders of magnitude, from below 2×10^{-4} to above 1.8×10^{3} in sum-of-squares units of each measure's score scale. Half of the gaps exceed 118, and the largest, 1,813.011, is on the 2024 Transitions of Care measure. Rescoring the trimmed samples against the optimal cut points instead of the Ward cut points changes 9,809 of the 47,752 contract-measure stars the 123 computations assign, about one in five (9,835 with the employer-only exclusion).

For star year 2023 CMS used different documented rules. Neither released version of that year's Technical Notes prescribes the Tukey outer-fence deletion, and 2023 is also the first year in which the guardrail bound cut-point movement. Because no fence is prescribed, there are no published fence values for 2023, whereas for 2024–2026 those values constrain the sample

Table 2: Medicare Advantage, star years 2024–2026: the clustering step, bonus status under the full-sample optimum, and CMS's own computation at sampled input orders. (a) The Ward grouping misses the stated minimum in most computations of every star year, and the minimizing grouping is unique wherever one exists. (b) With the later published stages applied, the full-sample optimum changes the bonus status of 60 contract-years, some upward and some downward; 34 of them are valued in the gross dollar estimate. (c) CMS's own resampled computation, rerun at 28 sampled input orders, leaves the bonus status of 118 contract-years dependent on the unpublished order. Per-year cells are given where a count is resolved by star year; the bonus for a star year is paid in the following payment year (2025, 2026, and 2027).

	Star year			
	2024	2025	2026	All three years
(a) The clustering step on the full trimmed sample[a]				
Cut-point computations	40	40	43	123
Ward grouping above the stated minimum	36	33	35	104 (16 attain it; 3 degenerate)
range over three reconstructions of the clustering sample				104–106
Minimizing grouping unique				120 of 120 non-degenerate
Largest gap, sum-of-squares units				1,813.011 (2024, Transitions of Care)
Contract-measure stars changed on rescoring against the optimal cut points				9,809 of 47,752
(b) Bonus status under the full-sample optimum, payment years 2025–2027[b]				
Contract-years whose bonus status changes				60
valued in the dollar estimate	11	15	8	34 (16 up, 18 down; 31 contracts)
counted without a dollar value				24
endpoint-conditional candidates, excluded				2
Enrollee-payment-years on the valued changes				2,291,034 (up 260,506; down 2,030,528)
Gross bonus payments at midpoint, $ million	$858	$206	$82	$1,146 (up $130.3; down $1,015.3)
band at $400–600 per enrollee-year, $ million				$916–1,375 (up $104–157; down $812–1,219)
Changes involving a guardrail-clamped measure				33 (18 up, 15 down)
changes absent without the guardrail				18, all up
MA-PD contract-years in the published ratings that cross 4.0 if one improvement-measure star changes	18	28	23	69 (57 with a published overall rating)
(c) CMS's resampled computation at 28 sampled input orders[c]				
Computations reproducing all four published cut points at an exhibited input order				50 of 114
Contract-years touching the bonus line under at least one sampled order				129 (98)
on both sides of the line under different orders				118 (92)
changed under every sampled order				11 (6)
Expected gross payments at stake per draw, $ million: Ward's method inside the groups				$787.2 ($629.7 to $944.6); 87 contract-years vary
exact optimum inside the groups				$98.4 ($78.7 to $118.1); 23 vary, 16 of them among the 87
Reconstructions of the group assignment from the published documents that match the published thresholds				none of 102 (best candidate, 21 of 77 cells)

[a] Clustering samples rebuilt from the documented rules with fitted score profiles; the two groupings are compared before resampling, the guardrail, and rounding. Star year 2023 (different documented rules; samples from the documented rules only): 36 of 38 computations miss the minimum, the optimum is unique in all 38, the largest gap is 1,181.666, and 2,930 contract-measure stars change on rescoring.

[b] Under the seven conventions of Section 2; dollar figures are gross estimates. The 34 are among the 1,724 published contract-year ratings of star years 2024–2026 (Table 3). The 24 counted without a dollar value: 10 stand-alone Part D, 2 Cost, 8 with no published overall rating, 4 undecidable under the June 17, 2026 recalculation. The 2 excluded candidates would add $8.9 million at midpoint. Of the 33 changes involving a clamped measure, 20 are valued (8 up, 12 down); the 18 absent without the guardrail include eleven of the 16 valued upward changes.

[c] The 28 orders are a sample (four conventional orderings and twenty-four random permutations), not an exhaustive set; counts in parentheses are contract-years with no other caveat. Of the 129, 17 have no published overall rating, 2 are Cost contracts, and 12 are stand-alone Part D.

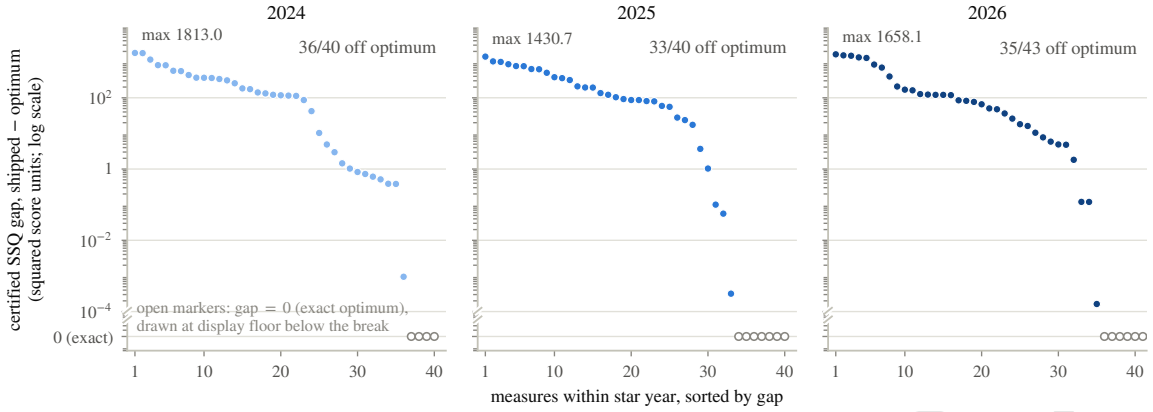

Figure 2: The sum of squares of CMS's Ward grouping is above the stated minimum in 104 of 123 cut-point computations for star years 2024–2026 (2024: 36/40; 2025: 33/40; 2026: 35/43), and the minimizing grouping is unique in all 120 non-degenerate computations. Clustering samples are reconstructed from the documented rules with fitted score profiles. Each panel plots, for one star year, the sum of squares of the Ward grouping (documented tie rule) minus the minimum, in squared units of each measure's published score scale, with measures sorted by gap; the axis is logarithmic because the gaps span seven orders of magnitude, and the largest is 1,813.011 (2024, Transitions of Care). Open markers below the axis break are the 19 computations with gap exactly zero (the 16 that attain the minimum and the three degenerate computations), drawn at a display floor. Gaps are exact rationals plotted as decimals.

reconstruction. The 2023 rosters can therefore only be rebuilt from the documented rules, with no check against a published quantity. On the samples rebuilt in this way, the Ward grouping misses the stated minimum in 36 of the 38 clusterable computations of that year, and the optimal grouping is unique in all 38. Thus CMS's clustering step misses the stated minimum in all four star years examined and under both sets of documented rules.

4.2 Effect on bonus status and payments

A difference in cut points affects payment only if it changes a contract's overall rating. Between the clustering step and the overall rating, the published method applies resampling, the guardrail, display rounding, and aggregation. To measure the effect on payment we apply the later stages (guardrail, rounding, and aggregation) to the full-sample optimal cut points under the conventions of Section 2 and compare the resulting ratings with the stars CMS assigned. We find that the optimal cut points change 5,800 contract-measure stars of the non-degenerate computations relative to the published stars. This count is smaller than the 9,809 of the clustering-step comparison, in which the trimmed samples are rescored against the optimal cut points without the later stages. We hold the three degenerate computations at their published stars.

One baseline, two counterfactuals: the baseline is always the stars CMS assigned, read from its own files (the published overall rating, together with the rating recomputed from the published measure stars under the documented rules). The exact optimum can replace Ward's method in either of two places. When the exact optimum is computed inside each of the ten resampling groups, everything else the Technical Notes specify is kept, but the result then depends on the order of the input file, which CMS does not publish. Under different sampled

orders, different contracts cross the bonus line. When instead the exact optimum is computed once on the full trimmed sample, the result is unique and does not depend on the order. This second counterfactual gives the bonus-status changes reported below. The two counterfactuals move different sets of contracts, and neither set contains the other: contracts exist that flip under every sampled ordering of the first and not at all under the second, and vice versa.

With the later published stages applied, the full-sample optimal cut points change the bonus status of 60 contract-years (Table 2, block b). Of these, 34 are valued in the dollar estimate (31 distinct MA-PD contracts), sixteen crossing the bonus threshold upward and eighteen downward, with 24 further flips counted and never priced and 2 endpoint-conditional candidates excluded. Among the 24, ten are stand-alone Part D contract-years (eight contracts), which have no bonus program, and two are Cost contracts, to which no quality bonus attaches. Eight more have no published overall rating ("Not enough data available" or "Plan too new to be measured"). The last four are star-year-2026 contract-years whose status changes against the published overall rating or against the recomputed rating but not both. That disagreement leaves them undecidable under the agency's upward-only recalculation of June 17, 2026. The two endpoint-conditional candidates would change status only if the prior-year cut point lay at the closed endpoint of its published display interval. We count them among the 60 but in no other total. Appendix B lists all 60, with dollar amounts for the valued ones.

