Fee for Service Adjuster and Payment Recovery for Contract Level Risk Adjustment Data Validation Audits - Technical Appendix

Development of the Fee for Service Adjuster (FFSA) was initiated as a response to the Medicare Advantage (MA) organizations' assertion that because the Centers for Medicare and Medicaid Services - Hierarchical Condition Category (CMS-HCC) risk adjustment model is calibrated on FFS diagnoses that are not validated by medical record documentation, there is an inaccuracy in payment that should be accounted for in Risk Adjustment Data Validation (RADV) audit recoveries. They have argued that the inaccuracy due to audit miscalibration introduces a systematic bias that causes underpayments to plans relative to FFS, which in turn is exacerbated by RADV payment error recoveries. [1]

In this section, we provide an expanded discussion on the theoretical, economic, and empirical grounds to evaluate the conception of the FFSA. The issues we focus on are as follows:

(1) Statistically, the foundations that would lead to such a payment bias are largely absent.
(2) Application of a FFSA to RADV recoveries is not an appropriate way to adjust for this bias without arbitrarily changing the payment methodology for plans deemed to owe recoveries vis-à-vis unaudited plans.
(3) Our empirical analysis does not provide evidence of a calibration error bias on the MA population.

I. Weak Statistical Foundations

The statistical foundation that would lead to a consistent bias is largely non-existent within the context of the risk adjustment modeling framework for MA. The discussion below examines how a bias can arise and why our circumstances are different.

Assume one wished to estimate an econometric model and apply it to the same population upon which the data was generated. 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 first look at the econometric framework given the above assumptions. A simplified CMS-HCC model can be the following

$$Exp = \alpha D + \beta HCC^{*} + e$$

[1] In the context of this discussion, "underpayments" and "overpayments" are evaluated relative to the expected true FFS costs of a given model coefficient. They are not RADV overpayments and underpayments.

where Exp are FFS expenditures and HCC^* has the value of one if the sole HCC is present and zero if the HCC is not present and D represents the existence of a demographic characteristic.

We measure HCC^* with error so

$$HCC = HCC^* + w$$

Which in turn implies

$$HCC^* = HCC - w$$

Here, w is the measurement error of the HCC.

Thus, instead of estimating $Exp = \alpha D + \beta HCC^* + e$, we in fact estimate

$$Exp = \alpha D + \beta(HCC - w) + e$$

Rearranging terms, we can see this is

$$Exp = \alpha D + \beta(HCC) + (e - \beta w)$$

or

$$Exp = \alpha D + \beta(HCC) + z$$

Where the actual error term $z = (e - \beta w)$.

Under Ordinary Least Squares (OLS) assumptions, the error term is not correlated with the independent regressors. In this case, we see that z (the error term) contains the parameters of the HCC regressors. Thus, the error term is correlated with the regressors and therefore ostensibly violates the OLS assumptions. This may lead to bias and inconsistent estimates.

The central issue is the covariance between z and the HCC.

It can be shown that this covariance is

$$cov(z, HCC) = -\beta\sigma_w^2$$

Now we see that the extent of the problem depends on the variance of the measurement error from observation to observation of the FFS data. This is an empirical question. There are circumstances where there would be no significant bias.

In practice it may not be unreasonable to assume that the measurement error is fairly constant for the HCC across FFS observations. For example, the probability of error of the HCC may be ten percent. Under this assumption, the variance would be close to zero (the variance of a constant is zero) and by extension the covariance would also be close to zero. This in turn implies that the OLS assumptions essentially hold and the regression procedure produces unbiased and consistent estimates. The variance may or may not be too small to have a substantial impact. The essential point is that, even if the problem conceptually exists, it may not have much effect in practice. The impact must be measurable, it cannot simply be assumed to exist.

More importantly, for payment purposes CMS is not interested in each HCC *per se*, it is interested in the sum of the HCCs along with demographic characteristics used to create the risk score. This means that positive and negative errors will generally offset, mitigating the impact of any bias.

Under this payment framework, while increasing the accuracy of the model may be appropriate, adjusting for underpayments while not adjusting for overpayments is inappropriate. Consider an example where HCCs have an average payment error of zero (i.e., unbiased, but a standard deviation of ten percent). It should be noted that plans are not paid for single HCCs. They are paid in the context of a portfolio of HCCs both within and across enrollees. In this circumstance, there will be some HCCs that have an underpayment of ten percent. However, because enrollees have a number of HCCs and plans have a number of enrollees, the error dramatically reduces by diversification.

