Addendum to the Fee-For-Service Adjuster Study

I. Introduction

On October 26, 2018, we published a study entitled Fee-For-Service Adjuster and Payment Recovery for Contract Level Risk Adjustment Data Validation Audits, along with a Technical Appendix to that study. (We will refer to these documents collectively as the FFS Adjuster Study.) Since that time, we have released (or made available through a data use agreement) data underlying the FFS Adjuster Study.

On April 30, 2019, we published a Federal Register notice announcing our intention to replicate that study, publish the results, and release associated data. We explained that certain intermediate data elements not saved in the implementation of the initial study would be preserved in the replication and released to the public. We have now completed that replication. Associated data has been released at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-Risk-Adjustment-Data-Validation-Program/Resources.html.

As explained in that Federal Register notice, the results of the replicated study are consistent with the initial implementation of the study. We present those results below, compare them to the results of the study's initial implementation, and clarify some aspects of the methodology used in both the initial implementation and the replication.

II. Results of Replicated Study

The replication of the FFS Adjuster Study followed the same methodology described in the initial papers we published in October 2018, which we elaborate on below.

The results of the replication, although not identical to the results of the initial implementation, were consistent with those initial results. Because of the use of randomization in our study methodology, we expected (and have observed) some variation between the two implementations of our study.

In the initial study we calculated, with a 95% confidence interval, that audit miscalibration results in a slight (0.07% to 0.09%) overpayment to plans in the aggregate. In the replication of that study we calculated, again with a 95% confidence interval, that audit miscalibration results in a slightly smaller (0.06% to 0.08%) overpayment to plans in the aggregate. *See* Table 1. In both instances, our methodology estimated the effect of the audit miscalibration to be negative and extremely close to zero.

As we explained in the initial Technical Appendix, and as the results of the replication similarly show, the chance of the effect in any one of the fifty (50) simulations comprising each iteration of the study being greater than zero is very unlikely. While there appears to be a negative bias, it is so negligible that it should be considered to be zero. *See* Table 2 and Figure 1. This shows that fee-for-service diagnosis error in the fee-for-service calibration data does not manifest a negative payment bias to Medicare Advantage plans.

