Online Appendix for

Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Reply

Hanming Fang   Qing Gong

In Section A of this appendix, we show that the results in Fang and Gong (2017) remain qualitatively the same after correcting the data and coding errors in our time-per-service estimation. In Section B, we discuss the noise and potential biases in Matsumoto's quantification of service over-counting in the utilization data.

A FG Results Revisited: Corrected Data and Methodology

In this section, we discuss the other comments Matsumoto (forthcoming) had on the time-per-service estimation in FG. We first show that there's a modest decline in the number of flagged physicians with the corrected data (Zuckerman et al., 2016) and methodology, although capturing the *level* of overbilling was never intended as a key objective of FG. More importantly, we show that the results in FG's subsequent analysis remain qualitatively the same, including those on the characteristics of flagged physicians, on the representativeness of specialties among flagged physicians, on the composition of workloads of flagged versus unflagged physicians, and on the coding patterns of flagged physicians.

A.1 Suggested changes to the time-per-service estimation

Matsumoto (forthcoming) pointed out the following issues with FG's estimation of the time needed to perform each service. We address each of them below:

No total service time estimates for type-I timed codes. Matsumoto noted that we did not estimate the total service time for the first type of timed codes, which differed from the description in the paper. We thank Matsumoto for identifying the inconsistency. We apologize for the inaccurate description, but we do not see a compelling reason to provide an estimate of the total service time because these codes already had a suggested or required time.

No "group average method" estimates for type-I timed codes. As much as we agree with Matsumoto that there was an inconsistency between the text and the codes, we are still hesitant to provide an estimated time for codes that already have a suggested or required time. Moreover, type-I timed codes are overwhelmingly "evaluation and management" (E/M) services. For codes in the E/M group, in particular, we do not find it necessary to re-estimate a service time if the code already has time in the definition, and do not find it appropriate to impose a time if the E/M code does not require a service time (e.g. code 99234 for "observation," which we assigned a time of 0 to err on the conservative side).

"Regression method" estimates only available for type-II timed codes. This was unintended and resulted from a typo in our Stata codes. We first obtained the time-per-service estimates using the group average method and named it `timeTotal1`. We then *intended* to obtain the estimates using the alternative, regression-based method if a service has a non-missing `timeTotal1`. But we mis-specified `timeTotal1` as `timeTotal`, which was only available for type-II timed codes.

Typos in data input. There were two typos when the table in Appendix B of Zuckerman et al. (2014) was typed into a spreadsheet.

- Code 22612 had a service time of 150 minutes, which was entered as 160 minutes. This had a negligible impact on the results.
- Code 44204 was entered as 44203 (the former had a wRVU of 26.42, the latter 4.44). Matsumoto correctly pointed out that this *could* bias the time-per-wRVU estimate upward. However, code 44203 was dropped from the sample in a later step, thus never used in the imputation and did *not* affect the results in any way.

Before we present the new results, we would like to also address another change Matsumoto made in the appendix E of his Comment. In E.1, he flagged physicians only using the intra-service hours; in E.2, he flagged physicians based on the *unique number of beneficiaries* on a given day. We find both changes to be overly conservative and rely on strong assumptions that (i) physicians do not spend any time on pre- and post-service and (ii) physicians only provide one unit of a particular service to each beneficiary on any given day. Matsumoto's findings under those assumptions could serve as lower-bound estimates, but our concern is that they can too conservative to be useful.

A.2 Results in FG remain qualitatively the same after corrections

We corrected the data and coding errors that Matsumoto pointed out in his Comment. We did not, however, provided estimates for timed codes for reasons discussed above. We find that the results in FG remain qualitatively the same after these changes.

Table A1 replicates Table 2 of FG. The number of flagged physicians in FG at the 100-hours/week threshold were 2,292 and 2,120 for the year 2012 and 2013, respectively. These flagged physicians were 2.71% and 2.55% of all physicians with at least 20 hours/week of Medicare Part B FFS services. After correcting the data and coding errors, we now flag 1,845 and 1,683 for these two years, a roughly 20% reduction. The fractions of flagged physicians among those with 20+hours/week are virtually the same as those in FG: they are now 2.67% and 2.50%, respectively. This is because the total number of physicians estimated to have worked 20+hours/week reduced from 96,033 in FG to 78,165 after the correction, a 19% reduction.