The 34 priced flips carry an estimated gross \$0.9–1.4 billion in bonus payments riding on the difference (using the stated rule relative to the assigned stars) summed over payment years 2025 through 2027.[1] These figures are gross estimates, and they imply nothing about the conduct of any contract or of CMS. Multiplying the enrollment of the valued contract-years by \$400–600 per enrollee per year [1, 24] and summing over the three payment years gives a band of \$916–1,375 million with midpoint \$1,146 million. The enrollment total is a lower bound, because three of the valued contract-years have zero reported enrollment. The sixteen upward changes account for \$130.3 million at midpoint and the eighteen downward changes for \$1,015.3 million (bands and enrollments in Table 2). Of the 34, seventeen already change status when a single run of Ward's method on the full sample replaces the resampled computation. The other seventeen, which change only under the exact optimum, account for \$272–408 million.

The dollar total is concentrated in a few contracts. The largest valued contract-year, H0524 in star year 2024, holds 59.2 percent of the valued enrollment (1,355,332 enrollee-years, \$542–814 million), and the eight largest contract-years (eight distinct contracts) hold 88.2 percent. H0524's published rating is only about 0.030 above the 3.75 rounding edge (the rating value at which the displayed star changes). Thus the largest component of the dollar total comes from a contract that qualified for the bonus by a narrow margin. One upward change, contract H0473 (\$6.8–10.2M), clears the rounding edge by only four parts in a hundred thousand of a rating point. All 34 remain when every fitted profile and removal of the sample reconstruction is dropped, and 20, H0524 and H0473 included, under every variation of the fitted values we tested (Appendix C.6).

Three contracts (H2816, H6988, and H2422) change downward in one star year and upward in the next. With the Part D pair S4802 and S5660 included, five contracts appear in two star years, and the enrollment and dollar totals are accordingly stated in enrollee-payment-years. We traced the largest changes measure by measure in the published files. For example, H0524's 2024 rating moves from 3.7801 to 3.5964 across six measure changes, two of them upward.

[1] Versions of this analysis circulated before September 10, 2026, including a revision of August 24, 2026, reported larger totals under conventions since corrected (Section 2); the present figures supersede them.

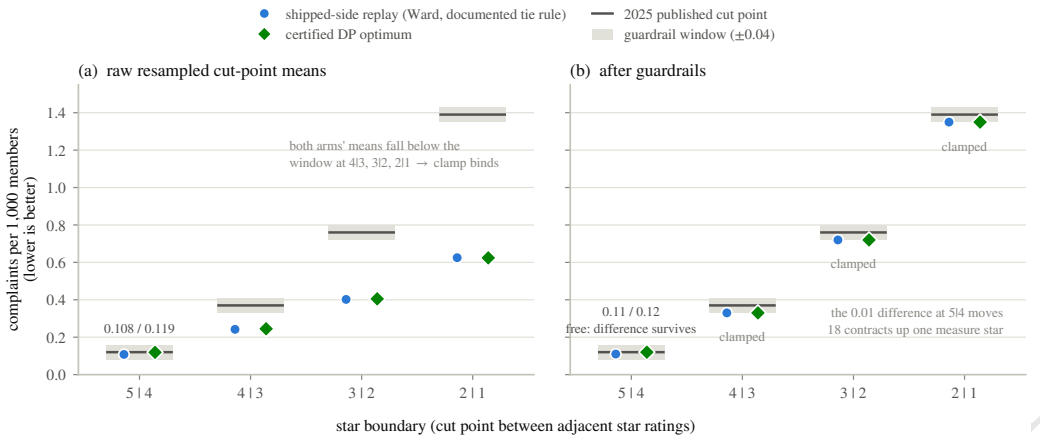

Figure 3: The guardrail clamps three of the four boundaries to the same value for both methods; at the fourth, where it does not bind, the two methods differ (0.11 against 0.12 in the assignment drawn here), and 18 contracts then move up one measure star under the optimum's cut point. The measure is Complaints about the Health Plan (Part C, star year 2026); the prior-year cut points are the published 2025 values (2025 measure C25, matched by name), and the shaded band is the guardrail window, the prior-year cut point ±0.04 (the documented cap, 5% of the prior year's restricted range). (a) The resampled cut-point means from Ward's method (documented tie rule) and from the exact optimum, both below the prior-year cut points at all four boundaries. (b) After the guardrail, the three means outside the window are clamped to its edge identically for both methods, and the 5|4-star boundary keeps its 0.01 difference. Both panels use one illustrative assignment of contracts to the ten groups, a realization from a candidate for the assignment procedure (Section 4.3). The unclamped difference and the count of 18 are specific to that assignment, whereas the clamp is the same at every assignment, including those of the identified procedure. Values are exact rationals shown as decimals.

The split between upward and downward changes does not come from optimal cut points that are uniformly harder or uniformly easier than the published ones. Across the 480 boundary cells of the 120 non-degenerate computations, the optimal cut points are harder than CMS's in 110, easier in 150, and equal in 220. The published final cut points, which also reflect resampling, the guardrail, and rounding, attain the optimum in 10 of the 120 computations when scored as whole partitions of the trimmed samples.

Where the guardrail binds, it clamps the cut point to the same value in the published computation and in the counterfactual (Figure 3). Of the 60 changes in bonus status, 33 involve at least one guardrail-clamped measure on which the optimum's star differs from the published star (Table 2). Without the guardrail, 18 changes disappear, all of them upward. Eleven are among the 16 valued upward changes. The rest are one contract-year with no published overall rating, the two excluded endpoint candidates, one contract-year left undecidable by the 2026 recalculation, and three Part D contract-years. These eighteen are not a subset of the 33, which are counted with the guardrail in place. For several of the eighteen the optimum's star and the published star agree on the clamped measures, and removing the guardrail changes the optimum's star there, which drops the rating back below the bonus line.

Because many contracts are rated close to 4.0, the bonus-status counts are sensitive to the convention that holds the improvement measures at their published stars. In the pub-

lished ratings themselves, a single-star change in one improvement measure moves 69 MA-PD contract-years across the threshold over the three star years (by year in Table 2). When both improvement measures may change together, 104 contract-years cross the threshold. No MA-only contract-year crosses (MA-only contracts lack Part D coverage), and stand-alone drug plans are excluded. On the recomputed side this freedom acts in both directions; separately, recalculating the aggregation's reward-factor thresholds from the changed stars instead of holding them as printed leaves 32 of the 34 and raises the dollar band (Appendix C.6).

The agency's own June 2024 recalculation of the 2024 star year, which followed the SCAN ruling, gives a sense of scale. Between the original and the recalculated release of the published files, 72 cut-point cells differ, as do 1,416 contract-measure stars across 678 contracts and the summary ratings of 239 contracts. These counts come from our comparison of the two releases, since CMS publishes no change log against which they could be checked. We do not add the recalculation's changes to those of the counterfactual, because the recalculation is the agency's own published action, whereas the counterfactual is our computation of what the same public data yield under the stated criterion. The recalculation does show that rating changes on a larger scale than the counterfactual's have already been made and published within the program.

4.3 Dependence on unpublished inputs

Because one randomized step is not documented fully enough to be reconstructed, the comparison between CMS's own computation and the exact optimum cannot be completed for every contract from the published files alone. The Technical Notes print the group-assignment step as SAS code with its seed [4–6]. No edition states how the procedure converts the seeded random-number stream into ten group labels or in what order it reads the contracts, and the labels depend on both. For star years 2024 and 2025 the Notes also leave open which of SAS's two random-number generator families was used, and we analyze both.

We can test candidate reconstructions of the group assignment against numbers CMS itself publishes, because each year's Technical Notes print the mean-resampling thresholds before the guardrail. We built 102 candidates, an 80-combination grid over the documented algorithm, generator, and ordering families plus 22 hand-built constructions. On this test the best grid candidate matches only 21 of the 77 published threshold cells, and none of the 22 hand-built constructions reproduces the published thresholds either (Appendix C). The 102 candidates are the interpretations of the procedure that we could construct from its documentation and observed behavior, and the conclusion is limited to them: from the published documents alone, the computation that produced the published cut points cannot be re-executed.

Instead, we identified the procedure empirically, by running the documented SAS group-assignment step ourselves on inputs of known size and order and reading off the assignments it produced. The step assigns the groups by a front-swap form of the Fisher–Yates shuffle applied to the random stream. The identified rule reproduces 1,240 of 1,240 assignment cells from those runs and is subject to four restrictions, stated together with the runs in Appendix C. Once the procedure is identified and the seed is known from the Notes, the group assignment is determined entirely by the unpublished input order (and, for the two earlier years, by the generator family).

The input order cannot be recovered from the published cut points either. Under the identified procedure the natural contract-identifier order matches 226 of 460 published final cut points. That is nominally the best score among the orders we tried, but since candidates built on procedures already rejected reach 222, the difference is not informative. This is because a rerun's agreement with the published cut points depends mainly on the clustering sample (the

factor CMS names in its caveat that exact replication is not possible) and only weakly on the group assignment. We therefore do not identify the input order. Only the 2026 Notes state the software release, which implies the modern generator, the one used in our identification runs; for 2024 and 2025 we retain both possibilities and make no claim about which ran.

For 50 of the 114 computations that our rerun of the resampled procedure covers, we found an input order at which the rerun reproduces all four published cut points. We exhibit one such ordering for each computation and make no claim about the others. Inside the ten resampling groups the exact optimum is less sensitive to the input order than Ward's method. Across 28 sampled input orders (four conventional orderings and twenty-four random permutations), 65 of the 114 computations are fully invariant to the order with the exact optimum, against 20 with Ward's method.

Under the identified procedure we recomputed every contract's bonus status at each of the 28 sampled input orders. In all, 129 contract-years change bonus status under at least one sampled order (Table 2 gives them by class). Among them, 98 carry no other caveat: they have a published overall rating and are neither Cost nor stand-alone Part D contracts. Of the 129, 118 fall on different sides of the bonus line under different orders, 92 of them with no other caveat. Their bonus status is undetermined by the published record, a statement about the published files and not about any contract. Eleven change status under every sampled order, six with no other caveat, and any one of them could still fail to flip under a twenty-ninth ordering. With further sampling these sets can only grow or be reclassified, and invariance over all orderings cannot be established from the public files.