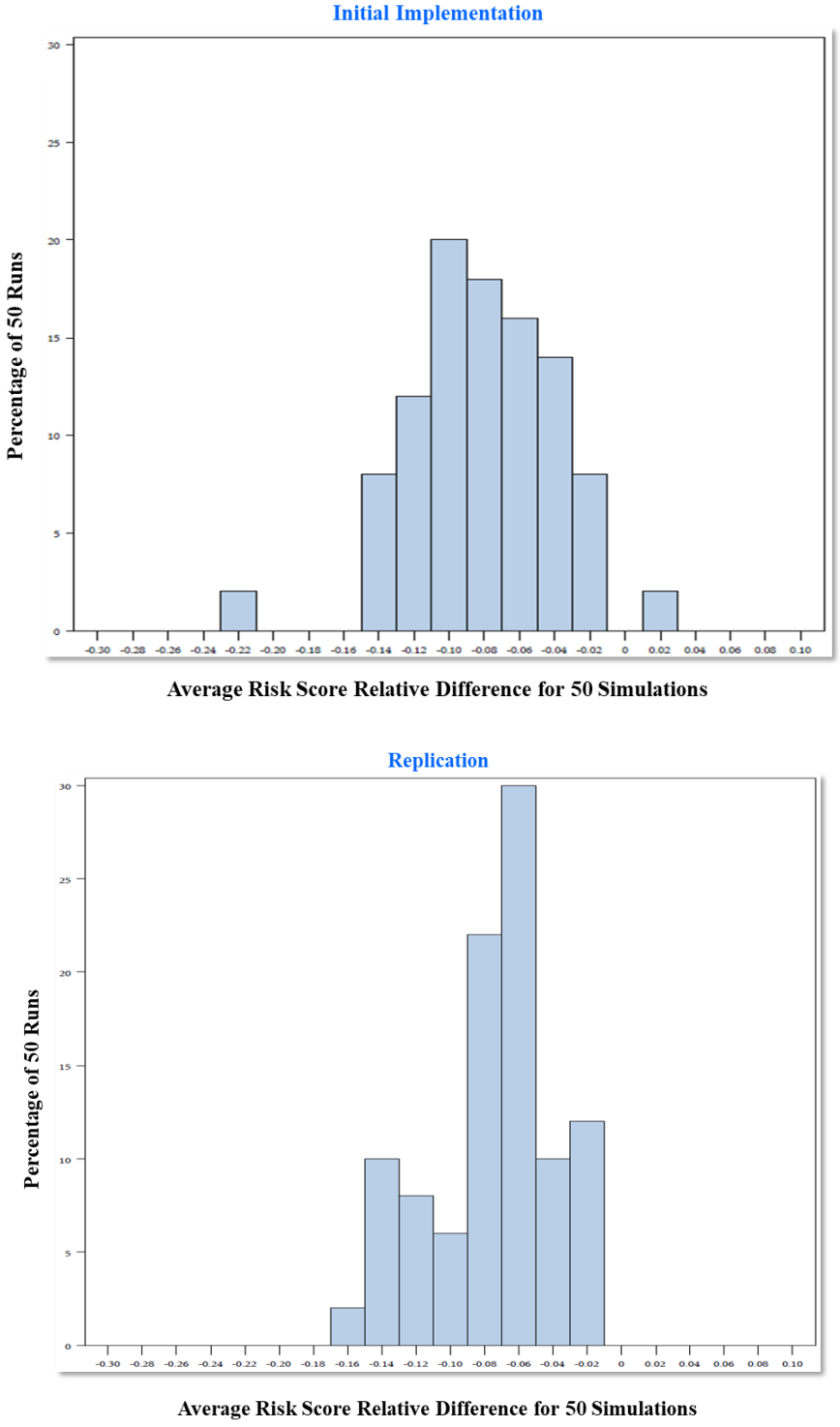
Table 1. Audit Miscalibration Estimate

Implementation	Mean (percent difference)	95% Upper Bound	95% Lower Bound
Initial	-0.08%	-0.07%	-0.09%
Replication	-0.07%	-0.06%	-0.08%

Table 2. Distributional Estimate of Bias

Distributional Statistic	Bias Estimate (Initial Implementation)	Bias Estimate (Replication)
Mean Relative Difference	-0.08%	-0.07%
Median Relative Difference	-0.09%	-0.06%
Minimum Relative Difference	-0.23%	-0.17%
Maximum Relative Difference	0.01%	-0.01%
25th Percentile	-0.10%	-0.09%
75th Percentile	-0.05%	-0.05%

Figure 1. Sampling Distribution of the Relative Difference for Initial Implementation and Replication



III. Study Methodology

As we explained in our October publications, the FFS Adjuster Study began by auditing 8,630 outpatient claims with CY 2008 dates of service. We reviewed the medical records associated with each claim (a small subset of the medical records associated with each beneficiary) to determine whether the diagnosis associated with the claim was supported by medical record documentation. In doing so, we excluded claims where providers refused to submit medical records, or did not provide sufficient documentation.

We drew our claims for audit from the 2008 Comprehensive Error Rate Testing (CERT) claims data. The qualifications of the medical record review contractors who reviewed the CERT medical records can be found at 75 Fed. Reg. 19,678, 19,747 (Apr. 15, 2010). Specific coding standards and guidelines apply. Medical records were coded according to the official conventions and instructions provided within the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), the ICD-9-CM Official Guidelines for Coding and Reporting, and guidance provided in the American Hospital Association Coding Clinic for ICD-9-CM, and our Medicare Advantage (MA) RADV Contract-Level coder guidelines.

