UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

ELEVANCE HEALTH, INC., *et al.*,

Plaintiffs,

v.

XAVIER BECERRA, *Secretary of Health and Human Services, et al.*,

Defendants.

Civil Action No. 23-3902 (RDM)

MEMORANDUM OPINION AND ORDER

Elevance Health, Inc. (formerly, Anthem Inc.) and six of its affiliated entities[1] (collectively, "Plaintiffs" or "Elevance") bring this suit under the Administrative Procedure Act ("APA"), 5 U.S.C. § 551 *et seq.*, to challenge the methodology that the Centers for Medicare and Medicaid Services ("CMS") used to evaluate the health plans that Plaintiffs offer to Medicare beneficiaries. Every year, CMS rates certain Medicare insurance plans on a one-to five-star scale to reflect the quality of care and services the plans offer. Plaintiffs contend that the way that CMS assigned "star ratings" to certain plans for 2024 was contrary to the agency's own regulations and was therefore arbitrary and capricious.

The parties have cross-moved for summary judgment on these claims. For the reasons explained below, the Court will **GRANT** in part and **DENY** in part Plaintiffs' motion for

[1] Elevance's affiliates include AMH Health, LLC; Anthem Healthchoice HMO, Inc.; Anthem Health Plans, Inc.; Anthem Insurance Companies, Inc.; Blue Cross Blue Shield Healthcare Plan of Georgia, Inc.; Community Care Health Plan of Louisiana, Inc.; Freedom Health, Inc; and Healthkeepers, Inc. Dkt. 23 at 4.

summary judgment, Dkt. 15, and will **DENY** Defendants' cross motion for summary judgment, Dkt. 17.

I. BACKGROUND

A. Statutory and Regulatory Background

The Medicare program provides healthcare for the elderly and disabled. *See* 42 U.S.C. § 1395 *et seq*. It is administered by CMS, a component within the U.S. Department of Health and Human Services. *See Johnson v. Becerra*, 668 F. Supp. 3d 14, 17 (D.D.C. 2023). Parts A and B of the program make up the traditional Medicare system under which CMS directly reimburses healthcare providers. 42 U.S.C. §§ 1395c, 1395j. Parts C and D of the program, in contrast, permit individuals to receive their Medicare benefits through private insurers. Part C, also known as the Medicare Advantage ("MA") program, permits Medicare beneficiaries to enroll in private health insurance plans for purposes of receiving their healthcare. *Id.* § 1395w–21(a)(1). Part D offers subsidized prescription drug insurance coverage (also known as prescription drug plans or "PDPs") to beneficiaries who enroll in traditional or Part C plans. *Id.* § 1395w–101(a)(1).

Companies that offer Medicare beneficiaries Medicare Advantage insurance plans are compensated on a per beneficiary basis. The amount the insurer receives varies depending in part on the demographic and health characteristics of each beneficiary and in part on the performance of the healthcare plan. *See generally UnitedHealthcare Ins. Co. v. Becerra*, 16 F.4th 867, 873 (D.C. Cir. 2021) (describing the statutory and regulatory framework that governs how CMS determines the payments Medicare Advantage organizations ("MAOs") receive for each Part C enrollee). Every year, CMS rates these insurance plans on a one-to-five-star scale as part of its Star Ratings program. 42 U.S.C. § 1395w–23(o). Plaintiffs, insurers that offer Part C

and Part D insurance coverage, bring this action to challenge the star ratings that they received for the 2024 contract year. *See generally* Dkt. 13 (Am. Compl.).

1. *Overview of the Star Rating Program*

In the fall of each year, CMS rates each Medicare Advantage and each PDP insurance plan on a scale from one to five stars. These "Star Ratings" are intended to reflect plan quality and performance; a five-star rating is the highest, and a one-star rating is the lowest. As CMS explains:

> The [Medicare Advantage] Star Ratings system is designed to provide information to the beneficiary that is a true reflection of the plan's quality and encompasses multiple dimensions of high quality care. . . . While encouraging improved health outcomes of beneficiaries in an efficient, person centered, equitable, and high quality manner is one of the primary goals of the ratings, they also provide feedback on specific aspects of care and performance that directly impact outcomes, such as process measures and the beneficiary's perspective. The ratings focus on aspects of care and performance that are within the control of the health plan and can spur quality improvement.

Contract Year 2019 Policy & Technical Changes to the Medicare Advantage Program, 83 Fed. Reg. 16440, 16520 (Apr. 16, 2018) (hereafter "Contract Year 2019 Final Rule"). Star ratings are released each October and apply to the upcoming contract year. Thus, as relevant here, the star ratings released in October 2023 are the star ratings applicable to the 2024 contract year ("2024 Star Ratings").

The Star Ratings program was originally introduced as a tool to better inform beneficiaries about the health insurance plans available to them through the Medicare Advantage program. Consistent with that purpose, CMS maintains a website known as the Medicare "Plan Finder," which displays information about available plans, including the plan's star ratings, to assist beneficiaries in choosing their coverage. *See* 42 C.F.R. § 422.166(h). Over the years, however, the Star Ratings program has evolved so that it now also affects the compensation that

high- and low-scoring plans receive under the Medicare program. *See* Contract Year 2019 Final Rule, 83 Fed. Reg. at 16520 (explaining that the Patient Protection and Affordable Care Act, as amended by the Healthcare and Education Reconciliation Act, provides for "quality ratings[] based on a 5-star rating system and [other information] to be used in calculating payment to MA organizations beginning in 2012").

In its most recent iteration, the Star Ratings program rewards high-performing plans by increasing the per-beneficiary payment that the plan is eligible to receive. Understanding how this works requires a brief primer on the Medicare Advantage program. To determine how much each Medicare Advantage insurer will receive each year for each beneficiary, CMS uses a bidding system. Under that system, CMS establishes yearly "benchmark" rates that represent the maximum CMS will pay a Medicare Advantage Organization to cover an average beneficiary in a given area. *See* 42 U.S.C. § 1395w–23(b)(1)(B), (n); 42 C.F.R. § 422.258. Plans then submit "bids" to CMS, representing what payment the insurer will accept to cover a beneficiary with an average risk profile in that area in the coming year. 42 U.S.C. § 1395w–23(a)(1)(B); 42 C.F.R. § 422.254. If a plan bids below the benchmark, then that bid becomes the plan's base payment (*i.e.*, the amount the MAO will be paid to cover a beneficiary of average risk) and CMS will return a portion of the savings to the plan as a "rebate," which the plan can use to fund additional benefits or reduce premiums. 42 U.S.C. § 1395w–23(a)(1)(E); *id.* § 1395w–24(b)(1)(C). If the plan bids above the benchmark, then the benchmark becomes the plan's base payment and the MAO must charge beneficiaries a premium to make up the difference. 42 U.S.C. § 1395w–23(a)(1)(B)(ii); § 1395w–24(b)(2)(A).

Under the Star Ratings program, high-scoring Medicare Advantage plans can receive important benefits, while low-scoring plans can suffer significant losses. As an initial matter,

higher scoring plans are likely to attract additional customers, while lower scoring plans may lose customers. But beyond those effects on the marketplace, the Star Rating program bears on the payments that plans receive from CMS. For one thing, plans that earn a rating of four stars (or higher) qualify for an increased benchmark against which to bid in the following contract year. 42 U.S.C. § 1395w–23(o)(1) (increasing, for qualifying plans, the applicable percentage that calculates the benchmark); *id*. § 1395w–23(o)(3)(A)(i) (a qualifying plan is one that earns a rating of four stars or higher). For another, star ratings can affect the rebate amount an insurer receives. Plans that earn a rating of four-and-a-half stars or higher receive a rebate of seventy percent of the difference between their bid and the benchmark, whereas plans that earn at least three-and-a-half but less than four-and-a-half stars receive a rebate of sixty-five percent of that difference. Plans that earn less than three-and-a-half stars are eligible for a rebate of fifty percent of that difference. 42 U.S.C. § 1395w–24(b)(1)(C)(v) (listing the "[f]inal applicable rebate percentage[s]" by rating); 42 C.F.R. § 422.266(a)(2)(ii) (same). Most severely, plans that receive consistently low star ratings—that is, a rating of less than three stars for three years in a row—of a certain type can be terminated by CMS. *See* 42 C.F.R. § 422.510(a)(4)(xi); *AvMed, Inc. v. Becerra*, No. 20-cv-3385, 2021 WL 2209406, at *1 (D.D.C. June 1, 2021); Contract Year 2019 Final Rule, 83 Fed. Reg. at 16520.