Table 1. Reduction in Error by Diversification

Number of HCCs	Mean	Error
1	0	0.100
2	0	0.071
3	0	0.041
4	0	0.020
5	0	0.009
6	0	0.004
7	0	0.001

We see that multiple HCC submissions dramatically reduce the error rate. In addition, if we attempted to adjust on only correct underpayments, we would introduce an overpayment bias. The regression estimates, for our purposes of estimating the relationship between Medicare payments and risk scores, must follow:

$$\overline{Exp} = \alpha\overline{D} + \beta\overline{HCC}$$

Where $\overline{Exp}, \overline{D}$ and $\overline{HCC}$ represent the model averages. The above equation implies that:

$$\alpha\overline{D} = \overline{Exp} - \beta\overline{HCC}$$

Thus, if the coefficients on the HCCs were biased downward, the demographic coefficients would be forced upward to maintain the expenditure average. Since risk scores sum the demographic factors as well as the HCCs, this would have a moderating effect on any bias. As a consequence, it is unlikely that a bias would occur and because of the purpose for which the model is used, econometric corrections might not be the appropriate remedy.[2]

II. Calibration Error Correction Limited to Recoveries is Economically Problematic

If there was a payment bias due to audit miscalibration, an adjustment targeted only to plans in a RADV audit would be arbitrary. An adjustment to original payments would be more appropriate.

Why Direct (Mean-Based) Application of the FFSA is Inappropriate

Assume that an enrollee's risk score is r_i^o. The enrollee is paid at the county rate C_i. The total original payment is $r_i^o C_i$. Upon audit, the risk score is recalculated to r_i^c. The enrollee's payment error (PE_i) is $(r_i^o - r_i^c)C_i$. Positive values indicate that the plan was overpaid and vice versa for negative values.

The calibration error (i.e., the error from using unaudited FFS data to develop the risk scores) presumably causes risk scores to have error due to the model being calibrated on unaudited FFS data. If Δ is the percent error in risk score due to calibration error, then the payment error, accounting for calibration error, is:

$$\left[\left(r_i^o(1-\Delta)\right) - \left(r_i^c(1-\Delta)\right)\right]C_i = [(r_i^o - r_i^c)C_i - \Delta(r_i^o - r_i^c)C_i] = PE_i - \Delta PE_i$$

Thus, the corrected payment error is:

$$PE_i^{**} = (1-\Delta)PE_i$$

In order for the direct adjustment to reduce the payment error, the Δ must be positive. But if Δ is positive, plans were overpaid in the first place. This leads to overpaid plans getting reduced recoveries. For example, if a representative MA population had an R^O equal to 1.045 and an R^c of 1.030, Δ would equal 1.5% ((1.045/1.030) – 1.0). Thus, the PE^* would need to be reduced by 0.985 (1-0.015):

$$PE^{**} = .985 * PE^*$$

A counterintuitive result arises when applying the adjustment in this scenario: the adjustment reduces the payment error amount. That is, MA plans were overpaid in the first place, and then they are being provided an offset to the payment error amount, which can be viewed as providing the plan an additional improper payment.

Now, consider the circumstance where plans are underpaid because the original unaudited model calibrated risk scores were too low. In this circumstance, Δ is negative

[2] For example, the FFS and MA populations are different. A bias may appear because of differences in the manner in which HCCs are utilized in the FFS and MA populations. This would have nothing to do with measurement error bias.

and the adjustment will increase the payment error. This is problematic as the original plan payments are not increased to compensate. Applying the adjustment only to the payment error would damage plans, as they would have been paid less than their enrollees' risk scores warrant and now the Δ adjustment (i.e., FFSA) would actually take back more money as well. This is again a counterintuitive result.[3]

In summary, if a plan's original payments represent consistent underpayments, the correct economic solution is to pay plans for their underpayments. Thus, eliminating a need to adjust recoveries. Moreover, if plans were consistently underpaid, and the attempt was made to collect the true economic value of the recoveries, those plans would have to return an amount over and above their original payment from CMS. In this case, plans would be returning payments to CMS that they never received which is counterintuitive.

To discount underpaid plans based directly on the audit miscalibration bias is contrary to economic reasoning. While its effect would moderate the impact of the original underpayment, the impacts would be economically arbitrary.

III. Empirical Analysis

This section describes the empirical analysis.

A. Input Data

This analysis requires three source data inputs, (a) FFS calibration data, (b) MA enrollee data and (c) the probability of HCC discrepancy for a RADV-like audit.

1. FFS data

The FFS data used in this analysis was the 2004-2005 file used for model calibration. There are 1,441,247 records in this data set.

