AMERICAN ACADEMY *of* ACTUARIES

Objective. Independent. Effective.™

January 2, 2020

Center for Consumer Information and Insurance Oversight (CCIIO)
Centers for Medicare and Medicaid Services (CMS)
Department of Health and Human Services (HHS)
Via email: CCIIOACARADataValidation@cms.hhs.gov

Re: December 2019 HHS-RADV White Paper

To Whom It May Concern,

On behalf of the Risk Sharing Subcommittee of the American Academy of Actuaries,[1] I appreciate the opportunity to provide comments on the HHS Risk Adjustment Validation (RADV) White Paper, released on December 6, 2019.[2]

To maintain stable individual and small group offerings, issuers need to be able to project the factors that will affect rates in the upcoming rating year. Risk adjustment transfers and any additional transfers made through RADV are two such factors. Although the base risk adjustment transfers can be quite large, most issuers can project these transfers fairly accurately, especially as they now have many years of experience with the program. The current RADV process, however, makes it extremely difficult for insurers to project RADV-related transfers and incorporate them into the rating process. In particular, the RADV approach has shown significant unpredictability and variability, which can lead to a sizable impact on risk adjustment transfers for some plans. Uncertain error rates and the uncertainty of whether there will be an outlier leads to uncertain effects on other plans in the state. In addition, environments in some states can lead to additional uncertainty and variability. For instance, states with a lot of provider capitation arrangements can have greater difficulty collecting the appropriate data; resulting error rates could be skewed when measured against a national benchmark error rate that is determined from mostly fee-for-service claims.

A RADV process should identify and hold accountable issuers that submit codes not reflecting enrollee diagnoses, thereby reducing incentives to overcode while also minimizing uncertainty and variability for other issuers. In the short term, it would be appropriate for CMS to focus on reducing issuer exposure to the variability and uncertainty in RADV-related transfers. In the long term, more significant changes to the program could be adopted that allow issuers more time to understand and adapt to changes. For example, a longer-term effort might remove or modify the

[1] The American Academy of Actuaries is a 19,500-member professional association whose mission is to serve the public and the U.S. actuarial profession. For more than 50 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.

[2] HHS Risk Adjustment Data Validation (HHS-RADV) White Paper, December 6, 2019.

confidence interval approach, such that if all issuers have error rates reasonably close to the mean, the approach would not force some issuers to be outliers. In short, the program should not punish issuers arbitrarily when both accuracy and precision are present.

The remainder of this letter makes more specific comments on the various sections of the white paper.

Glossary and Executive Summary

The error rate is defined as 1 – (stratum weighted sum of adjusted enrollee risk scores)/(sum of original External Data Gathering Environment (EDGE) enrollee risk scores). The denominator should also be weighted so that it equals the "stratum weighted sum of original EDGE enrollee risk scores."

The glossary makes the distinction between accuracy and precision, and the executive summary notes that a primary purpose of RADV is to "validate the accuracy of data submitted by issuers." However, the current methodology validates the precision of data submitted by issuers in the name of reducing the number of issuers identified by RADV.

Section 2: Changes to Sampling

CMS proposes several alternative sampling methods targeted at producing samples that are more representative of issuers' overall populations. Most of these procedures function by increasing the number of sampled individuals for larger issuers. Because mean group failure rates are currently determined using the frequency of Hierarchical Condition Categories (HCCs), increasing sample sizes for certain classes of issuers would give those issuers a disproportionate impact on the group failure rate. For example, combined with the current failure rate asymmetry, increasing the samples for outliers and near-outliers could lead to increased sample sizes for issuers with higher group failure rates. Inflated mean group failure rates could result, creating even larger negative error rate outliers than under the current approach. On the other hand, under 2017 HHS-RADV, larger issuers appear more likely to have negative error rates than smaller issuers; increasing sample sizes for larger issuers could reduce mean failure rates and cause even greater positive error rates than currently exist. In either case, these changes might be appropriate and more representative of actual patterns. Nevertheless, CMS should carefully evaluate the extent to which any changes to sampling impact mean group failure rates and the resulting error rates and issuer transfer adjustments. Even small changes in means and deviations can affect error rates and issuer transfers.