When the Star Ratings program was first introduced, it was not codified in the Code of Federal Regulations. Instead, CMS used its "authority to disseminate information to beneficiaries as the basis for developing and publicly posting the 5-star ratings system." Contract Year 2019 Final Rule, 83 Fed. Reg. at 16520. The statutory provisions governing the Medicare Advantage and Part D programs both require CMS to provide information about "plan quality and performance" to beneficiaries, and the Star Ratings program was developed to

accomplish this goal. *Id.* In 2017, however, CMS decided to formalize the Star Ratings program and to "improve transparency" within the program by codifying in a final rule the methodology that CMS uses each year to calculate the ratings. Contract Year 2019 Policy & Technical Changes to the Medicare Advantage Program, 82 Fed. Reg. 56336, 56376 (Nov. 28, 2017) (hereafter "Contract Year 2019 Proposed Rule"). Recognizing that "the rulemaking process [would] create a longer lead time for changes," CMS nonetheless concluded that it was worthwhile to "codify[] the Star Ratings methodology" because doing so would "provide plans with more stability to plan multi-year initiatives, because they will know the measures several years in advance." *Id.* CMS also concluded that it was "an appropriate time to codify the methodology, because the rating system ha[d] been used for several years now and [was] relatively mature so there [was] less need for extensive changes every year." *Id.*

Since then, the methodology that CMS employs to calculate each year's star ratings has changed, with each change going through a formal rulemaking process before being implemented. In its present form, the Star Ratings program scores each insurance plan or contract that offers benefits under Medicare Parts C or D on several dozen health and drug plan quality and performance measurements ("measures"). There are currently 30 measures for MA plans and 12 measures for Part D plans. These 42 measures fall into five categories: (1) outcome measures, which "reflect improvements in a beneficiary's health and are central to assessing quality of care;" (2) intermediate outcome measures, which "reflect actions taken [that] can assist in improving a beneficiary's health status;" (3) patient experience measures, which "reflect beneficiaries' perspectives of the care they received;" (4) access measures, which "reflect processes and issues that could create barriers to receiving needed care;" and (5) process measures, which "capture the health care services provided to beneficiaries [that] can assist in

maintaining, monitoring, or improvement their health status." CMS, *Medicare 2024 Part C & D Star Ratings Technical Notes* 9 (updated Mar. 13, 2024) [hereinafter *2024 Technical Notes*], cms.gov/files/document/2024-star-ratings-technical-notes.pdf.

Each plan receives a score for each applicable measure, which is then converted by CMS into a measure-specific star rating. These measure-specific star ratings for each plan are then combined into three "groups" or aggregate measures: domain star ratings, summary star ratings, and overall star ratings. *Id*. at 11–12. Domain star ratings "group together measures of similar services" for each plan using the "non-weighted average [s]tar [r]atings of the included measures." *Id.* at 11.[2] Summary star ratings group together the measure-specific star ratings applicable to MA plans and group together the measure-specific star ratings applicable to Part D ("PD") plans. *Id*. Accordingly, an "MA-PD" plan will receive two summary star ratings: one reflecting the plan's performance across the thirty MA measures and one reflecting the plan's performance across the twelve Part D plan measures. Finally, all plans that provide both MA and Part D benefits receive a single overall star rating, calculated using the weighted average star ratings of all included measures. *Id*. at 11–12.

CMS obtains the data necessary to create the star ratings from different sources, *id.* at 13, and the agency's methodology for calculating star ratings differs depending on the source of the information. Measures based on information collected from Consumer Assessment of

[2] There are five domain star ratings for MA-only plans and four domain star ratings for PD-only plans. The 5 domains for the MA Star Ratings are: (1) Staying Healthy: Screenings, Tests and Vaccines; (2) Managing Chronic (Long Term) Conditions; (3) Member Experience with Health Plan; (4) Member Complaints and Changes in the Health Plan's Performance; and (5) Health Plan Customer Service. The 4 domains for the Part D Star Ratings are: (1) Drug Plan Customer Service; (2) Member Complaints and Changes in the Drug Plan's Performance; (3) Member Experience with the Drug Plan; and (4) Drug Safety and Accuracy of Drug Pricing. 42 C.F.R. § 422.166(b)(1)(ii).

Healthcare Providers and Systems ("CAHPS") surveys[3] are calculated using relative distribution and significant testing methodology to "account [for] differences in the characteristics of enrollees across contracts that may potentially impact survey responses." *Id.* at 158; *see* 42 C.F.R. § 422.166(a)(3). These are referred to as "CAHPS measures." Non-CAHPS measures, in contrast, are calculated using a clustering methodology. *2024 Technical Notes* at 148. Because the parties' arguments focus on the methodology that CMS uses to convert individual measure-specific scores into individual measure-specific star ratings for non-CAHPS measures, *see* Dkt. 13 at 4 (Am. Compl. ¶ 3); Dkt. 17 at 6–7, the Court will provide additional background on that part of this process.

2. *Calculating Star Ratings for Non-CAHPS Measures*

To assign each plan a star rating for each measure, CMS creates five non-overlapping ranges of scores that are associated with each measure. To illustrate how this works in practice, consider the star ratings for the "Medication Adherence for Diabetes Medication" measure contained in the following table:

Table 1: 2024 Star Ratings for MA-PD plans
Medication Adherence for Diabetes Medications

1 Star	2 Stars	3 Stars	4 Stars	5 Stars
< 80%	80% to < 85%	85% to < 87%	87% to < 91%	≥ 91%

2024 Technical Notes at 149. Under this distribution, a plan with a score of 86% for 2024 would receive a rating of three stars that year, while a plan with a score of 91% would receive a rating

[3] CAHPS surveys "ask consumers and patients to evaluate the interpersonal aspects of health care" to "probe those aspects of care for which consumers and patients are the best and/or only source of information, as well as those that consumers and patients have identified as being important." *Id.* at 191.

of five stars. The point at which a score results in a higher or lower star rating is referred to as a "cut point." *Id*. at 148. In this example, the cut points are 91%, 87%, 85%, and 80%. When referring to a cut point for a specific measure (for instance, one that distinguishes between a three-star rating and a four-star rating for Medication Adherence for Diabetes Medications), CMS uses the term "measure-threshold-specific cut point" (in this example, 87%). *See, e.g.*, 42 C.F.R. § 422.162(a).

To identify the cut points for each non-CAHPS measure, CMS uses a clustering methodology. "Conceptually, the clustering algorithm identifies the natural gaps that exist within the distribution of the scores and creates groups (clusters) that are then used to identify the cut points that result in the creation of a pre-specified number of categories." *2024 Technical Notes* at 148. CMS does not require the same number of contracts to be within each group, *id.*; rather, "[t]he scores are grouped such that scores within the same Star Rating category are as similar as possible, and scores in different categories are as different as possible," *id.*; *see* 42 C.F.R. § 422.166(a)(2)(i). That is, CMS looks for naturally occurring clusters.

For the 2024 Star Ratings, the clustering methodology that CMS used to calculate the cut points for each measure proceeded in three steps: (1) first, CMS removed certain outlier scores, referred to as "Tukey outer fence outlier scores;" (2) then, CMS used mean resampling followed by hierarchal clustering to create a set of tentative cut points; (3) and finally, CMS applied a cap that prevents the cut points from significantly varying from year to year, referred to as "the guardrail."

a.

The first step in the process—the Tukey outlier deletion—"involves removing Tukey outer fence outlier contract scores." *2024 Technical Notes* at 150. Tukey outlier deletion is a

method used to identify and to remove extreme outliers in a dataset. The method defines outliers as scores that are below a certain point or above a certain point and removes those scores, but the method does not remove a score unless it meets the definition of a "true outlier." Contract Year 2021 and 2022 Policy & Technical Changes to the Medicare Advantage Program, 85 Fed. Reg. 9002, 9044 (Feb. 18, 2020) (hereinafter "Contract Year 2021 & 2022 Tukey Proposed Rule"). The definition that CMS employs for "true" outliers is "measure-specific scores outside the bounds of 3.0 times the measure-specific interquartile range subtracted from the 1st quartile or added to the 3rd quartile." *2024 Technical Notes* at 150. For the 2024 Star Ratings, CMS removed the Tukey outer fence outlier contract scores for non-CAHPS measures first before it conducted the mean resampling and hierarchical clustering analysis. Dkt. 15-1 at 7.

b.

The second step—mean resampling followed by hierarchal clustering—is the most important step in the process. Mean resampling separates the "measure-specific scores for the current year[] . . . into 10 equal-sized groups, using a random assignment process." *2024 Technical Notes* at 150. The hierarchal clustering algorithm, using those ten equal-sized groups, creates ten sets of four cut points for each measure.[4] For each threshold, that is, for each one of

[4] More specifically, the hierarchal clustering algorithm consists of four steps. First, CMS calculates the individual distances (specifically, the absolute value of the difference) between each contract's measure scores. This creates a matrix with all of the plans listed on each axis, so that there is a value representing the distance between each contract's score and every other contract's score. The matrix is the input to the next step of the hierarchal clustering algorithm. In that step, CMS starts with each contract as its own "cluster." Then, CMS merges two clusters at a time, each time trying to keep the most similar scores together in a cluster and those that are most different in separate clusters, until the algorithm produces one "single cluster containing all contracts' results." *Id.* at 151. At the conclusion of this step, CMS has created a "tree-like structure of cluster assignments, from which any number of clusters between 1 and the number of contract measure scores" could be derived. *Id.* The next step is a logical one: CMS wants five clusters (because it wants five star-rating ranges) so it derives from the tree-like structure it

the four cut points, CMS averages the ten cut points to arrive at a single measure-threshold-specific cut point. The set of four average cut points that emerges are the pre-guardrail-adjusted cut points for that year.

c.