We express the dollar consequence of the unpublished order as a spread, namely an expected value over one equiprobable draw from the 28 sampled orders. For each draw we take the contracts whose bonus status under the drawn order differs from their modal status across the 28 and sum their gross bonus payments at the \$500 midpoint of the per-enrollee band. The quantity was specified in writing before its values under the exact optimum were computed. With Ward's method inside the ten groups, as CMS runs it, the expectation is \$787.2 million per draw. This expectation arises from 87 valued contract-years whose bonus status varies with the order. The largest of them, H0524, holds bonus status in 12 of the 28 orders.

When the exact optimum replaces Ward's method inside the same ten groups, the expected payments at stake fall by a factor of eight, to \$98.4 million per draw, with 23 contract-years still varying and H0524 no longer among them. Both figures refer to the 28 sampled orderings, with the realization set, contract pools, published baseline, and pricing identical in the two computations. Each figure is one method's own variation with the order; at the same order the two methods disagree on the bonus status of 26 to 47 valued contract-years, six of them at all 28 orders (Appendix C.6). Within those orders the residual of roughly \$100 million arises from the resampling construction itself, which makes the group assignment depend on the order, whichever clustering method is used inside the groups. No published file contains the input order, and no rule requires CMS to publish it.

5 Patient-experience star ratings

The four patient-experience programs introduced in Section 2 assign the HCAHPS stars of hospitals, the Home Health Care CAHPS stars, the Hospice CAHPS stars, and the In-Center Hemodialysis CAHPS (ICH CAHPS) stars of dialysis facilities. Each program's technical notes prescribe Ward clustering with the within-cluster sum of squares of Eq. (1) as the quantity to

be minimized. The hospital survey's notes call this method "the same one employed by CMS to determine Medicare Part C and Part D Star Ratings." No payment depends on these stars, and we attach no dollar figure to them. The stars appear on Medicare's comparison sites, and for three of the four the underlying survey scores are also used in separate payment programs. For two of the four the public files suffice to test whether the published groupings attain the stated minimum.

5.1 Hospital and dialysis survey stars

The public HCAHPS file lists each hospital's per-measure stars beside the linear mean scores, rounded to integers, that the technical notes say are grouped. We examine five consecutive quarterly refreshes, April 2025 through May 2026, on the eight measures starred in all five. For each refresh and measure we compare the published grouping with the exact optimum on the published scores. This comparison does not recompute CMS's own clustering and takes the notes' description of the inputs as given. Nor would a recomputation from the file reproduce the published groupings, since Ward's method applied directly to its starred hospitals recovers only one of the eight on the May 2026 refresh, and none of the other seven under any input order or tie resolution we tried (Appendix C.6). Hence, under the Ward step the notes describe, the agency's inputs are not the file's starred set, and the divergences below combine input differences with suboptimality of the algorithm.

We find that the published HCAHPS grouping misses the stated minimum in 38 of the 40 refresh-measure cells and attains it in the other two. The exact optimum is unique in every cell. Regrouping each cell to its optimum moves 17,484 hospital-measure stars over the five refreshes, in both directions, 3,363 of them on the May 2026 refresh (2,118 up, 1,245 down). Two measures that CMS later retired, staff responsiveness and care transition, add six cells from the 2025 refreshes. Five of the six miss the minimum, and the sixth, staff responsiveness on the November 2025 refresh, attains it. Regrouping the five moves a further 2,518 stars, for a five-refresh total of 20,002. In Appendix C we compare each refresh's optimum with the cut points CMS adopted in the following quarter. Because those cut points are computed on newer data, their agreement with the earlier optimum reflects covariation of the two quarters' data.

Patients see one HCAHPS summary star per hospital, a published combination of the per-measure stars. On the two refreshes whose summary rule draws on six measures, the rule applied to the published per-measure stars reproduces every published summary star (3,176 of 3,176 in May 2026 and 3,183 of 3,183 in February 2026). When we apply the same rule to the optimal per-measure stars instead, the summary star changes for 413 of 3,176 hospitals (13.0%) in May 2026 and for 316 of 3,183 (9.9%) in February 2026. In the 2025 editions of the technical notes, the summary rule drew on ten starred measures, so we do not compute a summary-star counterfactual for those editions.

For the dialysis survey (ICH CAHPS) the April 2026 public file contains the clustered quantity itself, a linearized score for each of 2,594 rated facilities on six measures. Ward's method applied to this file reproduces five of the six published groupings identically and the sixth up to ties among its 46 distinct scores (all six under some input order; Appendix C.6), which confirms that the published scores are the agency's inputs. The dialysis divergences are therefore attributable to the algorithm and not to the inputs.

We find that none of the six published dialysis groupings attains the stated minimum, and each optimum is unique. Regrouping moves 4,061 facility-measure stars, about 26% of the 15,564 in the file. The share of facilities that move runs from 3% on the least-affected measure

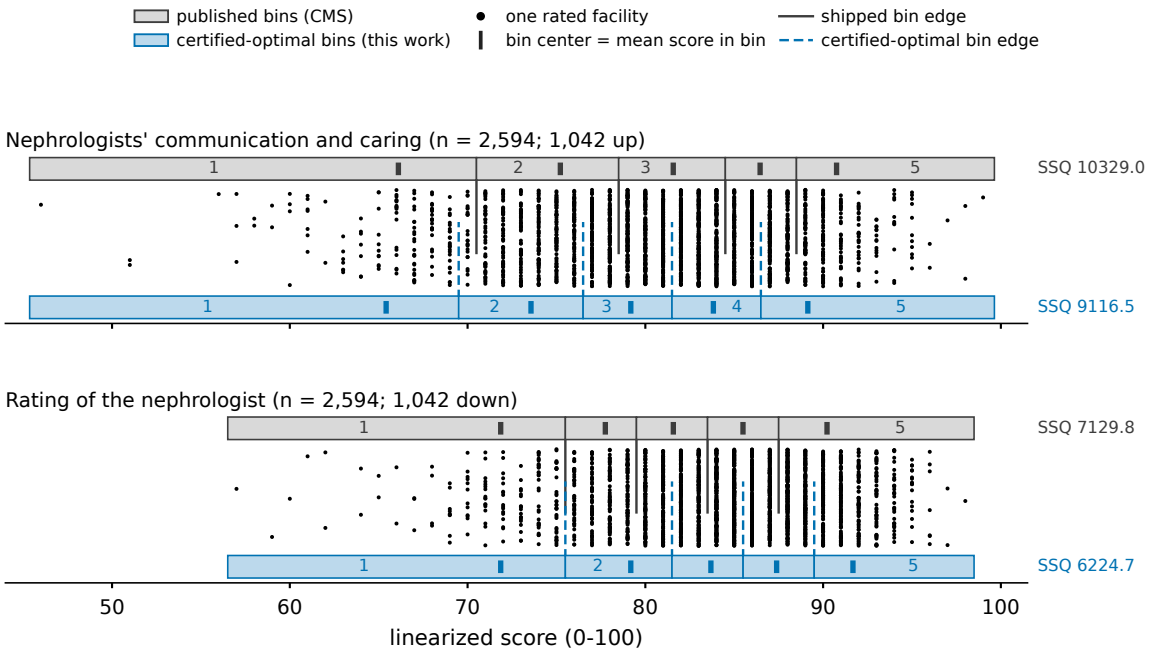

Figure 4: Regrouping to the exact optimum moves 1,042 of 2,594 dialysis facilities (40%) on each of the two most-affected dialysis-survey measures (ICH CAHPS, April 2026 file), every move upward on nephrologists' communication and caring and every move downward on rating of the nephrologist. Each facility is a dot at its published linearized score; the upper ruler shows the published bins and the lower ruler the optimal bins, drawn as in Figure 1, with the published and the optimal SSQ beside each panel. The optimal SSQ is strictly lower on both measures.

to 40% on each of two measures (Figure 4). These two are nephrologists' communication and caring (1,042 facilities, all moving upward) and rating of the nephrologist (1,042 facilities, all moving downward). On every measure the movement is in one direction only, upward on five and downward on one. For example, on the facility-rating measure, 74 facilities move from one star to two, 290 from two to three, 228 from three to four, and 158 from four to five.

The home health and hospice surveys cannot be checked from the public files. The home health file lists the stars but not the adjusted linear scores that the notes say are grouped. For the hospice survey, each measure's cut points are the average of cut points from four separate six-month clusterings, minus averaged adjustment factors; neither these intermediate quantities nor the per-measure stars appear in the public files. CMS does not publish the scores, so we cannot check; we suspect error rates similar to those in the systems examined here, where the share of hospitals whose displayed star changes runs from about one in fifteen to one in eight and a single dialysis measure reaches 40% of the facilities.