The white paper suggests expanding the exemption for RADV from the current 500 billable member months to 8,500 billable member months. It would be reasonable to use a billable member month threshold in lieu of a premium-based threshold for the materiality threshold for issuers participating in RADV every three years. Doing so would help align the RADV materiality threshold with statistical drivers of risk as opposed to the level of health care reimbursements in a state or the average benefit richness differences between issuers. The subcommittee recognizes CMS's concern about issuer gaming prospects if these issuers are instead always exempted from RADV. At the same time, the inherent volatility of results for

issuers with smaller enrollment and fewer enrollees with HCCs serves to limit the effectiveness of RADV as a method of ensuring coding accuracy for these issuers. CMS could consider performing some high-level metrics on smaller issuers' populations to determine whether broad EDGE HCC prevalence is significantly different from nationwide EDGE HCC prevalence, such as determining whether an issuer's percentage of enrollees with HCCs lies outside a 90 percent confidence interval for this ratio for all issuers ineligible for exemptions to either trigger a RADV sample or to put these issuers into a pool for targeted random sampling. Such an alternative for selecting issuers that are subject to targeted and random sampling could serve to provide an avenue for "typical" issuers to avoid the burdens of RADV while retaining enforcement authority over issuers who have more potential to be outside the bounds. Alternatively, CMS could consider granting exemptions based on having some share of business or normalized risk scores below particular thresholds, so that larger issuers selected to be exempt from RADV are more likely to have a minimal overall impact on transfers in the risk pool as a whole.

It would be appropriate to use HHS-RADV data to inform sample size selection given the differences between the Medicare Advantage (MA) population and the commercial population. Establishing a different standard for issuers that did not participate in prior HHS-RADV years is somewhat different than basing sample sizes on MA-RADV data, in that the sample size would be informed by another entity's data. Alternatively, HHS could consider using HHS-RADV average variation as opposed to issuer-specific variation. While this would remove some of the issuer-specific optimization, those optimizations would not address any changes made by the issuer in response to prior RADV years and so may already be less than optimal.

Section 3: Outlier Determination

CMS spends significant time in the white paper discussing alternatives for determining outlier status. We recognize CMS's reliance on the central limit theorem (CLT) in the assumption of normality of distributions. However, this theorem has limitations and care must be taken that is applied appropriately. In prior RADV methodology documents, CMS has explicitly referenced *t*-tests, which are robust primarily because the CLT holds fairly broadly. It may be more appropriate to reframe statements in the white paper in terms of *t*-tests, as these are a better analog to the statistical methodology used in RADV. While sample sizes appear to be large enough that 95 percent confidence intervals are unaffected, convergence to the normal distribution is slower as one goes further out into the tails, and use of *z*-scores from a normal distribution may not be as representative of the actual degree of magnitude of significant outliers.

We also note the importance of properly characterizing confidence intervals. A confidence interval serves to determine *when a sample observation is likely not the mean observation*. It might be inappropriate to characterize the mean of the samples as the population mean, particularly when there is significant variability in the variance of the sample means. Outlier detection serves primarily to determine values that are different than the average within the underlying population; one can make only limited inferences about the relative value of that outlier to the population mean. In this sense, correcting an outlier by moving it to the edge of the confidence interval can be viewed as moving it from a number that is probably not the mean to a number that is more likely to be the mean, and lies within the tolerance for what the mean is.

CMS notes that issuers with small numbers of HCCs are more likely to be determined as outliers. This is almost certainly driven by the limited applicability of the CLT to the validity of smaller issuers' results; while the overall distribution may be almost normal, the distribution for that issuer is less likely to be so. An ad-hoc approach, such as split confidence intervals for issuers with 30 or fewer HCCs, is one way to address this problem, but it would move the threshold of uncertainty to a smaller number of HCCs and would not address the statistical concerns underlying this methodology. Similarly, we agree with CMS's observation that a bootstrapping methodology would not meaningfully reduce the volatility in results for issuers with fewer HCCs relative to that option's computational complexity and lack of transparency.