Hours threshold	80+		100+		112+		168+	
Year	2012	2013	2012	2013	2012	2013	2012	2013
Number of physicians flagged	3250	2963	1845	1683	1395	1247	520	450
Fraction in physicians working 20+ hr/week	4.694	4.392	2.665	2.495	2.015	1.849	0.751	0.667
Fraction in all physicians	0.521	0.475	0.296	0.270	0.224	0.200	0.0830	0.0720
Fraction of zero-time codes (flagged)	10.59	10.56	8.564	8.779	7.820	7.773	5.987	6.304
Fraction of zero-time codes (unflagged*)	19.66	19.81	19.54	19.70	19.49	19.65	19.37	19.54
Fraction of wRVU from zero-time codes (flagged)	0.612	0.661	0.563	0.636	0.577	0.595	0.421	0.479
Fraction of wRVU from zero-time codes (unflagged*)	1.758	1.718	1.738	1.701	1.731	1.696	1.719	1.685
Fraction of volume from zero-time codes (flagged)	8.985	8.754	6.910	6.715	5.829	5.454	3.050	3.434
Fraction of volume from zero-time codes (unflagged*)	19.25	19.43	19.11	19.30	19.06	19.25	18.93	19.12
Fraction of revenue from zero-time codes (flagged)	5.124	5.511	4.140	4.418	3.442	3.453	1.948	2.101
Fraction of revenue from zero-time codes (unflagged*)	9.154	9.479	9.104	9.437	9.087	9.424	9.034	9.371
Total number of physicians working 20+ hr/week	78,165							
Total number of physicians	623,959							

Table A1: Number and fraction of physicians flagged

NOTES: The table reports the number and fraction of flagged physicians in 2012 and 2013. "Hours threshold" shows the cutoff number of hours billed per week above which a provider is flagged. "Fraction in physicians working 20+ hr/week" shows the fraction of flagged physicians among physicians who billed at least 20 hours per week in the same calendar year. "Fraction in all physicians" shows the fraction among all physicians in our sample, which covers the vast majority of physicians. "Zero-time codes" are codes for which positive time needed estimates are not available and account for 25 percent of all HCPCS codes. wRVU is the physician work RVUs that are specific to each HCPCS code and reflect the amount of work (primarily time) required to furnish each service. *Unless otherwise specified, "unflagged" refers to unflagged physicians whose estimated weekly hours worked are above 20 in 2012 or 2013 or both. All fractions are measured in percent.

Table A2 replicates Table 3 in FG, where we categorize the flagged physicians by their flag status in each calendar year. Just as in FG, we still find that roughly 31% of the physicians flagged in any year are only flagged in 2012, 25% of them are only flagged in 2013, and the remainder are flagged in both years.

Hours threshold	80+		100+		112+		168+	
Year(s) flagged	Count	Share (percent)	Count	Share (percent)	Count	Share (percent)	Count	Share (percent)
2012 only	906	27.88	578	31.33	474	33.98	204	39.23
2012 and 2013	2344		1267		921		316	
2013 only	619	20.89	416	24.72	326	26.14	134	29.78

Table A2: Flag patterns across time

NOTES: "Hours threshold" shows the cutoff number of hours billed per week above which a provider is flagged. "Count" columns report the number of physicians flagged (in 2012 only, in both years, or in 2013 only). "Share" columns show the fraction of physicians who are only flagged in 2012 (2013) among all physicians flagged in that year (percent).

Table A3 replicates Table 4 in FG, where we compared the characteristics of physicians by their flag status. We still find that physicians who have ever been flagged are significantly more likely to be male, not an MD, working in smaller practices, with fewer hospital affiliations, and provided fewer types of E/M services. The magnitude of the differences varies slightly from those in FG but are qualitatively the same. Again, these are similar to what Cutler et al. (2015) found to be the characteristics of physicians who “consistently and unambiguously recommended intensive care beyond those indicated by current clinical guidelines.”