5.2 The six programs together

For four of the six programs the public files include the scores as well as the stars, and in each of the four at least one published clustering fails to attain the stated minimum over the range examined. Table 3 gives the rates at which published stars in these programs differ from the exact optimum. The rates are not comparable across rows, since a changed bonus status is not

System	What the row counts	Share off the optimum
Medicare Advantage	bonus status of contract-years included in the dollar estimate, published stars against the full-sample optimum, star years 2024–2026	34 of the 1,724 published contract-year ratings (2.0%)
	cut-point computations off the stated minimum at the clustering step, star years 2024–2026, documented-rules clustering samples	104 of 123 (104–106 across three sample reconstructions)
Hospital overall stars	displayed stars, 2026	205 of 3,182 (6.4%)
	all ratings issued, 2021–2026, counted before the safety cap	1,146 of 18,493 (6.2%)
HCAHPS summary stars	current refresh (May 2026)	413 of 3,176 (13.0%)
	February 2026 refresh	316 of 3,183 (9.9%)
Dialysis survey stars	facility-measure stars, April 2026 file	4,061 of 15,564 (26%)
	per measure	3%–40%

Table 3: Each row counts a different thing, so the shares are not comparable across systems. In all four programs whose scores are public, some published groupings do not attain the stated minimum. The upper Medicare Advantage row counts contract-years whose bonus status changes under the full-sample optimum and the conventions of Section 2, out of all 1,724 published contract-year ratings, including the stand-alone Part D and Cost contracts, which carry no bonus. The lower row counts cut-point computations at the clustering step, before resampling. In the survey rows each published grouping is compared with the optimum on the published scores. The home health and hospice surveys, for which CMS publishes stars but not scores, are absent from the table.

a changed displayed star. Appendix A shows nineteen of the clusterings examined, drawn on their score lines.

Other federal star-rating systems whose published methodologies we reviewed assign stars by fixed thresholds or percentile cuts, and none states an optimality criterion. A companion paper [25] examines whether those systems apply their rules as published.

6 Discussion

In the four CMS rating programs whose scores are public we compared the published star groupings with the exact optimum of the stated clustering criterion. For two of them (the hospital ratings and Medicare Advantage) we also reran the agency's clustering from the published code and Technical Notes.

6.1 Reproducibility and optimality

To our knowledge the optimum had not been compared with the published ratings before, although an exact algorithm for minimizing the stated criterion dates from 1958 [11] and is available in standard statistical software [12]. CMS wrote the Medicare Advantage objective into 42 CFR 422.162 [2] and the Technical Notes [4–6], and published the hospital computation in full. Both programs use deterministic algorithms: the hospital code fixes its starting centers and the Medicare Advantage Technical Notes fix the random seed, so that a rerun on the same inputs in the same order returns the same grouping.

The natural test of a published computation is to reproduce it. We could reproduce every published hospital star, and could re-execute each documented stage of the Medicare Advantage computation except the group assignment, whose input order is unpublished. However, reproduction does not test the grouping against the criterion stated in the same documents. Therefore, in both programs the divergence between the published stars and the optimum could persist for years without an error by anyone. Bayles and Huang [17, 18] asked whether the k-means iteration used for earlier hospital ratings had converged. The code examined here does converge, but only to a local optimum fixed by its starting centers.

A published Medicare Advantage cut point that differs from the optimal one changes a contract's star only if its score lies between the two. The result is therefore not that the program rates too generously or too harshly in aggregate, but that specific contracts have a bonus status under CMS's computation that differs from their status under the full-sample optimum of the objective the regulations state.

6.2 Regulatory and legal context

CMS left the Medicare Advantage measure specifications to the Technical Notes, which 42 CFR 422.166 does not incorporate by reference [3]. The clustering objective (§422.162(a)) and the prescribed method (§422.166(a)(2)(i)), however, are both regulation text, and the tension we measure lies between two provisions of the regulations [2, 3]. The CY2027 final rule (91 FR 17384) adds 42 CFR 422.166(h)(2), effective June 2026. Under that provision, "During the second plan preview, CMS will display de-identified contract-level sample data for one of each type of measure needed to replicate the cut point methodology, as determined by CMS." From the documents published so far we could not carry out that replication (Section 4.3).

The court rulings that have altered the Medicare Advantage Star Ratings have so far each turned on one element of the methodology. *SCAN Health Plan v. HHS*, No. 1:23-cv-03910 (D.D.C. June 3, 2024), enforced the guardrail rule as written, and CMS then recalculated the 2024 ratings program-wide. *Clover Insurance Co. v. HHS*, No. 2:25-cv-142 (S.D. Ga. May 27, 2026), found ten measures outside the statute's data-source limits and treated the measure specifications as statements of policy under Section 1395hh(a)(2). The Clover rulings prompted the upward-only June 2026 recalculation of the 2027 bonus ratings. By contrast, *Humana v.*

HHS, 806 F. Supp. 3d 642 (N.D. Tex. 2025), had rejected the statement-of-policy theory, and the split is pending on appeal. Across these opinions no court has reached the question measured here, whether the clustering as CMS runs it attains the criterion the program's own rules state.

6.3 Limitations

Several limitations apply. We examine the six systems separately, applying the same criterion and method to each system's data where those data are public, and we do not pool the results. The Medicare Advantage dollar figures are estimates. They depend on the conventions of Section 2, on a per-enrollee value drawn from published estimates, and on a small number of large contracts, one of which accounts for more than half of the priced enrollment. Because we hold the improvement measures and the Medicare Advantage CAHPS measures at their published stars, the number of overall ratings that change is conditional on those stars and could be larger or smaller. For the improvement measures Appendix C.6 bounds the effect in both directions; for the CAHPS measures it is not quantified. For star year 2023 there are no published fence values to constrain the rebuilt clustering sample. The dollar figure for payment year 2027 uses July 2026 enrollment as a proxy for enrollment in that payment year. The bonus status of the contract-years we report as not determined by the published files depends on an input order that only CMS can supply. The Medicare Advantage clustering samples are themselves reconstructions from the public files, and some of the valued changes depend on the fitted profiles in them (Appendix C.6). The hospital comparison rests on code and input files obtained from public mirrors, and the two survey comparisons take the technical notes' description of the inputs as given.

6.4 Conclusion

For Medicare Advantage, publishing the clustering input file of one production run, that is, the scores that entered each computation in the order the procedure read them, or the group assignment itself, would make the agency's computation re-executable. The input order alone would not suffice, because the clustering sample is at present a reconstruction that the published fence values check but do not identify (Appendix C). For star years 2024 and 2025 the random-number generator family is also needed. Such a disclosure would settle the bonus status of each contract-year that the published files leave undetermined. The hospital ratings need no further disclosure, because a rerun of the mirrored public code already reproduces every published star.

CMS's CY2027 Advance Notice proposed replacing clustering with percentile-based cut points [26], and the Rate Announcement deferred that change to future rulemaking [27]. Thus clustering remains the method under the rules now in force. Wherever a program's rules state an optimality criterion for scores on a line, whether the published stars attain it can be determined by direct computation. In each of the four Medicare systems whose scores are public, at least some of the published stars do not attain it.

Data and code availability

All inputs are public CMS releases: the Care Compare hospital files [20]; the 2021–2026 hospital code and input files [8, 9], from the GitHub repositories klocey/stars-data-builder and Kentucky-Open-Science/CMS-Star-Rankings; the Medicare Advantage Star Ratings tables, Technical Notes [4–6] and enrollment reports; and the HCAHPS and In-Center Hemodialysis CAHPS public files.

Code and input files sufficient to reproduce each number in this paper are available from the paper's web page, together with the output of the SAS runs used to identify the group-assignment procedure (Appendix C).

A Clusterings drawn on their score lines

Every comparison we make concerns two groupings of scores on a line and can be drawn in full. Nineteen of them follow, in the format of Figure 1. The upper ruler and solid edges mark the bins of CMS's clustering, the lower ruler and dashed edges those of the exact optimum. Each edge extends two thirds of the way through the dots (one dot per rated entity at its score, jittered vertically). Thus the entities between a paired solid and dashed edge are those the two groupings rate differently. Thick ticks mark bin centers, and each ruler ends with its grouping's SSQ, Eq. (1), except in the eight hospital-survey panels.

The top panels of Figure A1 show the 2026 comparison for hospital peer groups 4 and 5. Peer group 3 is in Figure 1, and the fifteen peer-group computations of 2021–2025 (Table 1) are not drawn. The bottom panels show, for each Medicare Advantage star year, the computation with the largest gap in Figure 2, on the clustering sample reconstructed from the documented rules. We make the comparison before resampling, the guardrail, and rounding. The other 117 of the 120 non-degenerate computations appear in Figure 2 only as gap values, 16 of them at zero, where CMS's Ward step attains the minimum.

Figure A2 shows the fourteen patient-experience measures of Section 5.1: the hospital survey's eight starred measures on the May 2026 refresh and the dialysis survey's six on the April 2026 file. In both we compare the file's published grouping with the optimum on the same scores. The four earlier hospital-survey refreshes are not drawn. We draw no panels for the home health and hospice surveys, for which CMS publishes stars but not scores.