The final step in the process involves application of the guardrail to the cut points that are established through mean resampling and hierarchal clustering. The guardrail limits the degree to which cut point values for non-CAPHS measures can vary each year. It prevents the current year's cut points from varying by more than five percent from the prior year's cut points. Mechanically, the guardrail does this in one of two ways, depending on whether the measure is scored on a 0 to 100 scale.

For those measures scored on a 0 to 100 scale—that is, the score that a plan receives is a number between 0 and 100 for a given measure—the guardrail applies "an absolute percentage cap of 5 percentage point[s]." *2024 Technical Notes* at 156; *see* 42 C.F.R. § 422.162(a) (defining "absolute percentage cap" as "a cap applied to non-CAHPS measures that are on a 0 to

---

produced, five ranges of measures scores. Finally, the last step of the hierarchal clustering algorithm is to identify the thresholds between the star ratings, *i.e.*, to determine the cut points between each star rating. That cut point will be the lower bound of the range if the measure is one where a higher score is better, or it will be the higher bound of the range if the measure is one where a lower score is better.

This hierarchal clustering process is repeated ten times, "each time leaving out one of the 10 groups." *Id.* at 150. In other words, each time the hierarchal clustering algorithm is run, it includes only nine of the ten randomly selected groups from the mean resampling. This process produces "10 sets of measure-specific cut points, one for each of the 10 implementations of the hierarchical clustering algorithm." *Id.* at 152. The final set of thresholds (or cut points) are "calculated as the mean cut point for each threshold per measure from the 10 different cut point values." *Id.* In other words, the measure-threshold-specific cut point between a one-star and a two-star rating is the average of the ten thresholds at which a one-star rating becomes a two-star rating.

100 scale that restricts movement of the current year's measure-threshold-specific cut point to no more than the stated percentage as compared to the prior year's cut point"). "If the absolute difference between the current and prior year's cut point is less than or equal to 5 percentage points, the current year's cut point is used as the final cut point value." *2024 Technical Notes* at 156. But

> [i]f the absolute difference . . . is greater than 5 percentage points, a 5 percentage point cap is applied. That is, 5 percentage points are added to or subtracted from the prior year's cut point value (depending on the direction of movement for the cut point value in the current year) to obtain the final cut point for the current year.

Id. at 156–57.

If a measure is not scored on a 0 to 100 scale, CMS applies a restricted range cap. The restricted range is equal to "5 percent of the restricted range," 42 C.F.R. § 422.166(a)(2)(i); *see id.* § 422.162(a) (defining restricted range cap as "a cap applied to non-CAHPS measures that restricts the movement of the current year's measure-threshold-specific cut point to no more than the stated percentage of the restricted range of a measure calculated using the prior year's measure score distribution"), where the restricted range is "the difference between the maximum and minimum score values using the prior year's measure scores excluding outer fence outliers," *id.* § 422.162(a); *see 2024 Technical Notes* at 157 (explaining that "the restricted range cap is equal to the prior year's (maximum score value – minimum score value) * 0.05").

3. *Regulatory History*

When the Star Ratings program was first considered in a formal rulemaking process, the clustering algorithm consisted of only one step: hierarchal clustering. That method, however, was viewed as imperfect because it did not account for outlier scores, and it did not provide sufficient stability for the participating Medicare insurers. To address these concerns, CMS

made three significant changes to the clustering methodology that resulted in the process outlined above.

The first two changes occurred in the agency's rulemaking for the 2020 and 2021 contract years. As part of that rulemaking, CMS introduced mean resampling and it added the guardrail. Mean resampling was added to "attenuate the effect of outliers." Policy & Technical Changes to the Medicare Advantage Programs for Years 2020 and 2021, 84 Fed. Reg. 15680, 15754 (Apr. 16, 2019) (hereafter "Contract Years 2020 & 2021 Guardrail Final Rule"); *id.* at 15753 (explaining that mean resampling would "reduce[] the sensitivity of the clustering algorithm to outliers and reduce[] the random variation that contributes to fluctuations in cut points, and, therefore, improve[] the stability of the cut points over time"). The guardrail was added for a similar purpose: its aim was "to provide stability for cut points from year to year" while still "creat[ing] incentives for quality improvement." Contract Year 2019 Final Rule, 83 Fed. Reg. at 16569. As CMS explained in the final rule adopting this change:

> To increase the predictability of the cut points, we also proposed a second enhancement to the clustering algorithm: A guardrail for measures that have been in the Part C and D Star Ratings program for more than 3 years. We proposed a guardrail of 5 percent to be a bi-directional cap that restricts movement both above and below the prior year's cut points. A 5 percent cap restricts the movement of a cut point by imposing a rule for the maximum allowable movement per measure threshold; thus, it allows a degree of predictability. The trade-off for the predictability provided by bi-directional caps is the inability to fully keep pace with changes in performance across the industry. While cut points that change less than the cap would be unbiased and keep pace with changes in the measure score trends, changes in overall performance that are greater than the cap would not be reflected in the new cut points. A cap on upward movement may inflate the measure-level Star Ratings if true gains in performance improvements cannot be fully incorporated in the current year's ratings. Conversely, a cap on downward movement may decrease

> the measure-level Star Ratings since the ratings would not be adjusted fully for downward shifts in performance.

Contract Years 2020 & 2021 Guardrail Final Rule, 84 Fed. Reg. at 15754.

The third change, which added the Tukey outlier deletion methodology, was made in a subsequent rulemaking. By way of background, in response to the agency's notice of proposed rulemaking regarding contract years 2020 and 2021, some commenters expressed concern that "resampling would not be sufficient to address outliers" or that "resampling [would] not directly address year to year changes in cut points," and "[a] couple commenters supported removing outliers before clustering." *Id.* at 15755. At that time, however, CMS declined to finalize "a method to directly remove outliers prior to clustering" because the "methods were not included in the proposed rule," and thus "the public [did] not [have] an opportunity to comment on them specifically." *Id*. at 15756. Instead, CMS undertook to "continue to evaluate these and possibly other methods to directly address outliers and" explained that it would "consider proposing outlier deletion in [a] future rulemaking." *Id.*

Then, in February 2020, CMS issued a notice of proposed rulemaking, which proposed "to further increase the stability of cut points by modifying the cut point methodology for non-CAHPS measures through direct removal of outliers" using Tukey outer fence deletions. Contract Years 2021 & 2022 Tukey Proposed Rule, 85 Fed. Reg. 9002, 9043 (Feb. 18, 2020). In addition to creating increased stability, CMS noted that the new methodology would result in substantial savings to the Medicare program. When CMS ran simulations to see the effect that implementing the Tukey outlier deletion methodology would have on star ratings, it found:

> In general, there tend to be more outliers on the lower end of measure scores. As a result, the 1 to 2 star thresholds often increased in the simulations when outliers were removed compared to the other thresholds which were not as impacted. The effect of Tukey outlier deletion would create a savings of $808.9 million for 2024, increasing to $1,449.2 million by 2030.

Id. at 9044.

Of particular relevance here, when CMS proposed the addition of the Tukey outlier deletion methodology, it explained how it intended to implement the change. It wrote: "In the first year that [the change] would be implemented, the prior year's thresholds would be rerun, including mean resampling and Tukey outer fence deletion so that the guardrails would be applied such that there is consistency between the years." *Id*. That is, the agency planned to rerun the clustering analysis for the prior year using the Tukey outer fence deletion methodology so that it would create a set of simulated Tukey-adjusted cut points for the prior year.

In June 2020, CMS adopted the Tukey outlier deletion methodology in a final rule. Contract Year 2021 Policy & Technical Changes to the Medicare Advantage Program, 85 Fed. Reg. 33796, 33832–36 (June 2, 2020) (hereafter "Contract Year 2021 Tukey Final Rule"). In the final rule, CMS once again described how it would apply the change in the first year of implementation (applicable to the 2024 Star Ratings): it would "rerun the prior year's thresholds [that is, the cut points for the 2023 Star Ratings,] using mean resampling and Tukey outer fence deletion so that the guardrails would be applied such that there is consistency between the years." *Id.* at 33835; *see also id*. at 33833.

Consistent with these statements, when CMS removed Tukey outliers for the first time in the 2024 Star Ratings calculations, the agency reran the prior year's cut point calculations, using the Tukey outlier deletion methodology. As CMS explained in its *2024 Technical Notes*:

> When we proposed and finalized Tukey outlier deletion . . . , we described that in the first year of adding Tukey outlier deletion, the prior year's thresholds would be rerun, including mean resampling and Tukey outer fence deletion so that the guardrails would be applied such that there is consistency between the years. For the purposes of calculating the guardrails for the 2024 Star Ratings, the 2023 Star Ratings cut points were rerun including mean resampling, Tukey

> outlier deletion and no guardrails. These rerun 2023 Star Ratings cut points serve as the basis for applying the guardrails for the 2024 Star Ratings . . . .