2. MA sample

The MA sample has two million records that were sampled from the 2011 overpayment run. Approximately, one-half of the records are randomly sampled from the RADV eligible population and the other half are sampled from the non-RADV eligible population (excluding new, ESRD, and hospice enrollees).

3. Comprehensive Error Rate Testing (CERT)

[3] Say that the government purchased a $4 widget for $3. The government later wanted to get its money back because the widget was bad. A widget adjuster might say that there is a $1 bias underpayment. The correct economic value is (original price+bias = $3+$1) = $4. For the government to use the widget adjuster to correct the widget price to true economic value and demand $4 when they originally paid $3 is not equitable.

A sample of FFS claims and associated medical records were collected from the (CERT) audit. The CERT audit samples FFS claims for the purpose of reporting the error rate in FFS procedure code reporting. The sample consisted of a subset of claims with CY 2008 dates of service and the associated medical records.

B. Methodology

1. Probability of FFS Discrepancy

FFS Claim Level Discrepancy rates

Using the associated medical records, CPI performed a RADV-like review on the CERT data. We filtered the CERT data to create a subset of 8,630 unique claims to target for a RADV-like medical record review. The filtering (inclusion) criteria included:

- Outpatient claims paid through Medicare Part B. While CERT also samples Part A claims, they were not included in the FFS08 sample
- Claims that originated from an acceptable risk adjustment provider type
- Claims that contained ICD-9-CM diagnosis code(s) that mapped to one or more CMS-HCC

We will refer to the claims targeted for the audit as FFS08. CERT provided CPI with medical records for every claim in the sample. Using RADV coding guidelines, we performed a medical record review to develop discrepancy rates for the HCCs mapped from diagnoses on the claims. Table 2a summarizes the findings of the medical record review on FFS08.

Table 2a: FFS08 CERT Sample RADV HCC Discrepancy Counts at Claim Level

HCC	HCC Label	Sampled	Confirmed	Discrepant	%Discrepant
ALL	ALL	8,763	5,790	2,973	33.90%
HCC1	HIV/AIDS	22	16	6	27.30%
HCC2	Septicemia/Shock	51	30	21	41.20%
HCC5	Opportunistic Infections	12	3	9	75.00%
HCC7	Metastatic Cancer and Acute Leukemia	74	53	21	28.40%
HCC8	Lung, Upper Digestive Tract, and Other Severe Cancers	246	149	97	39.40%
HCC9	Lymphatic, Head and Neck, Brain and Other Major Cancers	275	170	105	38.20%
HCC10	Breast, Prostate, Colorectal and Other Cancers and Tumors	790	444	346	43.80%
HCC15	Diabetes with Renal or Peripheral Circulatory Manifestation	104	87	17	16.30%
HCC16	Diabetes with Neurologic or Other Specified Manifestation	123	98	25	20.30%
HCC17	Diabetes with Acute Complications	5	5	0	0.00%
HCC18	Diabetes with Ophthalmologic or Unspecified Manifestation	64	57	7	10.90%
HCC19	Diabetes without Complication	1,122	895	227	20.20%
HCC21	Protein-Calorie Malnutrition	11	5	6	54.50%

HCC	HCC Label	Sampled	Confirmed	Discrepant	%Discrepant
HCC25	End-Stage Liver Disease	4	3	1	25.00%
HCC26	Cirrhosis of Liver	19	14	5	26.30%
HCC27	Chronic Hepatitis	11	6	5	45.50%
HCC31	Intestinal Obstruction/Perforation	56	41	15	26.80%
HCC32	Pancreatic Disease	36	30	6	16.70%
HCC33	Inflammatory Bowel Disease	33	25	8	24.20%
HCC37	Bone/Joint/Muscle Infections/Necrosis	45	32	13	28.90%
HCC38	Rheumatoid Arthritis and Inflammatory Connective Tissue Disease	219	161	58	26.50%
HCC44	Severe Hematological Disorders	76	40	36	47.40%
HCC45	Disorders of Immunity	32	4	28	87.50%
HCC51	Drug/Alcohol Psychosis	4	2	2	50.00%
HCC52	Drug/Alcohol Dependence	16	9	7	43.80%
HCC54	Schizophrenia	197	111	86	43.70%
HCC55	Major Depressive, Bipolar, and Paranoid Disorders	448	216	232	51.80%
HCC67	Quadriplegia, Other Extensive Paralysis	4	0	4	100.00%
HCC68	Paraplegia	7	6	1	14.30%
HCC69	Spinal Cord Disorders/Injuries	8	7	1	12.50%
HCC70	Muscular Dystrophy	1	1	0	0.00%
HCC71	Polyneuropathy	81	43	38	46.90%
HCC72	Multiple Sclerosis	33	22	11	33.30%
HCC73	Parkinson's and Huntington's Diseases	69	51	18	26.10%
HCC74	Seizure Disorders and Convulsions	112	85	27	24.10%
HCC75	Coma, Brain Compression/Anoxic Damage	4	1	3	75.00%
HCC77	Respirator Dependence/Tracheostomy Status	0			
HCC78	Respiratory Arrest	7	5	2	28.60%
HCC79	Cardio-Respiratory Failure and Shock	234	170	64	27.40%
HCC80	Congestive Heart Failure	519	363	156	30.10%
HCC81	Acute Myocardial Infarction	41	29	12	29.30%
HCC82	Unstable Angina and Other Acute Ischemic Heart Disease	80	54	26	32.50%
HCC83	Angina Pectoris/Old Myocardial Infarction	103	54	49	47.60%
HCC92	Specified Heart Arrhythmias	900	523	377	41.90%
HCC95	Cerebral Hemorrhage	24	14	10	41.70%
HCC96	Ischemic or Unspecified Stroke	139	56	83	59.70%
HCC100	Hemiplegia/Hemiparesis	22	15	7	31.80%
HCC101	Cerebral Palsy and Other Paralytic Syndromes	3	1	2	66.70%
HCC104	Vascular Disease with Complications	105	70	35	33.30%
HCC105	Vascular Disease	444	297	147	33.10%