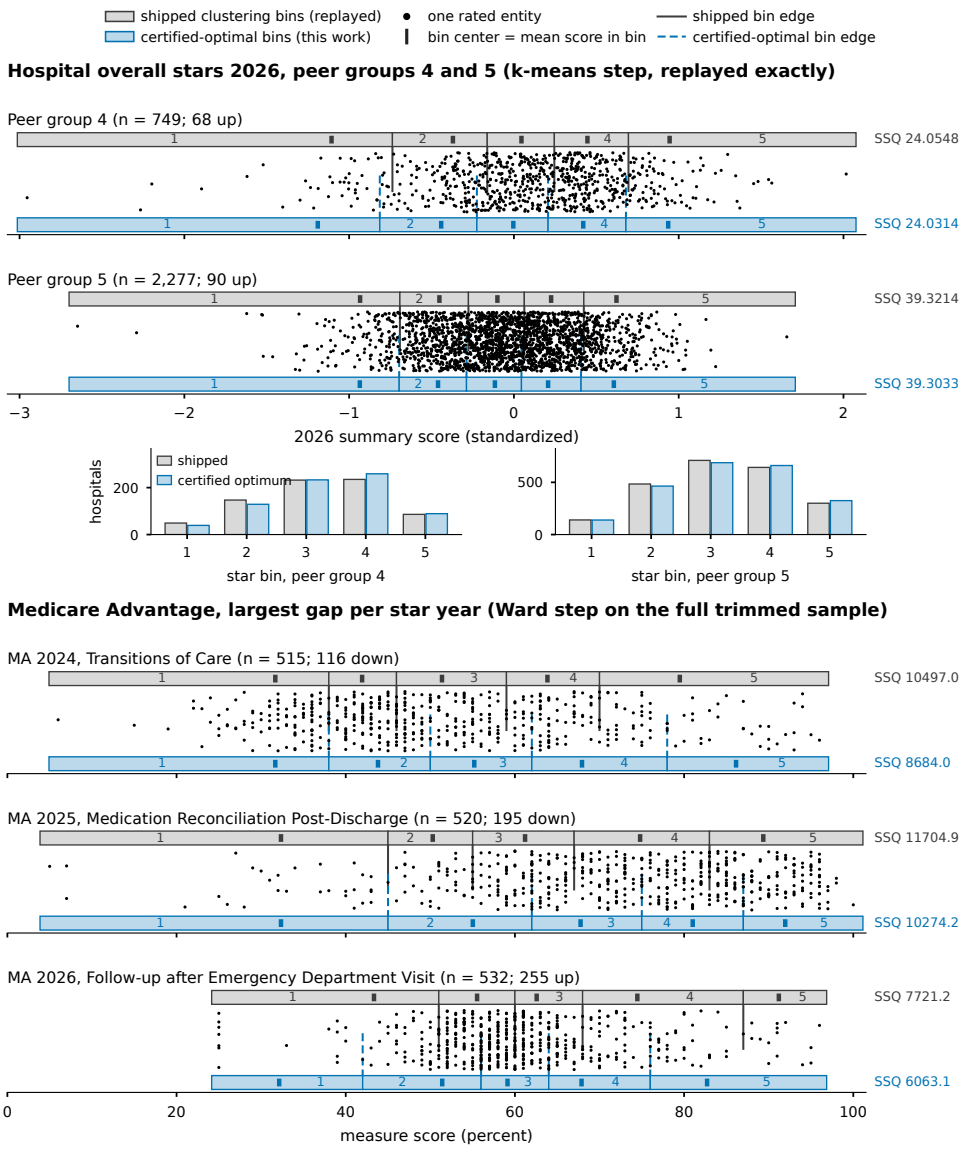

Figure A1: In all five panels CMS's cut points differ from the exact optimum's, and within each panel all regrouped entities move in the same direction. Top: hospital peer groups 4 and 5 in 2026 (749 and 2,277 hospitals) under CMS's k-means grouping before the safety cap. Their 68 and 90 regrouped hospitals all move up and, with peer group 3's 55, form the 213 differences counted before the cap. Bottom: the largest-gap computation of each Medicare Advantage star year on the documented-rules clustering sample. These are Transitions of Care in 2024 (515 contracts, 116 down), Medication Reconciliation Post-Discharge in 2025 (520, 195 down), and Follow-up after Emergency Department Visit for People with Multiple High-Risk Chronic Conditions in 2026 (532 of 533, one contract trimmed, 255 up). Solid edges mark CMS's Ward grouping of that sample, dashed edges the optimum. The published cut points, averaged over ten subsample clusterings, are not drawn.

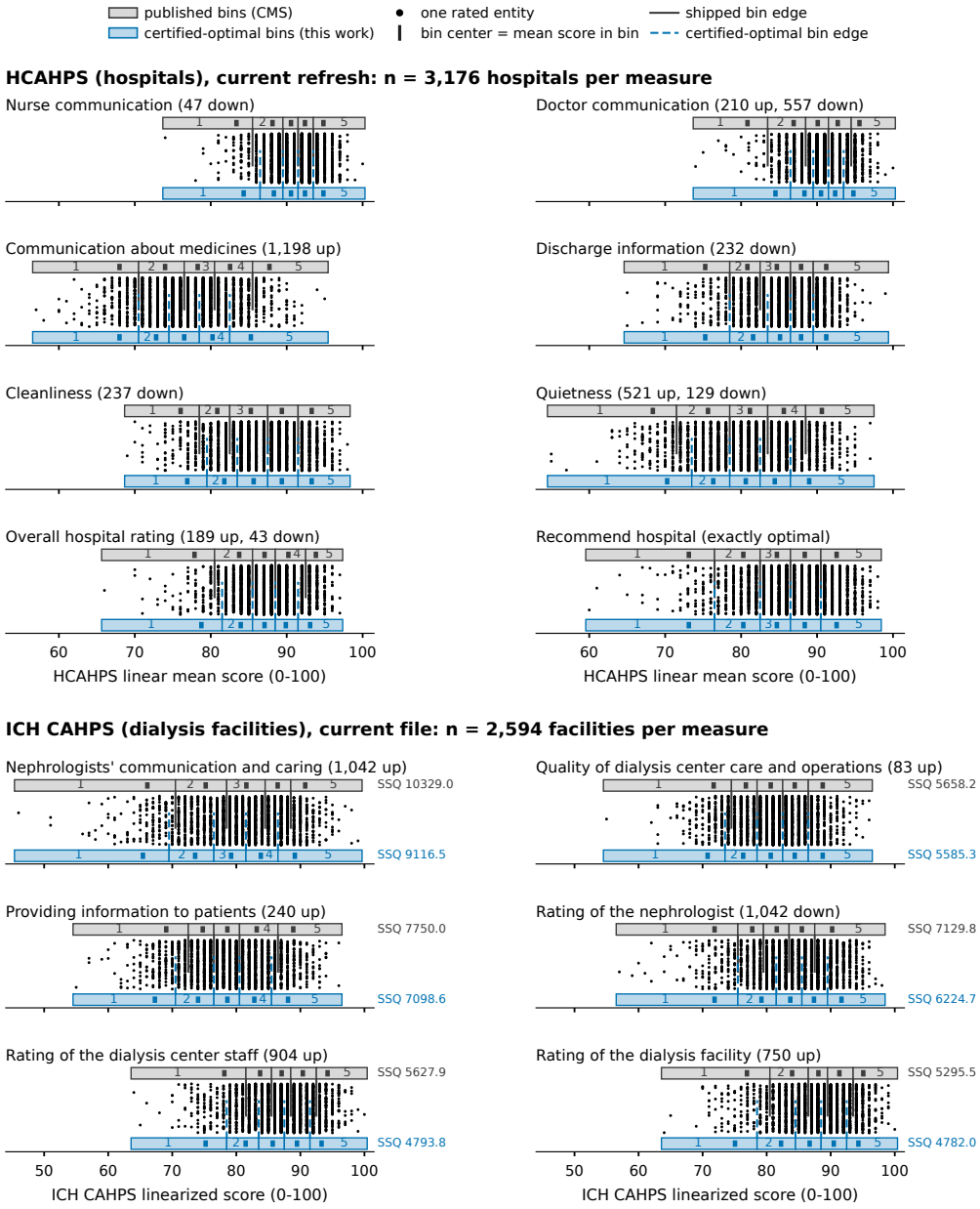

Figure A2: Thirteen of the fourteen patient-experience measure clusterings with public entity-level scores fail their stated criterion; the fourteenth, Recommend hospital, is exactly optimal, and its paired cuts coincide. Upper block: the eight starred measures of the hospital survey (HCAHPS) on the May 2026 public file, one of the five quarterly refreshes examined (the four earlier ones are not drawn). Lower block: the six measures of the dialysis survey (ICH CAHPS) on the April 2026 reporting file, built from the fall 2024 and spring 2025 surveys. Upper rulers show the published grouping read from the file, without rerunning CMS's code. Lower rulers show the optimum computed on the published scores, taken as the clustering inputs. Each panel heading gives the number of entities whose star changes, by direction. These numbers sum to the totals reported in Section 5.1. SSQ values for the two groupings are printed beside the dialysis panels only.

B Bonus-status changes by contract-year

Table B1 lists the 60 contract-years of Section 4.2 whose baseline rating and criterion rating lie on opposite sides of the 4.0 bonus line. We value 34 of them in the dollar estimate. We count another 24 without a dollar value: ten stand-alone Part D contract-years, two Cost contracts, eight with no published overall rating, and four that are undecidable under CMS's upward-only recalculation of June 17, 2026. The remaining two are endpoint-conditional upward candidates, listed in the table but excluded from all other counts. To obtain the criterion rating, we take the cut points of the full-sample exact optimum to the overall rating through the later stages of the published method (guardrail, rounding, and aggregation) described in Section 2. For star year 2026 the published baseline is the overall rating in CMS's July 22, 2026 summary file, which reflects the upward-only recalculation and was posted August 17, 2026.

Table B1: Contract-years whose bonus status differs between the published stars and the exact optimum of the stated criterion. Rows describe computations, not conduct. The Rating column gives baseline → criterion at display precision. The baseline is the published overall rating or, where none exists (no-overall and PDP rows; Cost contract H2462), the rating recomputed from published measure stars. The Class column abbreviates the block heading. PDP and Cost rows have no bonus program; recalculation rows change status against only one of the two baselines (published and recomputed); excluded rows change status only at the closed endpoint of the prior-year cut point's display interval. Enrollment is CMS's monthly count for December 2025 (star year 2024) or July 2026 (star years 2025 and, as a proxy, 2026), and is omitted for the PDP rows. H3749, H6502, and R3332 do not appear in CMS's enrollment report for their month and are valued at exactly zero; the H8095 row shows zero enrollment because CMS suppressed that contract's count in the report. QBP (Quality Bonus Program) dollars are enrollment times $400–600 per enrollee per year, rounded outward at one decimal, estimate-class, not certified; the certified object is the flip set. The seven conventions of Section 2 apply to every row.