2024 Technical Notes at 157. The benefit of rerunning the prior year's cut points, as CMS stated elsewhere, was that the current year's cut points and the prior year's cut points were calculated the same way so that the guardrail could apply in an apples-to-apples manner. *See* Dkt. 17 at 28.

B. Factual and Procedural Background

Elevance takes issue with CMS's decision to recalculate the 2023 cut points as part of the 2024 Star Ratings calculations. It argues that because CMS used simulated 2023 cut points instead of the actual 2023 cut points, several of Plaintiffs' insurance plans received lower star ratings than Plaintiffs believe they otherwise would have received. Dkt. 23 at 24; *see* Dkt. 23-1 (Abernathy Decl.). Elevance maintains, for example, that its overall star rating would have been 3.5 stars rather than 3 stars had the actual 2023 cut points been used for the 2024 guardrails. Dkt. 23-1 at 10–11 (Abernathy Decl. ¶¶ 27–28).

On December 29, 2023, Plaintiffs filed this lawsuit under the APA challenging the legality of CMS's 2024 Star Ratings calculations. Dkt. 1. Plaintiffs allege that CMS failed to follow its own rules in calculating the cut points for the 2024 Star Ratings year. *See* Dkt. 13 (Am. Compl.). More precisely, they maintain that "by calculating [their] 2024 Star Ratings [based on] simulated 2023 cut points using the Tukey statistical methodology and creating cut points that were more than 5 percentage points higher than the actual 2023 cut points," CMS acted contrary to law and arbitrarily and capriciously. *Id.* at 26 (Am. Compl.). Among other relief, they seek an order requiring the agency "to recalculate [their] Star Ratings without using the Tukey statistical methodology and by using actual 2023 Star Ratings." *Id.*

On March 8, 2024, Plaintiffs moved for summary judgment on their claims, Dkt. 15, and on March 29, 2024, CMS cross-moved for summary judgment, Dkt. 17. Both parties have

requested expedited consideration of their competing motions for summary judgment in light of Plaintiffs' impending deadline to submit their bids to CMS for the upcoming contract year. The Court granted that request, *see* Min. Order (Feb. 29, 2024), and held oral argument on the parties' cross-motions on May 29, 2024.

II. LEGAL STANDARD

"[W]hen a party seeks review of agency action under the APA . . . , the district judge sits as an appellate tribunal." *Rempfer v. Sharfstein*, 583 F.3d 860, 865 (D.C. Cir. 2009) (quoting *Am. Bioscience, Inc. v. Thompson*, 269 F.3d 1077, 1083 (D.C. Cir. 2001)). The general standard for summary judgment set forth in Rule 56 of the Federal Rules of Civil Procedure does not apply to a review of agency action. But summary judgment nonetheless "serves as the mechanism for deciding, as a matter of law, whether the agency action is supported by the administrative record and otherwise consistent with the APA standard of review." *Sierra Club v. Mainella*, 459 F. Supp. 2d 76, 90 (D.D.C. 2006) (citing *Richards v. INS*, 554 F.2d 1173, 1177 & n.28 (D.C. Cir. 1977)). In other words, "[t]he entire case on review is a question of law." *Marshall Cnty. Health Care Auth. v. Shalala*, 988 F.2d 1221, 1226 (D.C. Cir. 1993).

Under the APA, a reviewing court shall "hold unlawful and set aside agency action, findings, and conclusions found to be . . . arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law." 5 U.S.C. § 706(2)(A). An agency action will normally be set aside as "arbitrary and capricious" if the agency "has relied on factors which Congress has not intended it to consider, entirely failed to consider an important aspect of the problem, offered an explanation for its decision that runs counter to the evidence before the agency, or is so implausible that it could not be ascribed to a difference in view or the product of agency

expertise." *Motor Vehicle Mfrs. Ass'n of U.S. v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 43 (1983).

An agency must also, of course, "adhere to its own regulations," *Brock v. Cathedral Bluffs Shale Oil Co.*, 796 F.2d 533, 536 (D.C. Cir. 1986), and "an agency action may be set aside as arbitrary and capricious if the agency fails to 'comply with its own regulations,'" *Nat'l Env't Dev. Assoc.'s Clean Air Project v. EPA*, 752 F.3d 999, 1009 (D.C. Cir. 2014) (quoting *Env't, LLC v. FCC*, 661 F.3d 80, 85 (D.C. Cir. 2011)). "Although it is within the power of [an] agency to amend or repeal its own regulations, [an] agency is not free to ignore or violate its regulations while they remain in effect." *U.S. Lines, Inc. v. Fed. Mar. Comm'n*, 584 F.2d 519, 526 n.20 (D.C. Cir. 1978).

That said, an agency has wide latitude to interpret its regulation, and its interpretation is "controlling unless plainly erroneous or inconsistent with the regulation." *Auer v. Robbins*, 519 U.S. 452, 461 (1997) (internal citations and quotation marks omitted); *see also Bowles v. Seminole Rock & Sand Co.*, 325 U.S. 410, 414 (1945). The Court must therefore defer to CMS's interpretation of its regulation "unless an alternative reading is compelled by the regulation's plain language or by other indications of the [agency's] intent at the time of the regulation's promulgation." *Thomas Jefferson Univ. v. Shalala*, 512 U.S. 504, 512 (1994) (internal citations and quotation marks omitted). Finally, although an agency enjoys significant leeway when interpreting its regulation, that interpretation must bear the mark of consistency, for "agency action is arbitrary and capricious if it departs from agency precedent without explanation." *Ramaprakash v. FAA*, 346 F.3d 1121, 1124 (D.C. Cir. 2003).

III. ANALYSIS

A. Standing

Before considering the merits of Elevance's challenges to the 2024 Star Ratings calculations, the Court must first assure itself that Plaintiffs have standing to proceed with their claims. *See Bauer v. Marmara*, 774 F.3d 1026, 1029 (D.C. Cir. 2014) (explaining that standing "can never be forfeited or waived" and "'[w]hen there is doubt about a party's constitutional standing, the court *must* resolve the doubt, *sua sponte* if need be'" (alteration and emphasis in original) (quoting *Lee's Summit v. Surface Transp. Bd.*, 231 F.3d 39, 41 (D.C. Cir. 2000))). The "'irreducible constitutional minimum' of standing" requires that Plaintiffs demonstrate that they have "(1) suffered an injury in fact, (2) that is fairly traceable to the challenged conduct of the defendant, and (3) that is likely to be redressed by a favorable judicial decision." *Spokeo, Inc. v. Robins*, 578 U.S. 330, 338 (2016). Under the first element, injury-in-fact, a plaintiff's complained-of injury must be "concrete and particularized" and "actual or imminent, not conjectural or hypothetical." *Lujan v. Defs. of Wildlife*, 504 U.S. 555, 560 (1992) (internal quotation marks and citations omitted). Under the second element, causation, the injury must be "fairly traceable to the challenged action of the defendant, and not the result of the independent action of some third party." *Id.* at 560–61 (internal quotation marks, citation, and alterations omitted). And finally, under the third element, redressability, it must be "likely, as opposed to merely speculative, that the injury will be redressed by a favorable decision" of the Court. *Id.* at 561 (internal quotation marks and citation omitted).

Plaintiffs must satisfy the elements of standing "in the same way as any other matter on which the plaintiff bears the burden of proof, *i.e.*, with the manner and degree of evidence required at the successive stages of the litigation." *Id.* At the summary judgment stage,

therefore, a plaintiff must adduce "specific facts" "set forth by affidavit or other evidence" to establish its standing, *id.* (internal quotation marks omitted), such "that there exists no genuine issue of material fact as to justiciability," *Dep't of Com. v. U.S. House of Representatives*, 525 U.S. 316, 329 (1999). Moreover, because "standing is not dispensed in gross," *Town of Chester v. Laroe Ests., Inc.*, 581 U.S. 433, 439 (2017) (internal quotation marks and citation omitted), a plaintiff must "demonstrate standing for each claim [it] seeks to press" against each defendant, *DaimlerChrysler Corp. v. Cuno*, 547 U.S. 332, 352 (2006), and "for each form of relief sought," *Friends of the Earth, Inc. v. Laidlaw Env't Servs. (TOC), Inc.*, 528 U.S. 167, 185 (2000). "The same principle applies when there are multiple plaintiffs," that is, "[a]t least one plaintiff must have standing to seek each form of relief requested in the complaint." *Town of Chester*, 581 U.S. at 439.