A discrepancy rate for each CMS-HCC was calculated. For example, the data set contained 484 claims submitted with a diagnosis of chronic obstructive pulmonary disease, which is CMS-HCC 108. Of those diagnoses, 388 were supported by medical record documentation, and 96 were not, for a discrepancy rate of 19.8 percent. To account for the fact that the data set contained extremely small samples of many CMS-HCCs—for example, one diagnosis of extensive third degree burns and two diagnoses of severe head injury—we calculated a high, low, and baseline discrepancy rate. Each CMS-HCC was assigned one of these three mean discrepancy rates depending on its relationship to the baseline discrepancy rate: CMS-HCCs with a discrepancy rate significantly higher than the baseline were assigned to the high category, and those with a discrepancy rate significantly lower than the baseline were assigned to the low category. All other CMS-HCCs were assigned the baseline discrepancy rate. These rates were 46.2 percent, 33.8 percent, and 20.9 percent. These rates were established using a binomial test, which examines the likelihood that the number of successes of a given number of trials is consistent with a theoretical probability.

In a given year, multiple claims are submitted for Medicare Part B services received by a given beneficiary and associated with a given diagnosis. For example, an average beneficiary with metastatic cancer or acute leukemia, which is CMS-HCC 7, has seven claims associated with that diagnosis. Because we were interested in determining whether a given beneficiary had a documented diagnosis in a given year, and not whether any particular claim was associated with medical record documentation, we used the claim-level discrepancy rates described above to calculate beneficiary-level discrepancy rates. We used beneficiary-level discrepancy rates because an HCC is a beneficiary-level characteristic. A single unsupported claim therefore does not necessarily mean the HCC is unsupported such that the model calibration would be affected. For model calibration to be affected, all of a beneficiary's claims for a given HCC in a given year would have to be unsupported. Accordingly, the use of beneficiary-level discrepancy rates is appropriate.

After calculating this beneficiary-level discrepancy rate for each HCC, we ran fifty simulations in which we removed diagnoses from a data set of more than 1.4 million Medicare Part A and B beneficiaries at the HCC-specific beneficiary-level discrepancy rates. To determine which diagnoses to remove, a SAS Pseudo-Random number generator function generated numbers from a uniform distribution.

After removing diagnoses at the indicated rates, we used each simulated "corrected" data set to recalibrate the CMS-HCC risk adjustment model. As we explained in the Technical Appendix, we first estimated a proxy for the CMS-HCC model on the original uncorrected dataset. We estimated new dollar-denominated risk coefficients for each disease and demographic category, which we divided by the average FFS enrollee expenditure to calculate risk factors, which are expressed as percentages. We then applied these risk factors (the "original risk factors") to the MA sample, and estimated risk scores (the "original risk scores") for the uncorrected calibrating dataset. These original risk scores provided the baseline for our comparison with risk scores generated from the simulated "corrected" data sets.

Next, we carried out a series of steps to simulate the same MA enrollees' "corrected" risk scores—those that would have resulted from calibrating the model on a FFS dataset that was corrected in a RADV-like manner. We then estimated the CMS-HCC model on the simulated "corrected" data, calculating both dollar-denominated coefficients and risk scores expressed as percentages. For each of the fifty simulations, we then took the newly-calculated risk factors, applied them to the original FFS data set, and normalized these new sets of risk factors to one.

We performed this normalization because diagnoses cannot soundly be deleted in the CMS-HCC modeling context without making adjustments to account for the bias that the deletion procedure itself creates in expenditures. Without an appropriate adjustment, the predicted expenditures of the FFS population will be higher than they actually are for that population. If one were to perform a full RADV audit on the FFS data, certain unsupported diagnoses would result in unsupported claim amounts. As the regression fundamentally assumes that there is a relationship between claim dollars and diagnoses, the claim dollars must also be adjusted when deleting unsupported diagnoses, to simulate the unsupported claim amounts that would be removed from the data set in a full RADV audit. (The rationale for this adjustment is explained mathematically in Section IV.B, below, and the means by which the adjustment was calculated is explained in Section IV.C.)