The white paper discusses using a binomial distribution methodology to determine whether an issuer has inaccurate coding. For the most part, this methodology more intuitively aligns with the drivers of miscoding. However, this approach leaves several questions unanswered. First, how many issuers would be outliers for both newfound HCCs and unvalidated HCCs? Given that both metrics would result in about 5 percent of issuers being outliers, as many as 10 percent of issuers could be determined to be outliers if there is little overlap between issuers with newfound HCCs and those with unvalidated HCCs. If these metrics are further expanded across three HCC groups as is currently the practice, this methodology may not reduce the number of outliers. Moreover, the tendency to determine issuers with more HCCs (and thus likely with more members) as outliers without generating meaningfully different error rates could result in larger and more variable swings in transfer payments than under the current approach given the zero-sum nature of risk adjustment. Furthermore, the need to normalize the found rate metric against EDGE HCCs yields a less predictable element to the overall error rates.

The error rate adjustment proposed for newfound HCCs illustrates another challenge in the appropriate determination of outlier status. Taking a step back, RADV can perhaps best be thought of as the rate by which an issuer miscodes a diagnosis. Under this intuitive understanding, RADV should be evaluating variance relative to a "true" number of HCCs. But what would represent a true number? The current group failure rate metric measures variance relative to the number of EDGE diagnoses, but the whole premise of RADV is that EDGE might be incorrect, so this metric could itself be flawed. There are challenges to use of the initial validation audit (IVA) frequency as well, given that the number of validated HCCs is informed by the ability of the issuer to ensure auditors receive medical charts as well as the presence of supporting information for HCCs in those charts. That said, the use of IVA diagnoses is more consistent with the theory of RADV as a validation. Has CMS evaluated the impact of switching $Freq_{EDGE}$ and $Freq_{IVA}$ in the current group failure rate outlier determination methodology? Other parties are unlikely to have sufficient information to do so accurately. Meanwhile the binomial methodology uses terms that assume different "sources of truth" for correct coding, and as such might not capture meaningfully similar information about issuer practices with regards to newfound HCCs and unvalidated HCCs.

The white paper discusses McNemar's Test as another option; the underlying concept that the number of newfound HCCs be equivalent to the number of unvalidated HCCs is much closer to evaluating coding for net accuracy than other methodologies that evaluate for coding practices relative to other issuers. Although the test validates whether diagnosis errors balance out on a net

basis, it does not evaluate accuracy directly—an issuer with 100 of each incorrect coding type of both is treated equivalently to an issuer with none. As a metric focused more on absolute accuracy than relative accuracy, it is likely that more issuers would be affected. Given CMS's consideration of an acceptable level of validation failure, the target value of 0.5 could be adjusted to reflect that expected appropriate amount of failure. Such an adjustment may reduce the number of identified outliers and align with CMS's indicated concerns about an accuracy-based measure.

CMS references Bayesian methods as an approach to outlier determination. We share CMS's concerns that a Bayesian approach might be less likely to react with appropriate speed to changes in issuer coding practices in response to prior year audit results. However, a Bayesian methodology could be an appropriate supplement to identify small issuers that would otherwise be exempt. This would address some of CMS's concerns regarding a lack of transparency as the use of Bayesian techniques would not create any adjustments, although exactly how such an approach could be crafted is uncertain.

CMS has reviewed the use of machine learning (ML) algorithms to identify outliers and has determined that ML would be inappropriate. CMS's concerns are well founded. Insurers need a clear explanation of the outlier criteria for both pricing and risk adjustment program planning, but the ML methods investigated would not provide such clarity. It would also be preferable for insurers to know the outlier detection criteria at the start of the RADV process; ML methods would not produce criteria until after the data has been collected and the algorithm ran. Additionally, there are risks of swapping (identifying a normal data point as an outlier) and masking (multiple outliers concealing their presence). If a ML algorithm is selected in the future, a model validation should be performed.

The use of HCC hierarchies introduces some potential variation into the RADV methodology. It would be appropriate for an approach to include ranking HCC hierarchies as opposed to specific HCCs to determine HCC groupings. At the very least, HHS should consider grouping HCCs that are constrained to similar values to the same failure group.