Although neither party raised standing in their summary judgment briefing, the Court has an independent duty to assure that it has subject matter jurisdiction with respect to each claim before it and "*must* resolve" any doubt about standing before proceeding to the merits. *Bauer*, 774 F.3d at 1029 (emphasis in original) (quoting *Lee's Summit*, 231 F.3d at 41). Here, moreover, the Court raised the issue of standing at oral argument and provided Elevance and CMS with the opportunity to address the Court's concerns. Dkt. 29 at 51, 53, 56, 62–63. In response, CMS argued that the only plaintiff that has met its burden of showing that it has suffered or is likely to suffer an injury-in-fact that is traceable to the agency's alleged violations of its own regulations is Blue Cross Blue Shield Healthcare Plan of Georgia ("BCBS of Georgia"), which is the insurer that holds the H5422 contract. *Id.* at 51, 62; *see also* Dkt. 13 at 7 (Am. Compl.). As a result, CMS maintains that, if Plaintiffs prevail on the merits, the only Plaintiff that would be entitled to the relief Plaintiffs seek—that is, recalculation of their 2024 Star Ratings and a reassessment of

the payments owed to them, Dkt. 13 at 25–26—is BCBS of Georgia. At least on the present record, the Court agrees.

To meet their burden of demonstrating standing, Plaintiffs provided a declaration from Mark Abernathy, a certified public accountant who provides advice to clients on "operations, compliance, and strategic issues in the Medicare Advantage arena." Dkt. 15-1 at 1 (Abernathy Decl. ¶ 3).[5] In that declaration, Abernathy explains that he was asked by Plaintiffs' counsel to "analyze the cut points for one of the MA-PD measures[, the C25 measure,] for contract H5422" to "demonstrate the impact of CMS using simulated instead of actual 2023 Star Ratings cut points." *Id.* at 10 (Abernathy Decl. ¶ 27). He was also asked to analyze the cut points for the nine contracts held by the health plan Plaintiffs to determine how they would have fared had CMS declined to apply the Tukey outlier methodology to the 2024 Star Ratings. *Id.* at 12–14 (Abernathy Decl. ¶¶ 31–38).

The C25 measure reflects the number of complaints made to Medicare about a health plan. Each plan receives a score that equals the rate of complaints about the health plan per 1,000 members.[6] Because receiving fewer complaints is better than receiving more complaints, a plan that receives a lower score for the C25 measure receives a higher star rating for that measure. *2024 Technical Notes* at 76. The average score received by a plan for the C25 measure

[5] Abernathy is a Managing Director with the Berkeley Research Group. *Id.* (Abernathy Decl. ¶ 2). He has been appointed by "state and federal judges to provide operational and financial oversight of managed care plans, including both Medicaid and Medicare plans." *Id.* at 2 (Abernathy Decl. ¶ 4). Defendants have not objected to the expertise of Mr. Abernathy or the admissibility of his opinion as contained in the declaration.

[6] Specifically, the 2024 Technical Notes state that the score a plan receives for the C25 measure equals the "[(Total number of all complaints logged into the Complaints Tracking Module (CTM)) / (Average Contract enrollment)] * 1,000 * 30 / (Number of Days in Period)," where the Number of Days in Period equals the number of days in the year. *2024 Technical Notes* at 75.

in 2024 was 0.32 and the average star rating was 3.9 stars. *Id.* at 120. As relevant here, the H5422 plan held by BCBS of Georgia received a score for the C25 measure in 2024 of 0.37, Dkt. 15-1 at 12 (Abernathy Decl. ¶ 29), and in 2024, that score was given a rating of three stars, *2024 Technical Notes* at 76.

Table 2: 2024 Cut Points for C25
Member Complaints and Changes in Health Plan's Performance

1 Star	2 Stars	3 Stars	4 Stars	5 Stars
> 0.75	0.5 ≥ 0.75	0.32 ≥ 0.5	0.14 ≥ 0.32	≤ 0.14

Id. at 76.

According to the Abernathy declaration, if CMS had used the actual 2023 cut points instead of the simulated 2023 cut points, the H5422 plan would have received a higher star rating for the C25 measure in 2024. The Court agrees with Abernathy on this point. Moreover, although CMS disagrees with aspects of Elevance's calculations, *see, e.g.*, Dkt. 17 at 39–42 (challenging the way the restricted cap was calculated); Dkt. 26 at 21–23 (same), CMS does not dispute the bottom line: using actual as opposed to simulated cut points would result in a higher rating for the H5422 plan. *See* Dkt. 29 at 50–51. Following the methodology provided for in the 2024 CMS Technical Notes, the Court's own calculations confirm that conclusion.

The change in cut points for the C25 measure also, according to Abernathy's declaration, affects another measure: D02, which measures the complaints Medicare receives about a drug insurance plan. *See* Dkt. 15-1 at 2 (Abernathy Decl. ¶ 6). Abernathy attests that the H5422 contract also would receive a higher star rating on the D02 measure if the guardrail were applied in the manner that Plaintiffs advocate. *Id.* at 12 (Abernathy Decl. ¶ 29). When the higher star ratings for the C25 measure, the D02 measure, and the D01 measure (which measures foreign

language interpretation and TTY availability when members call the drug plan)[7] are taken into account, Abernathy attests that the overall star rating for H5422 in 2024 would increase from 3 stars to 3.5 stars. *Id.* (Abernathy Decl. ¶ 30). Plaintiffs allege that the H5422 plan's receipt of a lower star rating than deserved has "expose[d] [the Plaintiffs] to hundreds of millions of dollars in . . . loss." Dkt. 15 at 11; *see also id.* at 28–29. Considered in totality, the Court is persuaded that, among Plaintiffs, BCBS of Georgia has standing to challenge the agency's use of simulated 2023 cut points in the 2024 Star Ratings.

But that is as far as Plaintiffs' evidence goes. Notably, even after the Court raised concerns about their showing at oral argument, Plaintiffs have made no effort to supplement their evidentiary submission and have failed to proffer any evidence showing that any of the other plaintiffs in this suit have suffered, or are likely to suffer, any similar adverse effects resulting from CMS's use of the 2023 simulated cut points. At oral argument, moreover, Plaintiffs' counsel conceded that all that the evidence shows is that, if the Court grants Plaintiffs relief with respect to the guardrails issue, the H5422 contract will "move[,]" Dkt. 29 at 54; *see also id.* at 53, 55–56—but not that the contracts held by "the other companies," *id.* at 57, will move. Plaintiffs offer only three possible qualifications to this concession.

First, they argue that eight of the nine contracts identified in the Amended Complaint would receive different 2024 Star Ratings, if the Court directs that CMS refrain from applying the Tukey outlier deletion methodology *at all* in the 2024 Star Ratings—*i.e.*, if CMS had calculated the 2024 cut points the same way it calculated the 2023 cut points it released in 2022.

[7] Abernathy explains that "Elevance Health challenged a call in the reconsideration process relating to [the D01 star rating for] contracts H2593, H4036, H5431, R4487," which was settled in Elevance Health's favor. From Abernathy's review, he concluded that the challenge also benefits H5422. Dkt. 15-1 at 12 n.28.

Dkt. 15 at 21. The Abernathy Declaration, moreover, supports that contention. Dkt. 15-1 at 12–14 (Abernathy Decl. ¶¶ 31, 37–38 & tbl. 3). But, as far as the Court can discern from reviewing Plaintiffs' Amended Complaint and Motion for Summary Judgment, they are not challenging the application of the Tukey outlier deletion to the calculation of the 2024 cut points.[8] Plaintiffs' first claim for relief, for example, alleges that CMS improperly applied the Tukey statistical method in "determin[ing] cut points for 2024 Star ratings" because the agency "simulated the 2023 Star rating cut points assuming [that the] Tukey" methodology applied to the 2023 data. Dkt. 13 at 23 (Am. Compl. ¶ 67). Plaintiffs then allege that, "[a]s a result of the simulated cut points and the application of guardrails, the cut points increased by more than 5 percentage points from one year to the next—in violation of the plain regulatory language of 42 C.F.R. § 422.166(a)(2)." *Id.* (Am. Compl. ¶ 68). Significantly, Plaintiffs' first claim does not challenge the application of the Tukey outlier deletion to the 2024 data. And Plaintiffs second and third claims for relief merely track their first claim. *See id.* at 24–25 (Am. Compl. ¶¶ 73–82).

Nor would it matter if Plaintiffs had alleged that CMS unlawfully applied the Tukey outlier deletion to the 2024 data because they fail to identify *any* reason why it was improper for the agency to do so. To the contrary, CMS faithfully complied with its own regulations in applying Tukey outlier deletion to the 2024 dataset. *See* 42 C.F.R. § 422.166(a)(2)(i). If anything, it would have been unlawful for CMS to have failed to do so.