It was these normalized risk factors derived from simulated "corrected" data that we then applied to our data set of MA beneficiaries, comparing their original risk scores to the risk scores calculated with each model recalibrated with simulated "corrected" data, as described above. We found that the difference between the risk scores was very small, and that the recalibrated risk scores tended to be slightly lower than the original risk scores. Therefore, we concluded that diagnosis error in FFS claims data does not lead to systematic payment error in the MA program.

Our study examined calibration error due to attenuation bias. In the Technical Appendix, we characterized attenuation bias in this way: "In regression modeling, independent variables are assumed fixed and free of measurement error. If independent variables have measurement error, the affected regression estimates may be biased downward." We have examined whether

measurement error would have caused a downward bias in HCCs that MA plans disproportionately utilize. As explained in our initial papers, any downward bias in certain individual HCCs due to measurement bias would likely be offset by upward biases to other HCCs and demographic factors. Executive Summary at 5 n.9. The degree of offset is an empirical question, and whether the impact creates a bias in payment in the aggregate depends on both the distribution of an MA organization's enrollees' relative factors and the nature of the inaccuracies. *Id.* at 2, 5 n.9. It is conceptually possible that the MA population as a whole utilizes more of the risk factors that increase under the audited calibration methodology, and thus, that payment would increase in the aggregate under this methodology. Technical Appendix at 13. Such a bias, if it existed, would be distinct from and unaffected by the deletion bias correction related to claims dollars described above. The study found no significant calibration error due to attenuation bias, and thus no systematic effect from measurement error on MA payment.

IV. Mathematical Explanations

A. Comparison of Initial and Replicated Implementations

The difference between runs is an arbitrary pseudo random seed. Because of the random process, the resultant quantities of each run will not be exactly equal. The relevant question is whether the difference in resultant quantities exceeds an amount not purely accounted for by random variation. Accordingly, the following test examines whether the difference in the resultant means are different to a statistical degree of certainty.

We assume the two runs of the code are realizations of a single process that differ only randomly by the use of different random seeds. Under this assumption, it is to be expected that the differences in the distributions of the relative differences vary only by an amount due to random error. Since the two runs were conducted independently, this can be tested statistically by comparing the mean relative differences of the two runs to the standard error of the difference of the means:

$$\frac{Mean\ of\ Run\ 1 - Mean\ of\ Run\ 2}{SE(Mean\ of\ Run\ 1 - Mean\ of\ Run\ 2)}$$

Since the runs were conducted independently,

$$SE(Mean\ of\ Run\ 1 - Mean\ of\ Run\ 2) = \sqrt{SE^2(Mean\ of\ Run\ 1) + SE^2(Mean\ of\ Run\ 2)}$$

So, the test statistic is calculated as follows:

$$\frac{Mean\ of\ Run\ 1 - Mean\ of\ Run\ 2}{\sqrt{SE^2(Mean\ of\ Run\ 1) + SE^2(Mean\ of\ Run\ 2)}}$$

Inserting the actual numbers from the two runs, we get the following value for the test statistic:

$$\frac{-0.00082 - (-0.00074)}{\sqrt{(0.0000565685)^2 + (0.0000523259)^2}}$$

$$= \frac{-0.00082 + 0.00074}{\sqrt{0.0000000032 + 0.000000002738}}$$

$$= \frac{-0.00008}{\sqrt{0.000000005938}}$$

$$= \frac{-0.00008}{0.000077058}$$

$$= -1.03817$$

Therefore, the (rounded) test statistic value for this test comes in at -1.04 with a p-value of approximately .30. The conclusion is that there is no evidence of a statistical difference in mean relative differences between the two runs.

The results of this test are expected as it comports with statistical theory and *a priori* expectations.

B. General Expenditure Adjustment to Offset Deletion Bias

(1) Consider a regression of an explanatory matrix X of order $m \times p$ that has zeros and ones with an independent variable of expenditures.