Another alternative is changing the treatment of unvalidated HCCs. In many cases, issuers make good-faith efforts to obtain medical records from providers to validate diagnoses but are unable to do so. This is a relatively large problem that greatly contributes to the variance in failure rates, but there is no effort to address it in the current methodology or in any of the proposed methodologies. An HCC documented on a missing record is treated as a missing HCC, the same as a record that is unsupported by provided documentation. CMS could consider modifying the process so that issuers that document sufficient efforts to obtain documentation from providers for services indicated on EDGE claims (such as documentation of multiple attempts to obtain records including at least one attempt via certified mail at least 60 days prior to the end of the validation window) receive some form of credit for this effort. This credit could take the form of a decrement to the EDGE frequency count for HCCs that could have been validated by that record and which are not validated by another provider record, in effect excluding such HCCs from validation. Any such system would need to be constructed carefully to guard against insurers using the credit to avoid validation. Appropriate guardrails could include exclusions for

providers owned by related parties and a limit on the total percentage of providers for which these exemptions could be claimed.

Section 4: Error Rate Calculation

As with our other comments on the white paper and prior RADV-related comments, we stress the need for CMS to reduce the uncertainty and variability in magnitude of transfer adjustments associated with the current process. As CMS notes, the "payment cliff" is particularly problematic; it is a significant contributor to the volatility and uncertainty associated with the current RADV methodology and process. Immediate changes to the error rate calculation by eliminating or significantly reducing this impact are needed. The white paper presents several possibilities, two of which would be the most effective at addressing the payment cliff—the adjustment of outliers to the ends of the confidence intervals and sliding scale adjustment option one.

Adjusting the outliers to the ends of the confidence intervals would be the most effective and straightforward near-term "fix." This adjustment would greatly reduce the variability and magnitude of adjustments and greatly lessen the impact on other issuers with no adjustments. We understand that CMS is concerned that this approach would not do enough to remove the incentives for upcoding. However, the advantages of this approach in contributing to market stability and issuer confidence might greatly outweigh this concern. Because CMS considers risk scores of issuers with error rates falling within the confidence interval to be acceptable, it seems reasonable to adjust outliers to the acceptable level. Moreover, upcoding concerns are best addressed though the adjustment process. To the extent that errors are assigned because providers have not submitted medical records, it would not be appropriate to imply such errors are caused by upcoding. As we note above, automatically assigning errors to diagnoses for which the provider did not supply the medical record contributes to the uncertainty and variability of the current RADV process; there should be a more appropriate way to deal with this issue. To the extent that CMS is concerned about upcoding, we suggest that CMS evaluate upcoding through a different process. For example, CMS could take note of issuers that were consistently high outliers over multiple years or are extreme outliers as measured through the confidence interval process, and then evaluate such issuers through a separate audit process, and, if necessary, assess civil monetary penalties. Doing so should serve to provide reasonable disincentives for upcoding.

Option one of the sliding scale options, as shown in the Appendix to the white paper, would reduce adjustments more than the other sliding scale options and would eliminate the discontinuity at the boundary of the confidence interval. However, the method of adjusting outliers to the ends of the confidence interval would likely be more effective. As noted above, confidence intervals themselves do not calculate where the mean is, but rather where the mean is not. Even if this nominal national mean is used only for extreme outliers, the use of the "mean" reinforces this common statistical misunderstanding.

With regard to adjustment options for negative error rate outliers, it is appropriate to ensure that the methodology does not reward issuers that do not accurately code diagnoses—these issuers

have as much impact on confidence in risk adjustment results as positive error rate outliers but are instead rewarded for their practices by the RADV program. Placing a floor of zero on group failure rates seems to be a reasonable approach to address this concern; however, this approach could encourage issuers to overcode, as they are not rewarded for having coding practices that might be more strict than those used by the IVA or the second validation audit. As with charting, CMS should seek to balance those issuers that have more stringent coding practices and who may be targeting absolute accuracy versus those that do not engage in reasonable coding practices, trusting in RADV to compensate them and justify a low investment. This particular issue is challenging because of the current RADV focus on enforcing precision rather than accuracy—in the absence of an adjustment under the current methodology, accurate coders are penalized as noted by CMS. As such, an arbitrary floor of zero could unfairly penalize issuers with stringent coding practices that should be encouraged, though the resulting limitation on negative impacts to other non-outlier issuers would lead to increased stability in that state and market.