Star year	Contract	Direction	Rating	Enrollment	QBP dollars	Class
Valued: bonus-relevant, CMS-published overall rating, changes against both baselines						
2024	H0524	down	4 → 3.5	1,355,332	$542.1–813.2M	valued
2024	H3805	down	4 → 3.5	220,404	$88.2–132.2M	valued
2024	H0321	up	3.5 → 4	71,220	$28.5–42.7M	valued
2024	H5900	down	4 → 3.5	47,487	$19.0–28.5M	valued
2024	H0473	up	3.5 → 4	16,927	$6.8–10.2M	valued
2024	H6988	down	4 → 3.5	2,904	$1.2–1.7M	valued
2024	H2816	down	4 → 3.5	1,855	$0.7–1.1M	valued
2024	H1969	up	3.5 → 4	691	$0.3–0.4M	valued
2024	H3749	up	3.5 → 4	0	$0	valued
2024	H6502	down	4 → 3.5	0	$0	valued
2024	R3332	up	3.5 → 4	0	$0	valued
2025	H4004	down	4 → 3.5	102,368	$40.9–61.4M	valued
2025	H5211	down	4 → 3.5	79,359	$31.7–47.6M	valued
2025	H5591	down	4 → 3.5	62,630	$25.1–37.6M	valued
2025	H2322	down	4 → 3.5	40,309	$16.1–24.2M	valued
2025	H3347	up	3.5 → 4	26,840	$10.7–16.1M	valued
2025	H4227	up	3.5 → 4	22,975	$9.2–13.8M	valued
2025	H5262	down	4 → 3.5	18,984	$7.6–11.4M	valued
2025	H3653	down	4 → 3.5	15,257	$6.1–9.2M	valued

continued on next page

Bonus-status changes under the exact optimum (continued)

Star year	Contract	Direction	Rating	Enrollment	QBP dollars	Class
2025	H8189	down	4 → 3.5	12,924	$5.2–7.8M	valued
2025	H4623	up	3.5 → 4	10,953	$4.4–6.6M	valued
2025	H6988	up	3.5 → 4	8,077	$3.2–4.8M	valued
2025	H2422	down	4 → 3.5	5,775	$2.3–3.5M	valued
2025	H2816	up	3.5 → 4	1,829	$0.7–1.1M	valued
2025	H8547	up	3.5 → 4	1,553	$0.6–0.9M	valued
2025	H4172	down	4 → 3.5	1,345	$0.5–0.8M	valued
2026	H0354	up	3.5 → 4	75,390	$30.2–45.2M	valued
2026	H0174	down	4 → 3.5	54,139	$21.7–32.5M	valued
2026	H3080	up	3.5 → 4	9,590	$3.8–5.8M	valued
2026	H6202	up	3.5 → 4	6,703	$2.7–4.0M	valued
2026	H2422	up	3.5 → 4	5,775	$2.3–3.5M	valued
2026	H0978	down	4 → 3.5	5,652	$2.3–3.4M	valued
2026	H8010	down	4 → 3.5	3,804	$1.5–2.3M	valued
2026	H3777	up	3.5 → 4	1,983	$0.8–1.2M	valued
Endpoint-conditional upward candidates, excluded from every count						
2024	H1304	up	3.5 → 4	6,775	—	excluded
2025	H5273	up	3.5 → 4	11,065	—	excluded
Counted without a dollar value: no CMS-published overall rating						
2024	H2591	down	4 → 3	510	—	no overall
2025	H6852	down	4 → 3.5	8,164	—	no overall
2025	H3041	down	4 → 3.5	2,267	—	no overall
2025	H4490	down	4 → 3.5	768	—	no overall
2025	H9909	down	4 → 3.5	731	—	no overall
2026	H1619	up	3.5 → 4	4,126	—	no overall
2026	H8845	up	3.5 → 4	1,614	—	no overall
2026	H8095	down	4 → 3.5	0	—	no overall
Counted without a dollar value: 1876 Cost contracts						
2026	H2461	up	3.5 → 4	47,233	—	cost
2026	H2462	up	3.5 → 4	2,762	—	cost
Counted without a dollar value: undecidable under the June 17, 2026 upward-only recalculation						
2026	H4141	up	3.5 → 4	125,206	—	recalculation
2026	H1608	up	3.5 → 4	118,187	—	recalculation
2026	H1994	up	3.5 → 4	66,738	—	recalculation
2026	H7379	up	3.5 → 4	46,701	—	recalculation
Counted without a dollar value: stand-alone Part D						
2024	S2468	down	4 → 3.5	—	—	PDP
2024	S2893	down	4 → 3.5	—	—	PDP
2024	S5660	up	3 → 4	—	—	PDP
2024	S8841	up	3.5 → 4	—	—	PDP
2025	S4802	up	3.5 → 4	—	—	PDP
2025	S5660	up	3.5 → 4	—	—	PDP
2025	S5743	up	3 → 4	—	—	PDP
2025	S5877	up	3 → 4	—	—	PDP
2025	S6875	up	3.5 → 4.5	—	—	PDP
2026	S4802	up	3.5 → 4	—	—	PDP

C Computational and reproduction details

This appendix gives computational details in the order of the main text. It covers the hospital k-means code; for Medicare Advantage, the clustering-sample reconstruction, the group-assignment procedure, and the later stages of the published method; for HCAHPS, the cut points of the following quarter; and, in a final subsection, additional checks on the results of

Sections 3 to 5 in the same order.

C.1 Reproducing the hospital star-rating code

The 2021–2025 SAS packages [8] call the k-means routine FASTCLUS. Its documentation [21] leaves three cases unspecified: a hospital equidistant from two centers, two coincident initial centers, and a tie between cluster means when the clusters are ordered into stars. None of the three arose in the fifteen peer-group runs of 2021–2025, each of which converged within 38 of the 1,000 iterations allowed. In 11 of the 15 runs the documented exclusion option (STRICT) set aside between one and five hospitals by giving them a negative cluster number. A later step of the package re-included them by taking the absolute value of that number.

For 2026 we ran the comparison with the exact optimum on the binary floating-point scores of the R code [9] and on the fifteen-significant-digit decimal strings it writes out. The two representations differ in the least significant bits for 2,978 of the 3,203 scores (93%). Under both representations the same 213 hospitals have an optimal star that differs from the published one, and each of the 213 receives the same optimal star either way.

C.2 Reconstructing the clustering sample

The set of contract scores that enters each Medicare Advantage cut-point computation omits some scores that the public score file contains and includes some that the file suppresses. The Technical Notes [4–6] exclude scores that are missing, flagged, or voluntary, as well as scores covered by the disaster rules. We read these exclusions as extending to the elective 1876 Cost contracts. The public files, in turn, suppress the score cells of small-enrollment contracts, and no list of clustering inputs is published.

We test the reconstruction against the Tukey outer-fence cutoffs that Attachment K of the Technical Notes [4–6] prints for every measure. Wherever a fence falls inside the score range, the printed cutoff is unrounded. Because each interior cutoff is computed from the sample quartiles alone, any candidate sample must reproduce it. The three years yield 123 evaluable fence targets, one per computation. Under the documented exclusion rules alone, with nothing fitted, the reconstruction reproduces both published bounds for 82 of the 123 targets.

The published fences are consistent neither with excluding employer-only contracts from every computation nor with including them in every computation. The 2024 PDP hypertension-adherence fence requires the exclusion: the pair computed without employer-only contracts is $(80, 97.5)$, as published, and with them it is $(82, 96)$. By contrast, two 2026 PDP fences require employer-only contracts to be included: with them the computed pairs are $(0, 0.19)$ and $(81, 95)$, as published, and without them they are $(0, 0.21)$ and $(79.5, 97)$. We take the documented-rules sample as the primary basis and repeat each count on a second, employer-rule basis that adds the employer-only exclusion. On the primary basis we represent the employer exclusions that individual fences require through fitted score profiles.

We match the remaining targets by fitting score profiles for unpublished contracts, with values chosen to reproduce the published fences. Where a whole family of profiles fits, we take any one member, since a fitted profile stands in for suppressed score cells and is not identified with any particular contract. Across the three years the reconstruction adds 56 contracts and removes 10, with 336 fitted score cells in all (the clustering samples hold 20 to 701 contracts per computation). With the fitted profiles, the published fences are matched on 123 of 123 targets in both bases (Table C1). Nine fitted targets have both published bounds interior and unrounded,

Table C1: Reconstruction of the Medicare Advantage clustering sample by star year: the documented exclusion rules alone reproduce the published Tukey fences on 82 of 123 targets, and fitted score profiles for contracts whose scores the public files suppress bring the count to 123/123 in both bases. "Targets" are cut-point computations with an evaluable published outer fence (Attachment K, Tables K-5/K-6). "Parameter-free exact" counts targets reproduced with nothing fitted, on the documented-rules basis (which excludes voluntary scores, 1876 Cost scores, and scores covered by the disaster rules) and on the employer-rule basis, which adds an employer-only exclusion. "Fitted profiles added" and "Removals" count contracts; "cells" counts fitted score cells. "Two-sided" targets have both published bounds interior and unrounded and are classified on the employer-rule basis.

		Parameter-free exact		Fitted profiles added			Fitted layer	
Star year	Targets	doc.-rules	empl.-rule	doc.-rules	empl.-rule	Removals	cells (doc./empl.)	two-sided
2024	40	29	30	15	15	2	46 / 43	1
2025	40	24	24	23	22	8	150 / 149	4
2026	43	29	27	18	19	0	140 / 142	4
All	123	82	81	56	56	10	336 / 334	9

and in all nine the fit reproduces these two independent values at once. Six of the 123 targets are small stand-alone Part D computations. Four of them are reproduced with nothing fitted, and the two for star year 2025 match after four score cells are added to profiles already fitted for other targets.

The set of bonus-status changes of Section 4.2 is identical on both bases and invariant over 33 alternative classes of fitted profiles that we enumerated, taken singly or in combination. The enumeration is over classes, each represented by one constructed member found within a search of 4,000 candidates; alternatives within a class were not swept, and a feasible class for which the search found no member within that budget would be missed. Appendix C.6 reports a second search, which varied the fitted values within the classes.