HCC	HCC Label	Sampled	Confirmed	Discrepant	%Discrepant
HCC107	Cystic Fibrosis	1	0	1	100.00%
HCC108	Chronic Obstructive Pulmonary Disease	484	388	96	19.80%
HCC111	Aspiration and Specified Bacterial Pneumonias	29	18	11	37.90%
HCC112	Pneumococcal Pneumonia, Emphysema, Lung Abscess	12	6	6	50.00%
HCC119	Proliferative Diabetic Retinopathy and Vitreous Hemorrhage	24	19	5	20.80%
HCC130	Dialysis Status	18	15	3	16.70%
HCC131	Renal Failure	604	384	220	36.40%
HCC132	Nephritis	8	5	3	37.50%
HCC148	Decubitus Ulcer of Skin	65	49	16	24.60%
HCC149	Chronic Ulcer of Skin, Except Decubitus	185	148	37	20.00%
HCC150	Extensive Third-Degree Burns	1	1	0	0.00%
HCC154	Severe Head Injury	2	1	1	50.00%
HCC155	Major Head Injury	16	6	10	62.50%
HCC157	Vertebral Fractures without Spinal Cord Injury	41	32	9	22.00%
HCC158	Hip Fracture/Dislocation	135	77	58	43.00%
HCC161	Traumatic Amputation	2	2	0	0.00%
HCC164	Major Complications of Medical Care and Trauma	73	43	30	41.10%
HCC174	Major Organ Transplant Status	7	5	2	28.60%
HCC176	Artificial Openings for Feeding or Elimination	20	18	2	10.00%
HCC177	Amputation Status, Lower Limb/Amputation Complications	1	0	1	100.00%

One of the principle challenges of using FFS08 for this purpose is that the CERT sample was not designed to produce a representative sample of diagnoses. As a consequence, for many of the diagnoses and by extension, the HCCs, we have an insufficient sample size to develop reliable discrepancy rates at the HCC level. As shown in Table 2a, discrepancy rates ranged from 0-100%. As expected, sample size was an issue for a number of the HCCs. Nearly half of the HCCs had fewer than 28 observations. [4]

In this context, variation in HCC discrepancy rates could vary for two reasons. First, there are attributes intrinsic to a specific HCC that drive measurement error higher or lower. For example, the underlying diseases may be more difficult to evaluate and therefore the discrepancy rate or measurement error would be higher. In addition to measurement error, insufficient sample size can also contribute to the variation in the discrepancy rate. The purpose of this analysis is to determine the effects of the measurement error on the model calibration. Variation due to insufficient sample size confounds the analysis. We mitigated the impact of small sample size on the analysis as follows. While sample size is an issue in estimating HCC specific error rates, we do have sufficient data to calculate a reliable estimate of an overall HCC expected discrepancy

[4] The minimum number needed to determine a difference from zero at 90 percent confidence with a 0.15 margin of error.

rate. This is calculated by dividing the total number of HCCs sampled by the number of HCCs found to be discrepant. The overall discrepancy rate contains the most information as it is derived from all HCCs. Thus, in the absence of better information, it is the best measure of the probability of finding an HCC discrepant due to a RADV audit.
Using the overall discrepancy rate as our baseline estimate of HCC discrepancy, we can statistically determine, in a context that controls for sample size effects, if the HCC specific rates are significantly different from the overall rate. We apply a statistical test to each HCC level discrepancy rate to determine if it is statistically different at a 95 percent level of confidence from the overall rate. This approach confirms that differences from the overall discrepancy rate are likely due to intrinsic qualities of the HCC rather than small sample size.