(2) The sum of the estimated coefficients, b_i of the matrix must equal total expenditures. Let *j* index the number of the enrollee observations such that *j = 1, 2, …, m*. Let *i* index the CMS-HCC model disease grouping coefficients such that *i = 1, 2, … , p*. Whether a particular enrollee *j* has a particular disease *i* is defined by the indicator variable I_{ji}.

$$I_{ji} = \begin{cases} 1 \; if \; observation \; j \; has \; a \; diagnosis \; of \; HCC \; i \\ 0 \; otherwise \end{cases}$$

We see the total sum of the coefficients will equal the total expenditures.

$$\sum_{j=1}^{m} \sum_{i=1}^{p} b_{ji} I_{ji} = \sum_{j=1}^{m} E_j$$

where E_j is the expenditures for the *j*th enrollee, $E_j \geq 0 \; \forall \; j$;
with

$$\sum_{j=1}^{m} \sum_{i=1}^{p} I_{ji} = n^*$$
$$n^* \geq 0 \; and \; m \geq 1.$$

Intuitively, n^* is the total number of HCCs that are turned on.

(3) We can express the expenditures in terms of average coefficient per HCC that is turned on:

$$\bar{b} = \frac{\sum_{j=1}^{m} E_j}{n^*} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{p} b_{ji} I_{ji}}{n^*}$$

(4) If values of one in the X matrix where randomly deleted, then n^* would decrease to n'.

$$n' > 0 \; and \; n' < n^*$$

(5) The average coefficient is now

$$\bar{\bar{b}} = \frac{\sum_{j=1}^{m} E_j}{n'}$$

(6) It follows that

$$\bar{\bar{b}} > \bar{b} \; since \; n' < n^* \; and \; E_j \geq 0, \forall \; j.$$

(7) Multiplying both by the original number of ones n^*

$$\bar{\bar{b}} n^* > \bar{b} n^*$$

(8) Which implies the total using the "new" coefficients is greater than the actual expenditures.

$$\bar{\tilde{b}}n^* > \sum_{j=1}^{m} E_j$$

(9) Which in turn implies that the total using the "new" coefficients is greater than the original predicted expenditures.
Letting the predicted expenditures for individual j be $\sum_{j=1}^{m} E_j$. We know that $\sum_{j=1}^{m} E_j = \sum_{j=1}^{m}\sum_{i=1}^{p} b_{ji}I_{ji} = \sum_{j=1}^{m} \tilde{E}_j$. Therefore, $\bar{\tilde{b}}n^* > \sum_{j=1}^{m} \tilde{E}_j$

(10) Accordingly, deletion of ones in the X matrix MUST increase the predicted expenditures if it is applied to the matrix prior to the deletion.

(11) To maintain coefficient values such that predicted FFS expenditures do not increase, we need a factor Z such that

$$Z\left(\bar{\tilde{b}}n^*\right) = \bar{b}n^*$$

(12) Thus, the adjustment factor is

$$Z = \frac{\bar{b}n^*}{\bar{\tilde{b}}n^*} = \frac{\bar{b}}{\bar{\tilde{b}}} = \frac{m\bar{Y}}{m\bar{\tilde{Y}}} = \frac{\bar{Y}}{\bar{\tilde{Y}}}$$

Where $\bar{Y}$ are average expenditures of the FFS enrollees and $\bar{\tilde{Y}}$ are average predicted expenditures of FFS enrollees from the perturbed CMS HCC model.

(13) Which implies that the proper adjustment is

$$\frac{\bar{Y}}{\bar{\tilde{Y}}}\tilde{b}_i = \tilde{\tilde{b}}_i$$

where the perturbed regression coefficients/ risk factors are $\tilde{b}_i$. The renormalized coefficients are $\tilde{\tilde{b}}_i$.