C.3 Identifying the group-assignment procedure

The Technical Notes print the resampling step as a `SURVEYSELECT` call with the seed and the ten-group split stated (Section 2.4). The Mersenne-twister random-number stream we generate for a given seed reproduces the stream vectors in SAS's documentation for that seed at their full published precision. The Notes do not state the input order (the order of the rows in the data set), on which the assignments depend. Nor do they give the rule that converts the random stream into ten group labels, which we call the readout. Only the 2026 Notes name the SAS release that executed the step, and the release fixes which of SAS's two generator families was used [6].

None of the 102 candidate reconstructions of Section 4.3 reproduces the thresholds that the Notes publish for the mean-resampling step (the cut points before the guardrail). We tested the 22 hand-built constructions among them on the four most constraining published thresholds, where none matches more than 3 of 4. The one construction that we also tested on all 77 published threshold cells matches 16. We identified the readout instead by running the documented step on a SAS installation (SAS OnDemand for Academics, Base 9.4_M8 with SAS/STAT 15.3, on August 1, 2026).

From the SAS runs we obtained six assignment vectors through the `GROUPS=` option. They cover two seeds (one of them the published 8675309), group counts of 3 and 10, and data sets

of 17, 100, and 503 rows. In addition, we printed ten draws from the uniform generator to full 16-digit hexadecimal. The hexadecimal draws fix the conversion from the generator's 32-bit output y to a uniform variate u as one unit in the last place below $y \cdot 2^{-32}$. This conversion matches all ten draws to the last bit, and eight alternative conversions each fail all ten.

The rule we identified is a front-swap Fisher–Yates readout of the random stream. For n rows and G groups, the procedure lays out an array of n group labels in group order. Each group receives $\lfloor n/G \rfloor$ labels, and the remainder goes to the last groups. The procedure then reads the rows in input order. For the row at position w it draws one uniform u and selects the slot

$$j = w + \lfloor u\,(n - w) \rfloor \tag{C1}$$

among the slots from w to the end of the array. It assigns that row the label in slot j and moves the label at slot w into slot j. This rule reproduces all six vectors from the SAS runs (1,240 of 1,240 assignment cells). The group sizes observed in each vector are consistent only with the remainder going to the last groups.

We also tested 3,552 candidate rules drawn from the documented algorithm families and found no distinct alternative. Every candidate that fits the six vectors induces the same selection sequence as the identified rule. In exact rational arithmetic no selection index in these fits falls near a rounding boundary, and hence floating-point evaluation selects the same slots.

Four restrictions apply to this identification. It names the canonical member of an algebraic equivalence class of readouts that induce the same assignments, not the text of SAS's source code. It covers the generator path SAS uses for seeds not divisible by 8192, which includes the published 8675309. The conversion is fixed by the hexadecimal draws alone, because assignment vectors cannot distinguish conversions that preserve the selection floor. At the resolution of our runs the floor form cannot be distinguished from a ceiling form that differs only at exact-integer values of the product, of which the 1,240 draws contain none. Finally, the form we ran is the single-stratum `GROUPS=`G call that the Technical Notes print; list-valued and stratified variants are untested.

Given the readout and the printed seed, the group assignment depends only on the unpublished input order and, for 2024 and 2025, on the generator family. We identified the readout on the stream of the modern generator (the Mersenne twister), and we assume, without a direct test, that it carries over unchanged to the legacy stream. Our rerun of the resampled procedure matches 142 of 304 published cut-point values under the modern generator family and 127 of the same 304 under the legacy family; the modern family is nominally ahead but not decisive.

C.4 Numerical checks and reproduction of the later stages

The dynamic program of Section 2 [11, 12] agrees with exhaustive enumeration on 300 random instances, including the count of tied optima on 100 of them. It also agrees with the R package Ckmeans.1d.dp on 60 further random instances once that package's partitions are rescored in exact arithmetic. We compute the Tukey fences from the reconstructed clustering sample in exact rational arithmetic, using SAS's default quartile definition [23]. The prior-year cut points that enter the guardrail are public only at display precision. We represent each one as an interval of half a display unit either way and propagate that interval through the clamp (Figure 3).

We test our implementation of the aggregation stage by running it on the measure stars CMS published. It reproduces 1,754 of 1,763 published rating cells for 2024 at the original October vintage, 1,673 of 1,681 for 2025, and 1,645 of 1,653 for 2026 (between 99.48 and 99.53 percent).

The cells not reproduced all belong to Puerto Rico contracts, whose adherence-measure weights and categorical adjustment the Notes treat specially. At the recalculated 2024 vintage, the baseline for the bonus-status comparison of Section 4.2, the same implementation reproduces 1,705 of 1,763 cells. This shortfall shows that the recalculated ratings are not reproducible under the documented aggregation rules with the printed reward-factor thresholds, which CMS did not reissue. If instead the printed thresholds are shifted by the movement that the recalculated stars induce in the underlying percentiles, and no rating is allowed to fall below its October value (a rule we infer from the published outcome, in which no rating fell), the same implementation reproduces 1,756 of the 1,763 cells. Hence a contract-year enters the dollar estimate only when its bonus status changes against both the published overall rating and the recomputed rating (Section 2).

We determine the disaster hold-harmless from the two disaster-percentage columns CMS publishes per contract, using the column for two years before the star year and, for the Health Outcomes Survey measures, which the Notes key one year earlier, the column for three years before it. CMS prints the thresholds of the reward factor, a rating adjustment defined in the Notes, as six-decimal roundings of attained values, and we compare at that precision. For example, a contract whose exact mean is 137/37 must be treated as equal to the printed threshold 3.702703.

C.5 HCAHPS cut points in the following quarter

Each quarterly HCAHPS refresh is accompanied by an edition of the technical notes that lists the cut points used for that refresh (Section 5.1). For the four refreshes whose matching edition we obtained, all 32 printed cut points equal the cut points implied by the public file. Where a published cut point differs from the corresponding cut point of the exact optimum, we compare the next edition's cut point, computed on the next quarter's data, with that optimal value. Across the five refreshes there are 88 such cut points. Among them, the next edition's value equals the earlier optimal value 16 times, moves toward it 10 times, is unchanged 49 times, and moves away 13 times. In that tally we compare one group of 16 cut points with April 2026 cut points that we derived from the public file, because we have not obtained the April edition document. If we compare that group with the later July 2026 edition instead, the counts become 17, 10, 46, and 15, respectively. Under either pairing the next edition's cut points are computed on newer data. Thus their agreement with the earlier optimum reflects covariation of the two quarters' data and should not be read as a response to that optimum.

C.6 Additional checks

This subsection reports checks on the results of Sections 3 to 5 that the main text states in words, in the order of the main text.

Near-optimal hospital groupings. On the 2026 pre-cap scores of Section 3 the optimum is unique in each peer group, but the criterion is flat around it. The optimum lowers the SSQ of the published grouping by 4.67%, 0.097%, and 0.046% in peer groups 3, 4, and 5, and the second-best set of cut points, which moves one or two hospitals, lies within 0.026%, 0.004%, and 0.00008% of the optimal SSQ. For each hospital we also computed the smallest increase over the optimal SSQ at which some set of five cut points assigns it a different star. For 41, 40, and 88 of the 55, 68, and 90 reassigned hospitals in the three peer groups, a set of cut points strictly closer

to the optimum than the published grouping already returns the published star. Within the published grouping's own distance from the optimum a further 45, 22, and 75 hospitals whose published star agrees with the optimum could be moved as well. The differences reported in Section 3 are therefore those of the unique minimizer, not of every near-minimal grouping.

Fitted profiles and the Medicare Advantage results. The optimality gaps of Section 4.1 do not rest on the fitted profiles of Appendix C.2. Among the 82 computations whose sample the documented rules reproduce with nothing fitted, the Ward grouping misses the minimum in 65 and attains it in 14 (the other 3 are the degenerate computations); among the 41 that required fitted profiles it misses the minimum in 39. With every fitted profile and removal dropped, it misses in 103 of the 123. A fitted profile never receives a star: every measure star that changes in Section 4.2 belongs to a real contract's own published score, and the profiles act only through the cut points. When every cut point is recomputed on the documented-rules sample with no fitted profile and no removal, all 60 bonus-status changes of Section 4.2, the 34 valued ones included, remain at the published half-star precision on both bases, although 11 of the 129 changed measure stars behind the 34 revert, and two further downward crossings appear that we do not count. In particular, all six of H0524's measure-star changes and both of H0473's occur in computations whose sample the documented rules reproduce with nothing fitted, and both ratings are unchanged when every fitted profile and removal is dropped.

The enumeration of Appendix C.2 takes one member of each class of fitted profiles. A second search then varied the fitted values within the classes adversarially: 125,063 alternative profiles over the 41 fitted computations, swept once in the primary sample and again in two alternative reconstructions (2024 with no removals; 2025 with the forced removal carried out on the other eligible contracts), 58 sweeps in all, each profile reproducing both published fences exactly. Of the 34 valued flips, 20 keep their changed bonus status under every configuration we tested, among them H0524 and H0473. Each of the other 14 (two in 2024, nine in 2025, three in 2026) can be undone by some fence-exact choice of the fitted values, one of them only on the alternative 2024 reconstruction and three only by joint choices in two computations. In every such case the fitted scores fall inside the trimmed sample of a Members Choosing to Leave, Complaints, Diabetes Care or Reducing the Risk of Falling computation and move its optimal cut points by one to several display units. The 20 that survive carry 1,688,688 of the 2,291,034 valued enrollee-years and \$675–1,013 million of the \$916–1,375 million gross band of Section 4.2. The profiles searched are those our generator produced, not every admissible profile.