When the individual rate is found to be statistically greater than the baseline rate, we assign the HCC to a high category. Conversely, we assign the HCC to a low category when the individual rate is statistically lower than the baseline rate. Rates found to be neither higher nor lower are assigned to the norm category. The discrepancy rates used in the analysis would then be assigned as follows. HCCs in the high and low category would use the mean of the respective category. HCCs in the norm category would use the baseline rate. Following this approach, the discrepancy rates for the baseline, high and low were 0.34, 0.46, and 0.21 respectively. Table 2b, reports the adjusted claim level error rate for each HCC.

FFS Beneficiary Level Discrepancy rates

Each enrollee HCC potentially has multiple claims with independent supportive medical records. Consequently, more medical claims imply the existence of more medical records to confirm an HCC. Clearly, the effective probability of discrepancy depends on the number of claims.

Based on the above, we need to establish the probability that an HCC will be in error for an FFS enrollee given the expected number of associated claims. We will refer to this as the beneficiary level discrepancy rate. As previously described, we established base probabilities that each HCC j will be in error for a given claim, p_j. In table 2b, for each HCC j, we also establish the average number of underlying claims, $\bar{z}$. Accordingly, the beneficiary level discrepancy rate is:

$$p_j^{\bar{z}} = [Claim\ Level\ Discrepancy\ Rate]^{[Average\ Claims]}$$

Table 2c shows the distribution of the beneficiary level discrepancy rates. The adjustment of the claim level discrepancy rate to a beneficiary level discrepancy rate results in a much smaller probability of an HCC being in error as a result of FFS diagnoses error in the FFS claims. As shown in the table, seventy-five percent of the beneficiary level discrepancy rates are 5% or below. The average rate is 3% and the median rate is 2%.

Table 2b: Statistical HCC Base Discrepancy Probabilities and Effective Error Rates

HCC	Cat	Average Number of Claims per HCC* ($\bar{Z}$)	Claim Level Discrepancy Rate**	Beneficiary Level Discrepancy Rate ($p_j^{\bar{z}}$)
1	N	11.4	33.8%	0.000%
2	N	4.1	33.8%	1.223%
5	H	2.7	46.2%	12.259%
7	N	7.0	33.8%	0.050%
8	N	13.4	33.8%	0.000%
9	N	11.0	33.8%	0.001%
10	H	7.0	46.2%	0.432%
15	L	3.3	20.9%	0.581%
16	L	3.3	20.9%	0.581%
17	N	2.6	33.8%	5.716%
18	L	2.3	20.9%	2.888%
19	L	6.2	20.9%	0.006%
21	N	2.5	33.8%	6.814%
25	N	3.7	33.8%	1.739%
26	N	5.2	33.8%	0.349%
27	N	3.2	33.8%	3.127%
31	N	3.5	33.8%	2.278%
32	L	3.7	20.9%	0.285%
33	N	3.7	33.8%	1.898%
37	N	4.7	33.8%	0.621%
38	L	4.9	20.9%	0.045%
44	H	6.1	46.2%	0.891%
45	H	3.8	46.2%	5.332%
51	N	2.2	33.8%	8.717%
52	N	3.6	33.8%	2.084%
54	H	9.0	46.2%	0.098%
55	H	6.1	46.2%	0.889%
67	N	4.5	33.8%	0.726%
68	N	4.1	33.8%	1.225%
69	N	2.4	33.8%	7.163%
70	N	3.3	33.8%	2.851%
71	H	2.6	46.2%	12.931%
72	N	7.3	33.8%	0.037%
73	N	5.2	33.8%	0.340%
74	L	4.7	20.9%	0.063%