Second, Plaintiffs at least gesture at the contention that they do not know (and cannot know until CMS re-runs the data) whether other contracts might be affected by a decision

[8] Nor, for that matter, could Plaintiffs make such an argument as this is not a challenge to the rulemaking but is instead a challenge to CMS's adherence to its own regulations in calculating the 2024 Star Ratings, and it is clear from the face of the relevant regulations that the agency was required, at a minimum, to remove Tukey outer fence outliers from the 2024 dataset. *See* 42 C.F.R. § 422.166(a)(2)(i).

holding that CMS improperly applied the Tukey outlier deletion methodology to the 2023 dataset, resulting in changes in excess of those permitted by the guardrails. *See, e.g.*, Dkt. 29 at 57. But that effort fails for a variety of reasons. First and foremost, a plaintiff who has yet to establish standing—and merely posits that it might conceivably suffer an injury absent judicial relief—is not entitled to summary judgment. Second, it is difficult to reconcile Plaintiffs' uncertainty regarding all but the H5422 contract with their expert's ability to estimate the effects of application of the Tukey outlier deletions to the 2023 dataset on the H5422 contract. Plaintiffs offer no explanation for why their expert can make that calculation—and can also calculate the effect of the Tukey outlier deletion on the 2023 and 2024 datasets for all of the other contracts—but cannot determine whether and how all of the contracts, other than the H5422 contract, would be affected were the Court to hold that CMS erred in applying the methodology to create simulated 2023 cut points. Finally, Plaintiffs' counsel seemed to concede at oral argument that, "[t]o [her] knowledge," BCBS of Georgia is the only Elevance affiliate that would be affected by requiring CMS to use the actual, as opposed to the simulated, 2023 cut points. Dkt. 29 at 57. Thus, at a minimum, these vague assertions are too speculative to support a finding of standing. *See La. Env't Action Network v. Browner*, 87 F.3d 1379, 1384 (D.C. Cir. 1996); *Chamber of Com. of U.S. v. EPA*, 642 F.3d 192, 199 (D.C. Cir. 2011).

Third, Plaintiffs argue that the challenged methodology could have an effect on "Elevance's enterprise weighted average for new contracts." Dkt. 13 at 22 (Am. Compl. ¶ 61); *see also id.* at 24–25 (Am. Compl. ¶¶ 69–70, 76); Dkt. 29 at 53, 55–58. The problem with this argument is that Plaintiffs never explain the legal basis for their contention, and they offer no evidentiary support for this broader theory of relief. The word "enterprise" appears nowhere in their summary judgment briefs, *see* Dkt. 15; Dkt. 21, and their expert says nothing about this

theory, *see* Dkt. 15-1. Beyond those omissions, it is unclear whether this theory is just another way of saying that contracts other than the H5422 contract might conceivably be affected, and, regardless of whether it is a separate or distinct theory, it surfaces the same concerns about speculation that preclude the Court from finding that any plaintiff other than BCBS of Georgia has standing to sue. Finally, it is unclear whether Plaintiffs merely maintain that whatever relief BCBS of Georgia receives will flow through to other companies in the same enterprise or whether those other companies have their own stand-alone claim for relief. In short, the Court lacks sufficient argument or evidence to permit it to find on the present record that the other companies have standing to sue.

The Court, accordingly, concludes that Plaintiffs have satisfied their burden of establishing that BCBS of Georgia has standing to bring this action but have not carried their burden with respect to the remaining plaintiffs. By the same token, however, CMS has failed to demonstrate that the other plaintiffs lack standing and, because the agency did not raise standing in its cross-motion, the Court cannot—at least at this time—enter summary judgment in CMS's favor with respect to the standing of the remaining plaintiffs. The Court will, therefore, **DENY** both Plaintiffs' and CMS's motions for summary judgment with respect to all plaintiffs other than BCBS of Georgia. If Plaintiffs can better support their enterprise-ranking theory and require a separate judgment with respect to any other, affected members of the enterprise, they may promptly renew their motion for summary judgment with the appropriate legal argument and evidentiary support.

B. Use of Simulated 2023 Cut Points in 2024 Star Ratings

Even though the financial implications are substantial, the parties dispute on the merits is a narrow one. That dispute focuses on CMS's implementation of the Tukey outlier deletion

methodology for the 2024 Star Ratings. As explained above, CMS implemented the new methodology by removing Tukey outer fence outliers from the 2024 dataset, by recalculating what the measure-threshold-specific cut points would have been for the 2023 dataset had the Tukey outlier deletion methodology applied that year, and by then applying the guardrail caps to the differences between the simulated 2023 cut points and the actual 2024 cut points. According to Elevance, this was error for multiple reasons, most notably because the governing regulation, 42 C.F.R. § 422.166(a)(2)(i), does not authorize CMS to apply the Tukey outlier deletion method to star ratings issued before October 2023 (that is, before the 2024 Star Ratings), and, even more importantly, requires CMS to apply the guardrail in a manner that ensures that "the measure-threshold-specific cut points for non-CAHPS measures do not increase or decrease" by more than the relevant "cap from [one] year to the next," *id.* To avoid these errors, Elevance maintains that CMS should have, instead, implemented the guardrail for the 2024 Star Ratings by comparing the 2024 Tukey-outlier-adjusted cut points to the 2023 non-Tukey-outlier-adjusted cut points (*i.e.*, the cut points that were actually used for the 2023 Star Ratings) and then limiting the variation between those two sets of cut points (one adjusted, the other not) by the relevant cap. *See* Dkt. 23 at 26–28.

CMS counters that Elevance's reading of Section 422.166(a)(2)(i) is overly straightforward: To be sure, Section 422.166(a)(2)(i)'s guardrail limits the ability of measure-threshold-specific cut points to vary by more than the cap from year-to-year. But, on CMS's telling, the regulation does not specify how the measure-threshold-specific cut points are to be calculated for each year. Although it is possible to read the guardrail as limiting the degree to which the current year's cut points can vary from the prior year's cut points as actually calculated that prior year, CMS argues the better reading of the regulation—and the one it

adopted—is that the guardrail limits the year-to-year variation in cut points, where the cut points for both years are calculated using the current year's methodology. The latter interpretation, CMS argues, achieves two important purposes. First, CMS reads the preceding sentence, which makes the introduction of the Tukey outlier deletion methodology "[e]ffective for the Star Ratings issued in October 2023" (that is, the 2024 Star Ratings), to require the agency to apply the new methodology to both the 2024 and the 2023 datasets. 42 C.F.R. § 422.166(a)(2)(i). Reading the guardrail provision in the manner that CMS proposes, then, would permit the agency to give effect to its (equally expansive) reading of the provision making the Tukey outlier deletion methodology applicable to both the 2024 cut points and any comparative cut points used in applying the guardrails. Second, CMS's reading of the regulation "allow[s] for an apples-to-appl[es] comparison when applying the cap," while Elevance's reading would require an apples-to-oranges comparison. Dkt. 17 at 28.

To determine which party has the better argument, the Court starts with the text of the regulation, considering "both 'the language itself [and] the specific context in which that language is used.'" *Merit Mgmt. Grp., LP v. FTI Consulting, Inc.*, 583 U.S. 366, 378 (2018) (quoting *Robinson v. Shell Oil Co.*, 519 U.S. 337, 341 (1997)). Although many of the key definitions are found in Section 422.162, Section 422.166 sets forth the methodology for determining "Measure Star Ratings." 42 C.F.R. § 422.166(a). Section 422.166(a)(2), in turn, sets forth the methodology for determining cut points by applying the "[c]lustering algorithm."

Id. § 422.166(a)(2). Of relevance here, Section 422.166(a)(2)(i) contains three relevant steps, each set forth in a separate sentence. Those three steps bear on the question presented.

The meaning of the first sentence is not contested for present purposes. It provides:

> The method [or algorithm] 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.

Id. § 422.166(a)(2)(i) (first sentence). As explained above, mean resampling separates the "measure-specific scores for the current year[] . . . into 10 equal-sized groups, using a random assignment process," *2024 Technical Notes* at 150, which CMS uses to create ten sets of four cut points for each measure and then averages the ten cut points to arrive at a single measure-threshold-specific cut point. The meaning of the second and third sentences, in contrast, are hotly contested.

The second sentence adds the Tukey outlier deletion requirement to the cut-point calculation methodology:

> Effective for the Star Ratings issued in October 2023 and subsequent years, prior to applying mean resampling with hierarchal clustering, Tukey outer fence outliers are removed.

Id. § 422.166(a)(2)(i) (second sentence). This sentence introduces a step that *precedes* the step set forth in the first sentence. That is, "prior to applying mean resampling with hierarchal clustering," CMS must remove the Tukey outer fence outliers. *Id.* This step, moreover, did not take effect until October 2023—that is, for the determination of the 2024 Star Ratings.

Finally, the third sentence adds the guardrails to ensure that "the measure-threshold-specific cut points" do not vary by more than a specified cap from one year to the next. That sentence provides:

> Effective for the Star Ratings issued in October 2022 and subsequent years, CMS will add a guardrail so that the measure-threshold-specific cut points for

> non-CAHPS measures do not increase or decrease more than the value of the cap from [one] year to the next.

Id. § 422.166(a)(2)(i) (third sentence). The next sentence of Section 422.166(a)(2)(i) then specifies the size of the caps that apply, depending on whether the relevant measures are scored on a 0 to 100 scale or "[a] restricted range," and the final sentence provides that "[n]ew measures" will not be subject to the guardrails for the first three years. *Id.* § 422.166(a)(2)(i).