Table A4 replicates Table 5 in FG. First, recall that the *Specialty Flag Index* is defined as

$$\mathrm{SFI}_s = \frac{100 \times \Pr(s|\text{flagged})}{\Pr(s|\text{flagged}) + \Pr(s|\text{unflagged})} \tag{A1}$$

where the conditional probability $\Pr(s|\text{flagged})$ is defined as the fraction of flagged physicians in specialty s among all flagged physicians, and $\Pr(s|\text{unflagged})$ is the fraction of unflagged physicians in specialty s among all unflagged physicians. An SFI above 50 indicates over-representation of the specialty among the flagged physicians. Table A4 shows specialties with the highest SFIs *and* have at least 50 flagged physicians in any year. The resulting specialties are essentially the same: we still find optometry, dermatology, and ophthalmology to be the most over-represented among flagged physicians. The only difference from FG is the absence of pathology. This is because all but one pathology service codes get a time of 0 in our new time-per-service estimation. As a result, only one physician with a reported specialty of pathology is flagged.

Table A5 replicates Table 6 in FG, where we unpack and compare the services of flagged and unflagged physicians. Again, we still find that flagged physicians provide significantly more services (both in total and for each patient), but fewer services per estimated hour; they also have more patients (both in total and per calendar day), but fewer patients per estimated hour; they receive more Medicare payments for each service (noisily estimated) and per patient, but receive lower payments per estimated hour. The findings are qualitatively similar to those in FG and suggest that the flagged physicians are more likely to report higher intensity and/or more time-consuming services.

Table A6 replicates Table 7 in FG. Recall that *Overbilling Potential Factor 1* is defined as

$$\mathrm{OPF1}_i \equiv \frac{(\text{Total revenue})_i}{(\text{Fair revenue})_i} = \frac{(\text{Total revenue})_i}{(\text{Fair hourly revenue})_i \times (\text{Fair hours})}, \tag{A2}$$

where “Total revenue” is the observed annual Medicare Part B FFS payments of physician i; “Fair hourly revenue” is the predicted hourly revenue for physician i based on an OLS regression of the hourly revenues of *unflagged* physicians on observables, which include physician gender, credential, years of experience, and a full set of specialty, HRR, and year fixed effects; and “Fair hours” is

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Ever	2012	2013	2012 only	Both	2013 only	Never
1(Male)	0.019**	0.022***	0.018*	0.018	0.024**	0.002	0.861
	[0.008]	[0.008]	[0.010]	[0.012]	[0.010]	[0.017]	
1(MD)	-0.225***	-0.203***	-0.179***	-0.369***	-0.132***	-0.330***	0.838
	[0.034]	[0.037]	[0.034]	[0.049]	[0.037]	[0.037]	
Experience (years)	0.087	0.546*	-0.668**	2.294***	-0.243	-1.943***	24.531
	[0.274]	[0.294]	[0.289]	[0.524]	[0.306]	[0.614]	
# providers in group	-49.172***	-45.504***	-47.729***	-53.369***	-41.969***	-65.453***	73.335
	[5.543]	[5.492]	[5.792]	[10.120]	[5.460]	[10.665]	
# hospital affiliations	-1.572***	-1.546***	-1.524***	-1.734***	-1.467***	-1.707***	2.795
	[0.134]	[0.136]	[0.135]	[0.216]	[0.133]	[0.182]	
1(in Medicare)	0.011	0.013	0.011	0.011	0.014	0.001	0.863
	[0.007]	[0.008]	[0.008]	[0.011]	[0.009]	[0.018]	
1(in ERX)	-0.068***	-0.047**	-0.036*	-0.163***	0.002	-0.157***	0.528
	[0.018]	[0.020]	[0.019]	[0.026]	[0.021]	[0.027]	
1(in PQRS)	0.016	0.025	0.039**	-0.051**	0.060***	-0.024	0.364
	[0.015]	[0.016]	[0.017]	[0.023]	[0.018]	[0.027]	
1(in EHR)	-0.070***	-0.072***	-0.072***	-0.063**	-0.076***	-0.062***	0.466
	[0.014]	[0.016]	[0.015]	[0.025]	[0.018]	[0.024]	
Types of codes 2012	0.917	1.960*	2.246**	-3.250**	4.213***	-3.919***	21.325
	[0.900]	[1.009]	[0.876]	[1.363]	[0.975]	[0.922]	
Types of codes 2013	1.134	1.844*	2.833***	-4.126***	4.440***	-2.236**	21.178
	[0.920]	[1.017]	[0.901]	[1.288]	[0.988]	[1.018]	
Types of E/M codes 2012	-2.543***	-2.500***	-2.526***	-2.631***	-2.454***	-2.771***	7.113
	[0.223]	[0.217]	[0.231]	[0.352]	[0.218]	[0.347]	
Types of E/M codes 2013	-2.516***	-2.520***	-2.433***	-2.801***	-2.406***	-2.539***	7.076
	[0.228]	[0.219]	[0.240]	[0.341]	[0.225]	[0.371]	
Num. of physicians in group	2,261	1,845	1,683	578	1,267	416	75,904