HCC	Cat	Average Number of Claims per HCC* ($\bar{Z}$)	Claim Level Discrepancy Rate**	Beneficiary Level Discrepancy Rate $(p_j^{\bar{Z}})$
75	N	2.4	33.8%	7.669%
77	N	3.2	33.8%	3.100%
78	N	2.0	33.8%	11.297%
79	L	5.0	20.9%	0.040%
80	N	6.1	33.8%	0.138%
81	N	3.6	33.8%	2.057%
82	N	2.9	33.8%	4.205%
83	H	2.4	46.2%	15.435%
92	H	7.0	46.2%	0.446%
95	N	4.7	33.8%	0.630%
96	H	4.1	46.2%	4.186%
100	N	3.2	33.8%	3.188%
101	N	2.6	33.8%	5.672%
104	N	4.0	33.8%	1.241%
105	N	3.4	33.8%	2.389%
107	N	6.5	33.8%	0.084%
108	L	5.0	20.9%	0.041%
111	N	3.2	33.8%	3.248%
112	N	2.3	33.8%	7.966%
119	N	2.6	33.8%	5.710%
130	L	4.5	20.9%	0.090%
131	N	4.7	33.8%	0.593%
132	N	2.3	33.8%	8.185%
148	N	2.7	33.8%	5.642%
149	L	4.8	20.9%	0.051%
150	N	2.9	33.8%	4.372%
154	N	2.3	33.8%	7.897%
155	H	3.1	46.2%	8.908%
157	N	3.6	33.8%	2.114%
158	H	7.5	46.2%	0.295%
161	N	3.1	33.8%	3.593%
164	N	2.6	33.8%	5.886%
174	N	10.1	33.8%	0.002%
176	L	2.7	20.9%	1.359%
177	N	3.4	33.8%	2.495%

*The average claim per HCC was calculated using the underlying claims file for the 2004-2005 calibration data set.

**Rate as adjusted for sample size.

Table 2c: Summary Statistics for Effective Probabilities

MAX	15.435%
MIN	0.000%
MEDIAN	1.819%
AVERAGE	3.064%
1st QUARTILE	0.306%
3rd QUARTILE	5.092%

2. Audit Miscalibration Impact

Theoretical Impact to Risk Scores

A key assertion justifying the FFS adjuster has been the assumption that if the underlying FFS claims were subject to a RADV-like audit that resulted in one or more beneficiary HCC changing from 1 to 0, the HCC risk factors from the re-calibrated CMS-HCC model would increase. The argument would then follow that the increased risk factors would increase a plan's payment.

To understand why this is incorrect, we must recognize how the CMS-HCC model is used in payment. First, CMS derives the base capitation rates by county and largely independent of the CMS-HCC model.[5] This establishes the fundamental payment. The CMS-HCC model is then used to adjust those payments by predictive disease-cost load. For example, assume that in a particular county the capitated payment is $\$X$.[6] Plans in that county have the incentive to directly or indirectly select enrollees that will cost less than $\$X$ and exclude those likely to cost more than $\$X$ (adverse selection). Noticeably, plans that have a greater number of likely more expensive enrollees are at greater financial risk. Finally, this adverse selection will increase the cost to the Medicare Trust Fund. To combat this adverse selection and financial risk, CMS created a risk score (R) for each enrollee that will adjust base payments to pay more for likely more expensive enrollees and less for likely less expensive enrollees. So a plan receiving $\$X$ for each enrollee will instead receive $\$X*R$. Since an average cost enrollee will have a risk score of one (R=1) the plan will still receive $\$X$ ($\$X*R=\$X*1=\$X$) for an average enrollee. The CMS-HCC model bases the risk score (R) on the possession of demographic and chronic disease characteristics. Although fundamentally based on expenditures, the regression is adjusted such that the HCC and demographic factors will provide an average risk score of one on the calibrating FFS dataset. Thus, by definition of the CMS-HCC model, the average FFS enrollee will never have a risk score of greater or less than one.

On the other hand, the average MA enrollee risk score is not constrained to any particular value. If the MA population is generally unhealthy relative to the FFS population, the average risk score for an MA enrollee will be greater than one. If the MA population is

[5] The county rates are primarily derived from FFS data and the bidding process.

[6] To simplify the discussion, for illustrative purposes we refer to a single county rate. In the actual payment there is a specific rate for each plan per county.

generally healthier than the FFS population, the average risk score for an MA enrollee will be less than one.

Collectively, the calibration of the CMS-HCC model cannot, by definition, cause the risk factors to increase such that the average risk score of the calibrating FFS data is greater than one. Therefore, calibrating the CMS-HCC model on an audited dataset would simply change the risk factors relative to each other. It is conceptually possible that the MA population utilizes more of the risk factors that increase under the audited calibration methodology and thus increase MA payments. However, there is no theoretical reason for this to be the case. Nonetheless, we will examine the issue empirically.