Improvement stars on the recomputed side. Convention (1) of Section 2 holds the two improvement measures at their published stars, and Section 4.2 gives the sensitivity of the published ratings to a change in those stars. In the recomputed ratings the same freedom acts in both directions. When the two improvement stars range over every value from one to five, 15 of the 34 valued changes cannot be reversed by any combination and 23 survive every single one-star change; the other 19 are reversible by some combination, 8 of them only by changes of two or more star steps. The same variation could carry up to 155 further contract-years across the bonus line under the valuation conventions, 43 of them by a single one-star change. These are possibility bounds, since the improvement scores under the optimal cut points are not computable from the public files.

Reward-factor thresholds. On the counterfactual side of Section 4.2 we hold the reward-factor thresholds at their printed values. They are percentiles of the contract-level rating means and weighted variances (the 65th and 85th, and the 30th and 70th), so changed measure stars move them. Our recomputation of the printed tables from the published stars is exact for the stand-alone Part D column of 2026 but not for the MA columns, whose population we could not identify from the public files; we therefore shift each printed threshold by the movement that the counterfactual stars induce in our recomputed percentile (149 of the 160 printed cells move). Under that treatment 32 of the 34 valued contract-years of Section 4.2 remain, two leave (one of them a zero-enrollment row) and four enter, and the gross band moves from \$916–1,375 million to \$1,003–1,505 million; under two other treatments of the thresholds 30 and 27 of the 34 remain and the band again rises. Holding the thresholds fixed therefore does not enlarge the estimate.

Contracts affected by disasters in two consecutive years. Convention (6) of Section 2 keeps a disaster-affected contract's measure star at or above its prior-year published star. For contracts affected in two consecutive years the Notes compare instead against what the prior-year star would have been without that year's own disaster adjustment; resolving that star from the earlier public files changes none of the 34 valued contract-years of Section 4.2 and could add at most one further change.

Direct comparison of the two resampled computations. The two expected payments at stake of Section 4.3, one with Ward's method inside the ten resampling groups and one with the exact optimum inside them, measure each method's own variation with the input order, not the difference between the methods. Compared directly at the same order, the two assign different bonus status to between 26 and 47 valued contract-years per order (median 34.5) over the 28 sampled orders. Six of them, with 375,095 enrollees and \$187.5 million in gross payments at the midpoint, differ at every one of the 28, and four of those six are among the 34 valued full-sample flips of Section 4.2.

Input order and tied merges in the survey files. The HCAHPS scores of Section 5.1 are integers with 24 to 37 distinct values per measure, so Ward's method meets tied merges and its grouping can depend on the input order. For the May 2026 HCAHPS refresh we reran Ward's method on the file's starred scores under 68 input orderings per measure and under every resolution of the tied merges. One of the eight published groupings is recovered, under 39 of the 68 orderings, and none of the other seven under any ordering or tie resolution, the closest of the seven differing from the published grouping at 315 of 3,176 hospitals. On the dialysis file the same orderings recover all six published groupings identically under some ordering: four under every one of the 68, and the two whose tied merges depend on the order (one has only 46 distinct scores among the 2,594 facilities) under 40 and 11 of them.

References

[1] Jeannie Fuglesten Biniek, Nancy Ochieng, and Meredith Freed. Medicare will spend more than \$13 billion on the Medicare Advantage quality bonus program in 2026. KFF, July 2026. `https://www.kff.org/medicare/medicare-will-spend-more-than-13-billion-on-the-medicare-advantage-quality-bonus-program-in-2026/`, accessed 2026-07-18.

[2] Code of Federal Regulations. 42 C.F.R. § 422.162: Medicare Advantage Quality Rating System. https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-B/part-422/subpart-D/section-422.162, . Definition of clustering at paragraph (a). Accessed 2026-08-26.

[3] Code of Federal Regulations. 42 C.F.R. § 422.166: Calculation of Star Ratings. https://www.ecfr.gov/current/title-42/section-422.166, . Accessed 2026-07-18.

[4] Centers for Medicare & Medicaid Services. Medicare 2024 Part C & D Star Ratings Technical Notes. Technical report, 2024. Updated 03/13/2024. https://www.cms.gov/files/document/2024-star-ratings-technical-notes.pdf.

[5] Centers for Medicare & Medicaid Services. Medicare 2025 Part C & D Star Ratings Technical Notes. Technical report, 2024. Updated 10/03/2024. https://www.cms.gov/files/document/2025-star-ratings-technical-notes.pdf.

[6] Centers for Medicare & Medicaid Services. Medicare 2026 Part C & D Star Ratings Technical Notes. Technical report, 2025. Updated 09/25/2025. https://www.cms.gov/files/document/2026-star-ratings-technical-notes.pdf.

[7] Yale New Haven Health Services Corporation / Center for Outcomes Research & Evaluation. CMS Overall Hospital Quality Star Rating: Comprehensive Methodology Report (v5.1). Report prepared for the Centers for Medicare & Medicaid Services, April 2026. https://qualitynet.cms.gov/files/69eb9be1f1becb2be7183b08?filename=Star_Rtngs_CompMthdlgy_v5.1.pdf.

[8] Centers for Medicare & Medicaid Services. Overall Hospital Quality Star Ratings: 2025 SAS Pack Software Documentation. QualityNet, July 2025. Distributed with the star-rating SAS package (SAS_Pack_2025Jul.zip) and its input file; the 2021–2024 rating years' SAS packages are distributed the same way. https://qualitynet.cms.gov/files/688d19c02844d4a92b85f900?filename=StarRatings_SASPackDoc_Jul2025.pdf.

[9] Centers for Medicare & Medicaid Services. Overall Hospital Quality Star Ratings: 2026 R Pack Software Documentation. QualityNet, April 2026. Distributed with the 2026 star-rating R package (R_Pack_2026Apr.zip) and its input file.

[10] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009. https://doi.org/10.1007/s10994-009-5103-0.

[11] Walter D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958. https://doi.org/10.1080/01621459.1958.10501479.

[12] Haizhou Wang and Mingzhou Song. Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29–33, 2011. https://doi.org/10.32614/RJ-2011-015.

[13] Hayley Rogers and Matthew Smith. Recalculating Medicare Advantage: Potential SCAN and Elevance ruling implications for MA stakeholders. White paper, Milliman, June 2024.

https://www.milliman.com/en/insight/recalculating-medicare-advantage-scan-elevance-ruling-implications, accessed 2026-07-18.

[14] Michael Darden and Ian M. McCarthy. The star treatment: Estimating the impact of star ratings on Medicare Advantage enrollments. *Journal of Human Resources*, 50(4):980–1008, 2015. https://doi.org/10.3368/jhr.50.4.980.

[15] Andrew Anderson and Mark K. Meiselbach. Fluctuating Star Ratings and Medicare Advantage bonuses. *JAMA Health Forum*, 6(10):e254398, 2025. https://doi.org/10.1001/jamahealthforum.2025.4398.

[16] Benjamin Vatter. Quality disclosure and regulation: Scoring design in Medicare Advantage. *Econometrica*, 93(3):959–1001, 2025. https://doi.org/10.3982/ECTA21182.

[17] Bo Bayles. The fault is not in our stars, but in ourselves. Web post, https://www.bbayles.com/index/cms-star-ratings, June 2017. Accessed 2026-07-18.

[18] Ren-Huai Huang. rstarating: CMS Hospital Compare Star Rating SAS Pack Replica. Software repository, https://github.com/huangrh/rstarating, 2017. Accessed 2026-07-18.

[19] Matthew E. Barclay, Mary Dixon-Woods, and Georgios Lyratzopoulos. Concordance of hospital ranks and category ratings using the current technical specification of US hospital star ratings and reasonable alternative specifications. *JAMA Health Forum*, 3(5):e221006, 2022. https://doi.org/10.1001/jamahealthforum.2022.1006.

[20] Centers for Medicare & Medicaid Services. Care Compare hospital data: archived dataset snapshots. Provider Data Catalog, data.cms.gov, 2021–2026. Dataset-archive snapshots, one per rating year, e.g. https://data.cms.gov/provider-data/sites/default/files/dataset-archives/theme/hospitals/hospitals_2026-05-13.zip for the 2026 comparison.

[21] SAS Institute Inc. *SAS/STAT 14.1 User's Guide: The FASTCLUS Procedure*. SAS Institute Inc., Cary, NC, 2015. https://support.sas.com/documentation/onlinedoc/stat/141/fastclus.pdf.

[22] SAS Institute Inc. *SAS/STAT 14.2 User's Guide: The CLUSTER Procedure*. SAS Institute Inc., Cary, NC, 2016. https://support.sas.com/documentation/onlinedoc/stat/142/cluster.pdf.

[23] Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996. https://doi.org/10.1080/00031305.1996.10473566.

[24] Medicare Payment Advisory Commission. The Medicare Advantage program: Status report. Chapter 11 in Report to the Congress: Medicare Payment Policy, March 2025. https://www.medpac.gov/wp-content/uploads/2025/03/Mar25_Ch11_MedPAC_Report_To_Congress_SEC.pdf.

[25] Matthew D. Schwartz. The unwritten rules: rule as published versus rule as run in Medicare's payment formulas. Companion paper, in preparation, 2026.

[26] Centers for Medicare & Medicaid Services. Advance notice of methodological changes for calendar year (CY) 2027 for Medicare Advantage (MA) capitation rates and Part C and Part D payment policies. Technical report, January 2026. Attachment IV, Section K. https://www.cms.gov/files/document/2027-advance-notice.pdf.

[27] Centers for Medicare & Medicaid Services. Announcement of calendar year (CY) 2027 Medicare Advantage (MA) capitation rates and Part C and Part D payment policies. Technical report, April 2026. Attachment VI, Section K. https://www.cms.gov/files/document/2027-announcement.pdf.