Section 5: Application of HHS-RADV Results

The Academy's Individual and Small Group Markets Committee previously commented on RADV timing considerations following the final 2020 Notice of Benefit and Payment Parameters (NBPPs). As noted in that letter, the prospective application of 2017 RADV results to 2018 risk scores and the recognition of amounts on 2021 medical loss ratio (MLR) reports and potential inclusion of amounts in 2021 premiums present significant challenges to issuers.[3]

Prospective application of 2017 benefit year RADV results to 2018 risk scores presents the possibility for distortion of transfers due to differences in issuer experience between the two years as well as changes to the makeup of the state, including statewide average premium levels. Concurrent application of 2017 benefit year RADV results to 2017 risk scores would better align the program with experience. A primary reason to apply 2017 RADV to 2018 risk scores was the ability to recognize transfers upon receipt of results. With the timing delays in the recognition of results due to the extended appeals window, this factor no longer applies. As such, concurrent RADV application would better address stability by reducing the number of extraneous influences on transfers. Concurrent application would not create additional workload for CMS as CMS already re-adjudicates 2017 risk adjustment results for RADV adjustments to exiting issuers. CMS notes this possibility but questions the ability to shift from prospective application to concurrent application because one year would, in effect, be adjusted twice. We note the following two counterpoints. First, CMS is not collecting transfer adjustments for 2017 RADV until 2021. This delay theoretically would give CMS the ability to modify payments to reflect concurrent application of RADV results with significant time for issuers to react to the changes prior to payment. If this were done as part of the 2021 NBPPs or along a similar time frame, then issuers would be able to reflect these changes in 2021 pricing if states permit them to do so. Second, while the switch would apply two years of RADV to a single year of experience, the RADV experience represents two different years. Exiting issuers are already subject to this dynamic, and, ultimately, other issuers still would see two years of impact. As long as the timing as to when RADV transfer adjustments are paid or received is unchanged, the impact likely

[3] For more discussion of the issues we noted, see our comments related to RADV timing here: https://www.actuary.org/sites/default/files/2019-10/RADV_Timing_Comments_100319.pdf.

would not unduly disturb the market (at least not to a greater extent than under the current RADV process).

Lastly, the subcommittee notes that none of the application provisions affect the timing considerations, which are significant. At the very least, CMS should consider aligning recognition of RADV adjustments in the MLR calculation with state treatment of RADV adjustments in pricing to avoid MLR distortions due to amounts being reflected in premiums differently than they are treated in the loss ratio determination.

Other Considerations

One challenge to external quantitative analyses of the impacts of any of the alternatives contained in the white paper is the relative lack of available data. As the only recipient of all validation results, CMS is in a unique position to be able to propose changes to the model and estimate their impacts. Third parties have limited ability to inform the process because the ability to quantify specific proposals is limited.

The current RADV process has some areas of apparent ambiguity and inconsistency between this white paper, the RADV protocols, and information contained in the annual NBPPs. One example is the error rate applied to exempt issuers. The 2018 NBPP references the "national average negative error rate, or the state negative error rate if lower," while the 2019 NBPP indicates that exempt issuers' transfers "will not have their risk adjustment transfers adjusted." Meanwhile, the white paper notes that exempt issuers "are not exempt from transfer adjustments," suggesting that exempt issuers will receive a 0 percent error rate. Another example is the calculation of error rates. The 2019 NBPP and white paper suggest through notational definitions of $EdgeRS_{i,e}$ and $AdjRS_{i,e}$ that only the HCC portion of risk scores is included in the calculation of the issuer error rate, while the RADV protocols indicate that weighted risk scores also include EDGE demographic information not referenced in the other documents. Both of these ambiguities lead to different RADV results and, further, foster uncertainty around the impact of the program. We suggest that HHS determine which is the authoritative description of the RADV program and ensure that definitions in the protocols and those documents align.

****

We would welcome the opportunity to speak with you in more detail and answer any questions you have regarding these comments. Please contact David Linn, the Academy's senior health policy analyst, at 202-223-8196 or linn@actuary.org to facilitate further discussions.

Sincerely,

Alfred A. Bingham Jr., MAAA, FSA
Chairperson, Risk Sharing Subcommittee
American Academy of Actuaries