Impact Measurement

The objective of the analysis is to compare the risk scores that come out of the original uncorrected data model to those that come out of the RADV-like corrected data model. The difference, when applied to MA enrollees, would be an estimate for the bias. To perform the analysis of the impact of the RADV-like audit, we proceed as follows. First, we estimate a proxy for the CMS-HCC model on the original uncorrected dataset and estimate the original factors. We then apply these original factors to the MA sample and estimate original risk scores for the uncorrected calibrating dataset.

Next, we carry out a series of steps to simulate the same MA enrollees' corrected risk scores that would have resulted from calibrating the model on a FFS dataset that was RADV-like corrected.

We apply these probabilities to the calibrating dataset in such a manner that each existing HCC of each enrollee is subject to a $p_j^{\bar{z}}$ probability of being redefined as discrepant. We then estimate the CMS-HCC model on the simulated corrected data. In the next step, we take the new coefficients and apply them on the original FFS data set, normalizing a new set of relative factors to one. The new relative factors are then applied to the beneficiary profiles in the MA sample to calculate new risk scores based on coefficients unaffected by FFS diagnoses error in the calibration. All other things being equal, the difference between the error-free risk score and the original risk score is the impact of FFS diagnoses error in the calibration. The normalization process is illustrated as follows:
A simplified risk adjustment model is

$$E = \alpha DEM + \beta HCC1 + \delta HCC2 + \epsilon$$

Here, E is the enrollee expenditures from period (t), DEM is a demographic characteristic (such as sex (male=1, female=1)), *HCC1* and *HCC2* are equal to one if the particular condition is present and zero if it is not present at time (*t-1*) and ϵ is a error term satisfying the assumptions of the Ordinary Least Squares Model.

Since E is dollar expenditures, the estimates coefficient of the model $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ will also be dollar expenditures. $\hat{\alpha}$ is the incremental dollar cost of being female over male; $\hat{\beta}$ is the incremental dollar cost of having HCC1 and $\hat{\delta}$ is the incremental dollar cost of having

HCC2. We start with an enrollee profile of five FFS enrollees. The actual FFS expenditures for enrollees 1-5 are E1, E2, E3, E4 and E5.

$$\begin{bmatrix} E & DEM & HCC1 & HCC2 \\ E1 & 1 & 0 & 1 \\ E2 & 1 & 1 & 1 \\ E3 & 0 & 1 & 1 \\ E4 & 1 & 0 & 1 \\ E5 & 1 & 1 & 0 \end{bmatrix}$$

We estimate the model and obtain estimated coefficient of $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ which are dollar estimates. To create the risk scores, we apply these estimates to the enrollee profiles derived from the uncorrected FFS data.

$$\widehat{\boldsymbol{X}} = \begin{bmatrix} Bene & DEM & HCC1 & HCC2 \\ 1 & \hat{\alpha} & 0 & \hat{\delta} \\ 2 & \hat{\alpha} & \hat{\beta} & \hat{\delta} \\ 3 & 0 & \hat{\beta} & \hat{\delta} \\ 4 & \hat{\alpha} & 0 & \hat{\delta} \\ 5 & \hat{\alpha} & \hat{\beta} & 0 \end{bmatrix}$$

Then we sum the expected expenditure for each enrollee:

$$\begin{array}{ccc} Bene & Sum & Estimate \\ 1 & \hat{\alpha} + \hat{\delta} & = \hat{E}1 \\ 2 & \hat{\alpha} + \hat{\beta} + \hat{\delta} & = \hat{E}2 \\ 3 & \hat{\beta} + \hat{\delta} & = \hat{E}3 \\ 4 & \hat{\alpha} + \hat{\delta} & = \hat{E}4 \\ 5 & \hat{\alpha} + \hat{\beta} & = \hat{E}5 \end{array}$$

To normalize into risk scores, we first obtain the average predicted expenditure of the enrollees:

$$\bar{E} = \frac{\hat{E}1 + \hat{E}2 + \hat{E}3 + \hat{E}4 + \hat{E}5}{5}$$

Next, we divide the expected expenditures for each enrollee by the average expenditure of all enrollees:

$Bene$	Sum	$Risk\ Score$
1	$\frac{\hat{E}1}{\bar{E}}$	$= R1$
2	$\frac{\hat{E}2}{\bar{E}}$	$= R2$
3	$\frac{\hat{E}3}{\bar{E}}$	$= R3$
4	$\frac{\hat{E}4}{\bar{E}}$	$= R4$
5	$\frac{\hat{E}5}{\bar{E}}$	$= R5$