Table A3: Characteristics of flagged physicians vs. unflagged physicians, conditional on Hospital Referral Region (HRR)

NOTES: The table summarizes the difference in physician characteristics between flagged subgroups and the never-flagged subgroup (means reported in the last column) conditional on HRR. We restrict the sample to physicians billing at least 20 hours per week in at least one year. The number in each cell is the estimated coefficient from an OLS regression on the subset of physicians who are either never flagged, or have the flag status indicated by the column heading. We use the physician characteristic in the corresponding row as the dependent variable, and the flag status dummy as the explanatory variable together with HRR fixed effects. Physician experience is imputed from the year of graduation. # providers in group refer to the number of providers in the group practice where the billing physician works and it is 1 if the billing physician works in a solo practice. The number of hospital affiliations is top coded at 5 in the data. 1(in Medicare) is an indicator that the physician accepts Medicare-approved payment amount. 1(in ERX) is an indicator for participation in the Medicare Electronic Prescribing (eRx) Incentive Program, which encourages eRx. 1(in PQRS) is an indicator for participation in the Medicare Physician Quality Reporting System Incentive Program, which provides financial incentives to physicians who report quality measures. 1(in EHR) is an indicator for participation in the Medicare Electronic Health Record (EHR) Incentive Program, which uses financial incentives to reward the adoption of certified EHR technology. Standard errors clustered at the HRR level are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

		Num. unflagged		Num. flagged		SFI	
Specialty\Year	Fraction in all	2012	2013	2012	2013	2012	2013
Optometry	2.563	1437	1541	567	463	94.23	93.18
Dermatology	6.401	4413	4393	591	611	84.71	86.34
Ophthalmology	10.11	7637	7654	265	248	58.94	59.55
Nephrology	6.381	4915	4923	73	65	38.06	37.50
Cardiology	12.73	9878	9899	71	50	22.92	18.67
Internal Medicine	15.60	12117	12122	78	73	21.03	21.49
All physicians		76320	76482	1845	1683		

Table A4: Physician specialties and flag status

NOTES: The table shows seven specialties with the highest SFI, defined in equation (A1), among the specialties with at least 50 flagged physicians. "Fraction in all" shows the percentage of physicians in a specialty among all physicians in our sample (restricted to physicians billing at least 20 hours per week in at least one year). The last row labeled "All physicians" shows the number of flagged and unflagged physicians by year in our sample.