(14) We note that the study did not attempt to estimate the impact of FFS overcoding *per se*. Such an analysis would have to examine FFS enrollees' entire disease profiles for both overcoding and undercoding. The net impact would cause bias. Because CMS could not do this, the overcoding that resulted from our deletion methodology was neutralized. We also note that this may have been overly conservative, as some studies have found that there is net undercoding of FFS enrollee profiles, as one might expect from the different economic incentives of the MA and FFS payment systems. *See* Kronick and Welch, *Measuring Coding Intensity in the Medicare Advantage Program*, MEDICARE & MEDICAID RESEARCH REVIEW (MMRR), 2014, Vol. 4, No. 2, at E3–E5.

C. Explanation of the Inflated Post-Audit Risk Score (IPARS) Histogram

This section summarizes the methodology used, based on the calculations and adjustments described in Section B above, to calculate an adjustment to offset the bias that the deletion procedure itself creates in expenditures. We refer to this adjustment the Inflated Post-Audit Risk Score (IPARS). For each of the 50 simulations, an IPARS was calculated using the process outlined below:

- We calculated predicted expenses for each sample beneficiary using the dollar coefficients calculated from the simulated "corrected" data set, which we then applied to the un-perturbed FFS data. We call this "mean predicted expenses."
- We calculated the average, actual expenses for each sample beneficiary. We call this "mean expenses."
- The IPARS adjustment factor is the mean predicted expenses divided by the mean expenses.

We thus have an IPARS adjustment factor for each of the 50 simulations. The histogram shows the distribution of the 50 IPARS adjustment factors. The mean IPARS adjustment factor is 1.0089048 and the median is 1.0089006. The summary statistics for the IPARS adjustment factor are summarized in the table below.

Table 3. Distribution of Average IPARS Over 50 Simulations

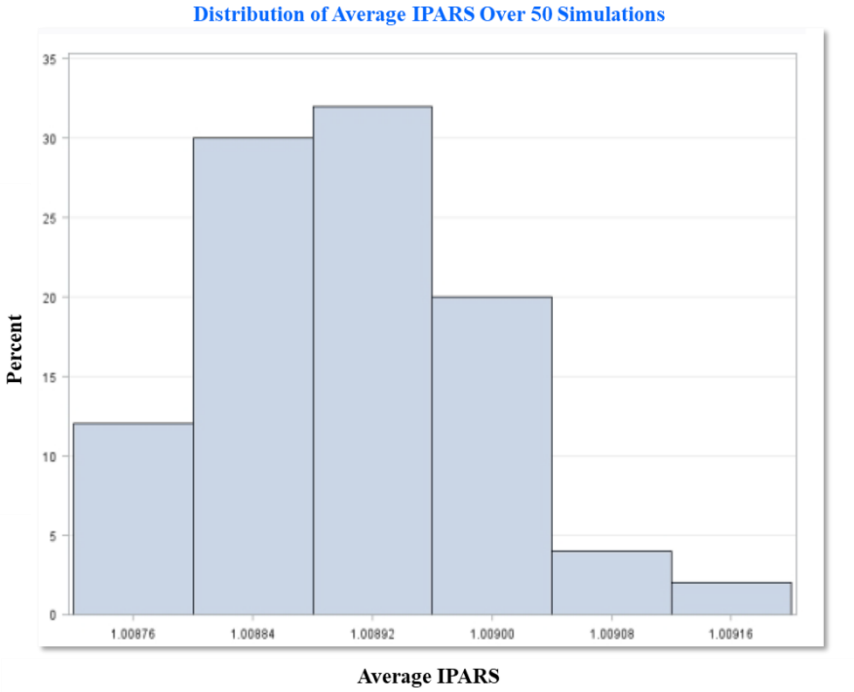


Figure 2. Summary Statistics of Average IPARS Over 50 Simulations

Summary Statistics of Average IPARS Over 50 Simulations

Minimum	Maximum	Median	Mean	25th Percentile	75% Percentile
1.0087524	1.0091804	1.0089006	1.0089048	1.0088274	1.0089653