The normalized risk scores $R1, R2, R3, R4, R5$ must, on average, equal one. It is important to note that this is true regardless of the enrollee profile used to calibrate the model and the level of "correctness" of the enrollee profile. It also follows that the estimates $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ will always be such that the calibrating dataset's sum of actual expenditures will be equal to the expected expenditures. Consistent with the above, the risk factor for *DEM* is $\frac{\hat{\alpha}}{\bar{E}}$, for *HCC1* is $\frac{\hat{\beta}}{\bar{E}}$ and for *HCC2* is $\frac{\hat{\delta}}{\bar{E}}$. Finally, corrected factors are applied to the MA enrollee sample to establish the corrected MA risk scores. For each enrollee *k* in the calibrated dataset, we calculate the percentage difference in risk score from the original score:

$$d_k = (corrected\ risk\ score_k - original\ risk\ score_k)/original\ risk\ score_k$$

If there are N enrollees in the MA sample dataset, then the average of the percentage difference in risk score across these enrollees is an estimate of the bias,

$$\bar{d}_k = \frac{1}{N}\sum_{k=1}^{N} d_k$$

For the bias to be impactful against the MA plans, it should be a large, positive percentage.

3. Simulation and Empirical Results

In order to estimate variation in $\bar{d}_k$, we ran a simulation calculating 50 cases of the impact analysis. For each case, using a random number generator, we applied the beneficiary level discrepancy rates to get 50 independent corrected FFS data files. Each file was then used to re-calibrate the CMS-HCC model, re-normalize a new set of risk scores and estimate a new value of $\bar{d}_k$. This resulted in a distribution of 50 bias estimates, $\bar{d}_k^1,, \bar{d}_k^2,, \bar{d}_k^3, \cdots,, \bar{d}_k^{50}$.

This distribution of relative differences is reported below. The mean bias estimate of these 50 simulations was -0.08% and the mean bias estimate was between -.07%,-.09% at a 95% level of confidence.

An examination of Table 3 and Figure 1 makes it clear that the chance of any one of these simulations 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. This shows that FFS diagnosis error in the FFS calibration data does not manifest a negative payment bias to MA plans.

Table 3. Distributional Estimate of Bias

Distributional Statistic	Bias Estimate
Mean Relative Difference	-0.08%
Median Relative Difference	-0.09%
Minimum Relative Difference	-0.23%
Maximum Relative Difference	0.01%
25th Percentile	-0.10%
7th Percentile	-0.05%

Figure 1 Sampling Distribution of the Relative Difference

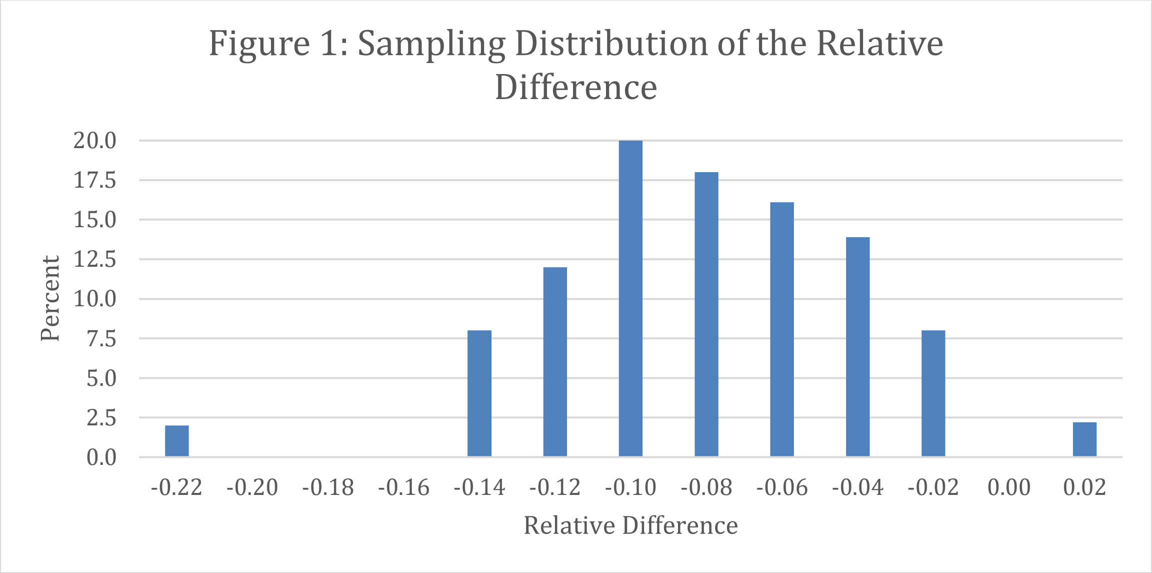
Figure 1: Sampling Distribution of the Relative Difference
Percent
Relative Difference