	Flagged		Unflagged	
Year	2012	2013	2012	2013
Num. of services provided	6439.077***	6358.656***	4726	4654
	[536.548]	[446.440]		
Num. of services per patient	1.889***	1.438***	2.300	2.264
	[0.276]	[0.201]		
Num. of services provided per hour	-1.592***	-1.639***	3.050	3.072
	[0.069]	[0.063]		
Num. of patients	1939.796***	2137.074***	2438	2423
	[258.149]	[227.409]		
Num. of patients per day	5.300***	5.855***	6.661	6.638
	[0.705]	[0.623]		
Num. of patients per hour	-1.037***	-1.046***	1.611	1.632
	[0.044]	[0.042]		
Medicare payment per service ($)	1.008	4.547	78.36	77.08
	[3.860]	[3.620]		
Medicare payment per patient ($)	28.036***	28.175***	155.6	152.3
	[6.818]	[6.372]		
Medicare payment per hour ($)	-74.018***	-73.287***	183.1	180.2
	[5.238]	[4.397]		
N	1,845	1,683	76,320	76,482

Table A5: Volume of services supplied conditional on Hospital Referral Regions: flagged vs. unflagged physicians

NOTES: The table compares the volume of services furnished by physicians of different subgroups. We restrict the sample to physicians billing at least 20 hours per week in at least one year. The first two columns report the estimation results from OLS regressions using the volume measure in that row as the dependent variable, and the flag dummy as the explanatory variable, together with HRR fixed effects. Standard errors clustered at the HRR level are in brackets. ** $p < 0.05$, *** $p < 0.01$. The last two columns report the means of the two unflagged groups as references. "Num. of patients" is an overestimation of the actual number of distinct patients due to data limitation, because it is the physician-level sum of the number of distinct patients for each code the physician billed. Hence a patient receiving more than one type of service will be counted multiple times. "Num. of patients per day" is the average number of patients per day assuming 366 (365, respectively) working days in year 2012 (2013, respectively). "Per hour" statistics are calculated using the estimated total hours worked of each physician.

set to be 8 hours per day multiplied by 365 days. An OPF1 above 1 captures the excess revenue relative to the predicted "fair" amount that is not explained by observed physician and local market characteristics.

Overbilling Potential Factor 2 is defined under the assumption that the goal of overbilling is to achieve the same revenue with fewer *actual* hours. For each flagged physician i, we have:

$$(\text{True Hours})_i \ \times \ (\text{Fair hourly revenue})_i = (\text{Reported hours})_i \times (\text{Reported hourly revenue})_i.$$

Thus,

$$\text{OPF2}_i \equiv \frac{(\text{Reported hours})_i}{(\text{True hours})_i} \equiv \frac{(\text{Fair hourly revenue})_i}{(\text{Reported hourly revenue})_i}, \tag{A3}$$

where, as in equation (A2), "Fair hourly revenue" is the predicted hourly revenue for physician i based on an OLS regression of the hourly revenues of *unflagged* physicians on observables, which include physician gender, credential, years of experience, and a full set of specialty, HRR, and year fixed effects; "Reported hourly revenue" is simply the total revenue received by physician i divided by the total hours reported by i, which we estimated based on i's claims.

We find OPFs that are highly similar to those found in FG. In particular, flagged physicians have OPFs that are several times the OPFs of their unflagged counterparts.

	Flagged Physicians	Unflagged Physicians
Reported hourly revenue ($)	102.173	178.166
	(1.464)	(0.270)
Predicted hourly revenue ($)	138.796	178.184
	(1.057)	(0.154)
Overbilling Potential Factor 1	1.690	0.558
	(0.025)	(0.001)
Overbilling Potential Factor 2	6.991	1.178
	(0.186)	(0.004)
N	3,528	152,802

Table A6: Hourly revenues and Overbilling Potential Factors (OPFs)

NOTES: The table compares the hourly revenues and OPFs (defined in equations (A2) and ()) between flagged and unflagged physicians. We restrict the sample to physicians billing at least 20 hours per week in at least one year. Reported hourly revenues are total revenues divided by total hours reported in one calendar year. Predicted hourly revenues are obtained by first regressing reported hourly revenues on observables (gender, credential, years of experience, a full set of specialty, HRR, and year fixed effects) using the unflagged sample, and then predicting a "fair" hourly revenues for all physicians based on the regression estimates. Standard errors are reported in parentheses.