The Court starts with the third sentence, which Elevance features in its argument. On Elevance's reading, that sentence precludes CMS from varying any measure-threshold-specific cut point by an amount that exceeds the specified cap—typically five percent—from one year to the next. What that means according to Elevance is that CMS cannot increase or decrease any *actual* cut point from one year to the next, and CMS's use of *simulated* cut points for the benchmark year—here, the 2023 Star Ratings—cannot be reconciled with the plain language of the regulation. The Court agrees.

Although the phrase "measure-threshold-specific cut points" is not defined in the regulations, the meaning of that phrase is not hard to discern; as discussed above, "cut points" are high and low scores that separate one star from two, two from three, and so on, and "measure-threshold-specific cut points" are the cut points for each of the forty-two specific measures, as opposed to the aggregate measures (domain star ratings, summary star ratings, and overall star ratings). Nothing in Section 422.166, moreover, suggests that "measure-threshold-specific cut points" refers to anything other than the actual cut points set for the relevant year.

Indeed, if anything, the regulatory definitions confirm that the guardrails regulate variances between the actual cut points "from [one] year to the next." *Id.* Section 422.162(a), for example, defines "[g]uardrail" as "a bidirectional cap that restricts both upward and downward movement of a measure-threshold-specific cut point for the current year's measure-

level Star Ratings *as compared to the prior year's* measure-threshold-specific cut point." 42 C.F.R. § 422.162(a) (emphasis added). "Cut point cap" is defined as "a restriction on the change in the amount of movement a measure-threshold-specific cut point can make as compared to *the prior year's* measure-threshold-specific cut point." *Id.* (emphasis added). "Absolute percentage cap" is defined to mean "a cap . . . that restricts movement of the current year's measure-threshold-specific cut point . . . as compared to *the prior year's cut point*." *Id.* (emphasis added). And "[r]estricted range cap" is defined as "a cap applied to non-CAHPS measures that restricts movement of the current year's measure-threshold-specific cut point to no more than" a value "calculated using *the prior year's* measure score distribution." *Id.* (emphasis added).

The most natural way to read these references to the "prior year's" cut points is as a reference to actual cut points used for the prior year. In ordinary parlance, a reference to something from the prior year is often a reference to the historical fact of what happened that prior year. *Prior*, def. 1, Oxford English Dictionary (3d ed. 2007) ("That precedes in time or order; earlier, former, anterior, antecedent."). For instance, a rent control law that limits the year-to-year increase in the rent a tenant owes by comparing the current year's rent to the rent from the prior year is commonly understood to limit the amount charged this year as compared to the amount actually charged the prior year. Here, CMS asks the Court to depart from this ordinary usage and to read "the measure-threshold-specific cut points" for "[one] year," 42 C.F.R. § 422.166(a)(2)(i) (third sentence), or "the prior year's cut point," 42 C.F.R. § 422.162(a), to refer to the prior year's cut as recalculated using the current year's methodology. But that is not what the regulation says, and it is not how the relevant text is best—or even reasonably—construed. As Judge Nichols recently observed in a case raising this exact same interpretive question: "The best and most natural reading is that this regulation refers to the *actual* cut points

in the initial year just as it refers to the *actual* cut points that will be created for the next year." *SCAN Health Plan v. Dep't of Health and Human Services*, No. 23-cv-3910, 2024 WL 2815789, at *5 (D.D.C. June 3, 2024) (emphasis in original). And, as Judge Nichols further observed: "Had CMS wanted to apply a policy designed to increase the predictability of cut points to anything other than *actual* cut points, one might have expected it to say so explicitly." *Id.* (emphasis in original).

CMS resists this plain reading of the regulation. The agency argues that even if the guardrail provision, when read in isolation, is best construed to limit the current year's cut points from differing significantly from the prior year's *actual* cut points, when the guardrail provision is read in conjunction with the Tukey provision—that is, the second sentence—an ambiguity emerges that can be resolved by reading the guardrail provision—that is, the third sentence—to require use of *simulated* cut points for the prior year. This is because the Tukey provision, in CMS's view, requires the agency to remove the outer fence outliers from all years' data—not just the current year's data—as part of its 2024 Star Ratings calculations. Thus, to comply with the Tukey provision, CMS argues that it had to recalculate the 2023 cut points without the Tukey outer fence outliers.

CMS's argument suffers from several problems. To start, by arguing that the most natural reading of the guardrail sentence must give way to the plain language of the Tukey sentence, CMS has it backwards. The language of the guardrail sentence is unambiguous, and even if the Court assumes that the language of the Tukey sentence can be read as CMS suggests, that reading is by no means compelled and cannot create an ambiguity that does not otherwise exist in the guardrail sentence. Rather, to the extent any tension exists between the two sentences, it is most naturally resolved by reading the Tukey sentence to require application of

the outer fence outlier deletions to the dataset for the current year—that is, the 2024 Star Ratings. That reading of the Tukey sentence, moreover, is not only a reasonable one, which coheres with the language found in the guardrail sentence, but it is the only reading that makes sense when the Tukey sentence is read in light of the first sentence of Section 422.166(a)(2)(i). That sentence requires CMS to apply "mean resampling with . . . hierarchal clustering" to "the current year's data." 42 C.F.R. § 422.166(a)(2)(i). The Tukey sentence then requires CMS to remove outliers "prior to applying mean resampling with hierarchal clustering"—that is, it requires CMS to remove Tukey outliers from "the current year's data" before applying "mean resampling with . . . hierarchal clustering" to *that* data. *Id.* Again, Judge Nichols reached this same conclusion in reading this same regulatory text. *SCAN Health Plan*, 2014 WL 2815789, at *6.

Understood in this way, the first three sentences of Section 422.166(a)(2)(i) work together in a coherent manner. The first sentence requires "mean resampling with the hierarchal clustering of the current year's data" to identify the cut points. The second sentence requires CMS—starting in October 2023—to use the Tukey methodology to remove outliers from that dataset before applying the required "mean resampling with hierarchal clustering." And the third sentence requires CMS to apply "a guardrail so that the" cut points for the new year do not increase or decrease by more than the cap "from [one] year to the next." Read in this logical manner, the Tukey outliers are removed from the "current year's data," but the prior year's cut points are not recalculated using the prior year's dataset but the current year's methodology.

Finally, this reading avoids a further problem with the approach that CMS proposes. If CMS were correct, and if the second sentence required the removal of Tukey outliers not just from the current data, but from all data used to determine the current-year cut points, the agency might have been required to recalculate the cut points not just for the 2023 Star Ratings—as

CMS did—but also for the 2022 Star Ratings—which it did not do. This is because under CMS's reading of the regulation, actual cut points are replaced with simulated cut points derived from the prior year's dataset but applying the current year's methodology. But in order to recalculate the 2023 cut points using the current year's methodology, CMS would need to apply the 2022–2023 guardrails to ensure that the 2023 cut points would not increase or decrease from the 2022 cut points by more than the cap. To make that calculation, however, the agency would have been required to apply the Tukey outlier deletion methodology to the 2022 cut points, given CMS's contention that the second sentence requires CMS to apply Tukey not only to the current year's dataset, but to all data used—directly or indirectly—in the process. Although this cascade would carry back only to the 2022 cut points, because the guardrails did not take effect until October 2022, CMS offers no convincing explanation for why it failed to carry its reading of the regulation to its logical extreme. Indeed, even beyond that difficulty, CMS simply resets the guardrails beginning with the 2023 Star Ratings, but it offers no explanation for how that reset can be squared with the text of the third sentence, which requires application of the guardrails starting with "the Star Ratings issued in October 2022." 42 C.F.R. § 422.166(a)(2)(i). To the contrary, the entire premise of the guardrails is to ensure that the cut points do not increase or decrease too quickly over time.

CMS answers these concerns—and, indeed, premises much of its argument—on language that appeared in the preamble to the proposed and final rule that adopted the Tukey outlier deletion methodology. As described above, that preambulatory language stated that "for the first year (2024 Star Ratings)" that the Tukey outlier deletion would be in effect, CMS would "rerun the prior year's thresholds, using mean resampling and Tukey outer fence deletion so that the guardrails would be applied such that there is consistency between the years." Contract Year

2021 Tukey Final Rule, 85 Fed. Reg. at 33835; *see also id.* at 33833 ("We requested comments on our proposal to use Tukey outer fence outlier deletion as an additional step prior to hierarchal clustering. We explained that under our proposal in the first year of implementing this process, the prior year's thresholds would be rerun, including mean resampling and Tukey outer fence deletion, so that the guardrails would be applied such that there is consistency between the years."). This would, of course, be a very different case if the language contained in the preamble also appeared in the regulatory text, or if the preamble did not directly conflict with the only reasonable reading of the regulatory text. But for the reasons explained above, the preamble says something very different from the regulation itself. That conflict precludes CMS's argument.

As CMS itself acknowledges, it is axiomatic that "language in the preamble of a regulation is not controlling over the language of the regulation itself." *Entergy Servs., Inc. v. FERC*, 375 F.3d 1204, 1209 (D.C. Cir. 2004) (quoting *Wyoming Outdoor Council v. U.S. Forest Serv.*, 165 F.3d 43, 53 (D.C. Cir. 1999)); *see also* Dkt. 26 at 7 (acknowledging that when preamble text is "directly contradicted by text in the Code of Federal Regulations" it is "evidence that the agency did not intend the text to be binding" and thus the preamble text cannot be understood to have binding legal effect); Dkt. 29 at 29–31. In other words, if the preamble and the regulation conflict, it is the regulation that controls, even if in other circumstances the preamble might be considered binding on the agency. *See AT&T Corp. v. Fed. Commc'ns Comm'n*, 970 F.3d 344, 351 (D.C. Cir. 2020) ("[W]here, as here, there is a discrepancy between the preamble and the Code, it is the codified provisions that control."); *Nat. Res. Def. Council v. EPA*, 559 F.3d 561, 564–65 (D.C. Cir. 2009) (finding where the preamble described the regulation as applying to "high wind events" but the regulation itself, as published in the Code,

did not, the preamble statement was a nullity); *see also Kennecott Utah Copper Corp. v. Dep't of Interior*, 88 F.3d 1191, 1222–23 (D.C. Cir. 1996) (explaining that there is no "categorical bar to judicial review of a preamble;" rather, "[t]he question of reviewability hinges upon whether the preamble has independent legal effect, which in turn is a function of the agency's intention to bind either itself or regulated parties"). Because § 422.166(a)(2)(i) is best read to require CMS to apply the guardrail to limit the variation between the current year's cut points and the actual cut points issued the prior year, CMS's statements in the preamble to the contrary cannot have binding legal effect.

CMS's reliance on the preamble suffers from a second problem as well: even if the preamble was given the force of law, and even if it could trump the best reading of the regulatory text, it offers no justification for the agency's decision to reset the guardrails for the simulated cut points. As discussed above, the third sentence of the regulation prevents the measure-threshold-specific cut points from increasing or decreasing from year-to-year by more than the cap. In recalculating the cut points for 2023, however, CMS failed to apply the caps to ensure that those, simulated cut points did not increase or decrease by too large a margin from 2022. And, even if the preamble might justify applying the Tukey outlier deletion methodology to the simulated 2023 cut points, it says nothing about disregarding—or resetting—the guardrails that would otherwise limit changes in the cut points between 2022 and 2023. Rather, the preamble merely states that "for the first year (2024 Star Ratings), we will rerun the prior year's thresholds, using mean resampling and Tukey outer fence deletion so that the guardrails would be applied such that there is consistency between the years." Contract Year 2021 Tukey Final Rule, 85 Fed. Reg. at 33835. The preamble does not specify, in other words, that for the 2024 Star Ratings the guardrail would be applied as if it was introduced that year, notwithstanding its

introduction the prior year—and notwithstanding the regulatory command to "add a guardrail" "[e]ffective for the Star Ratings issued in October 2022." 42 C.F.R. § 422.166(a)(2)(i).

Nor is the Court persuaded that Elevance's proposed reading of the regulation would lead to absurd results or is so at odds with common sense that the Court should pause before giving the regulation its most natural meaning. *Cf. Cook v. Food & Drug Admin.*, 733 F.3d 1, 9 (D.C. Cir. 2013) ("We may . . . in rare instances depart from the plain text when adherence to the plain text leads to an 'absurd' result." (internal quotation marks omitted)). The agency explains that it reran the prior year's cut points so that the guardrail was applied in an "apples-to-apples" fashion; that is, the 2024 cut points were compared to cut points that were calculated using the same methodology. This was important, the agency further explained, because by rerunning the prior year's cut points without Tukey outliers prior to applying the guardrail improved the accuracy of the current year's cut points. The cut points were not dragged down, so to speak, by the earlier year's less precisely calibrated cut point calculation.

The Court recognizes that the accuracy of the cut points is important. But that is not the only factor in play; indeed, CMS has always understood that application of the guardrails would come at a cost to the accuracy of the cut points. In the final rule adopting the guardrail provision, the agency recognized that "[t]he trade-off for the predictability provided by bi-directional caps is the inability to fully keep pace with changes in performance across the industry." Contract Years 2020 & 2021 Guardrail Final Rule, 84 Fed. Reg. at 15754. As CMS further explained:

> A cap on upward movement may inflate the measure-level Star Ratings if true gains in performance improvements cannot be fully incorporated in the current year's ratings. Conversely, a cap on downward movement may decrease the measure-level Star Ratings since the ratings would not be adjusted fully for downward shifts in performance.

Id. The agency, nevertheless, determined that the cost to the accuracy of the star ratings in any given year was worth the benefits that the guardrail would provide in the form of stability and predictability.

Nor did CMS leave itself without options should it later conclude that the guardrails are unduly limiting the ability of the cut points to keep up with industry performance. CMS explained that "[i]f cut points are not keeping pace with the changes in scores over time, [it] may propose in the future how to adjust the cut points to account for significant changes in industry performance." *Id.* at 15757. And in a notice of proposed rulemaking issued by the agency on December 27, 2022, CMS proposed removing the guardrails altogether, explaining that "[w]hile [the agency] recognized the possibility at the time [it] finalized the guardrails policy" that "there may be an inability for thresholds to fully keep pace with changes in performance across the industry," "we now have evidence from the 2022 and 2023 Star Ratings that shows that unintended consequence of the policy" and the "importan[ce] [of permitting] cut points to adjust for unforeseen circumstances that may cause overall industry performance to either increase or decrease." Contract Year 2024 Policy & Technical Changes to the Medicare Advantage Program, 87 Fed. Reg. 79452, 79625 (Dec. 27, 2022). In its final rule, CMS declined to adopt this change and, instead, deferred the issue for "a later final rule." Contract Year 2024 Policy & Technical Changes to the Medicare Advantage Program, 88 Fed. Reg. 22120, 22121 (Apr. 12, 2023). But the point is that this is a policy choice that the CMS can pursue as it deems appropriate. The agency may not, however, amend the existing rule through a position taken in litigation.

For all of these reasons, the Court concludes that the correct reading of the regulation, *see Biden v. Nebraska*, 143 S. Ct. 2355, 2378 (2023) (Barrett, J., concurring); *Kisor v. Wilkie*, 588

U.S. 558, 589–90 (2019) (Courts "must make a conscientious effort to determine, based on indicia like text, structure, history, and purpose, whether the regulation really has more than one reasonable meaning"), requires CMS to apply the guardrails in a manner that limits variations in actual—as opposed to simulated—cut points by the relevant cap "from [one] year to the next." 42 C.F.R. § 422.166(a)(2)(i). It then follows that CMS's decision for the 2024 Star Ratings to apply the guardrail in an apples-to-apples fashion—that is, to recalculate the 2023 cut points using the Tukey outlier deletion methodology and to then compare those simulated cut points to the actual 2024 cut points—was contrary to the agency's own regulations and thus contrary to law and arbitrary and capricious. *See Battle v. FAA*, 393 F.3d 1330, 1336 (D.C. Cir. 2005) ("[A]gencies may not violate their own rules and regulations to the prejudice of others."); *United States ex rel. Accardi v. Shaughnessy*, 347 U.S. 260, 266–68 (1954).

C. Remedy

That leaves the question of remedy. Elevance requests the Court to "set aside CMS's unlawful actions" and to "order Defendants to recalculate Plaintiffs' 2024 Star Ratings by using actual 2023 Start Rating cut points." Dkt. 23 at 35–36. Although CMS disagrees on the merits, it does not argue for a different remedy. The only competing consideration, which neither party addresses, is whether and how a remedy that is limited to Elevance—and, more precisely, to BCBS of Georgia—might affect third parties. That concern is potentially significant, moreover, because star ratings allow healthcare consumers to compare the quality of MA plans, and an adjustment that affects one (or perhaps two, *see SCAN Health Plan*, 2024 WL 2815789, at *7) MA plans could skew the star rating system more generally. But because no party (or third party) has raised this concern with the Court, and because courts should, where possible, leave it to administrative agencies to determine in the first instance how best to implement a judicial

decision that alters the relevant legal framework, *see Center for Biological Diversity v. Regan*, No. 21-cv-119, 2024 WL 1602457, at *44 (D.D.C. Apr. 12, 2024), the Court will simply set aside the 2024 Star Ratings for BCBS of Georgia and will order CMS to redetermine those star ratings in a manner consist with this opinion. CMS, in turn, is free to decide whether other MAOs should receive similar relief in the administrative process, and, if necessary, any MAO suffering a cognizable injury in fact can pursue judicial relief to the extent appropriate.

CONCLUSION

For the foregoing reasons, the Court hereby **GRANTS** in part and **DENIES** in part Plaintiffs' motion for summary judgment and **DENIES** Defendants' cross motion for summary judgment. CMS's 2024 Star Ratings for BCBS of Georgia are hereby **SET ASIDE**, and CMS is **ORDERED** to redetermine those ratings in a manner consistent with this opinion.

SO ORDERED.

/s/ Randolph D. Moss
RANDOLPH D. MOSS
United States District Judge

Date: June 7, 2024