Finally, Table A7 replicates Table 8 in FG, where we examined how the *distribution of intensities within the same cluster of codes* differed between flagged and unflagged physicians. Similar to the findings in FG, columns (1) through (4) show that flagged physicians are more likely to report mid- and high-intensity codes, *conditional on the code cluster (i.e. the service).* Additionally, this tendency only persists when the marginal gain from "upcoding" is relatively high (column (6)), but is reversed when the marginal gain from doing so is relatively low (column (5)).

	(1)	(2)	(3)	(4)	(5)	(6)
	$K = 3$	$K = 4$	$K = 5$	All K	All K & below average	All K & above average
Flagged	333.7***	417.6***	26.88*	225.8***	220.5***	200.9***
	[82.05]	[161.2]	[13.83]	[17.27]	[27.36]	[20.44]
Intensity=2	260.5***	170.3***	23.47***			
	[3.020]	[9.342]	[3.575]			
Intensity=3	136.7***	162.6***	254.0***			
	[2.642]	[9.942]	[3.761]			
Intensity=4		-59.14***	245.2***			
		[10.31]	[3.591]			
Intensity=5			16.16***			
			[3.413]			
Flagged × (intensity=2)	664.0***	116.5	136.6***			
	[103.1]	[216.8]	[19.22]			
Flagged × (intensity=3)	159.0*	362.5	181.7***			
	[93.06]	[236.8]	[23.72]			
Flagged × (intensity=4)		403.5	20.84			
		[260.1]	[22.34]			
Flagged × (intensity=5)			33.91			
			[26.94]			
Mid-intensity				249.9***	20.71***	352.3***
				[1.866]	[1.339]	[2.994]
High-intensity				152.6***	28.85***	181.4***
				[1.597]	[1.294]	[2.633]
Flagged × Mid-intensity				139.9***	-105.4***	314.1***
				[26.55]	[27.79]	[39.05]
Flagged × High-intensity				-39.82*	-99.11***	35.69
				[23.73]	[27.56]	[34.08]
HRR	Y	Y	Y	Y	Y	Y
Code cluster	Y	Y	Y	Y	Y	Y
Year	Y	Y	Y	Y	Y	Y
Adjusted R^2	0.208	0.057	0.178	0.164	0.154	0.085
Observations	371,947	54,167	516,165	942,279	470,431	471,848

Table A7: Billing patterns and code intensity level

NOTES: The table reports OLS estimates of the partial effects of code intensity on the number of times the code is filed. We restrict the sample in all specifications to physicians billing at least 20 hours per week in at least one year, and HCPCS codes in the 28 well-defined clusters. Furthermore, Columns (1) to (3) are only using the subsamples of code clusters with 3, 4, and 5 levels of intensities, respectively. Column (4) pools codes in all clusters together, and re-classify intensities to low, middle, and high as specified in Table 9 of Fang and Gong (2017). Columns (5) and (6) use the subsample of codes with below- and above-average marginal increase in work RVUs between two adjacent intensity levels, respectively. Physician characteristics, HRR fixed effects, code cluster fixed effects, year fixed effects, and a constant term are included in all specifications but not reported. Standard errors clustered at the physician level are in brackets. ** $p < 0.05$, *** $p < 0.01$.

B Noises and Potential Biases in Matsumoto's Hours Estimator

B.1 Matsumoto's quantification of service over-counting with 5% claims data

Matsumoto (forthcoming) discussed possible ways of service over-counting in the utilization data used by FG. He then used the Part B FFS claims on a 5% sample of Medicare beneficiaries (*5% claims data* henceforth) to quantify the extent of potential over-counting in the following steps:

1. *Without service count adjustments,* estimate the total service time for each physician i, $\hat{T}_i^u$, as follows:

   (a) Estimate the time-per-wRVU, $\hat{\alpha}_c$, for each service code c (same as FG).

   (b) Calculate the sum of time needed for the *sampled* services of i:

   $$\hat{t}_i = \sum_{c \in \mathcal{C}_i^s} wRVU_c \times \hat{\alpha}_c \times n_{ic}^u \tag{A4}$$

   Note that, taking $\hat{\alpha}_c$ as given, the "true" total service time is

   $$\hat{T}_i^* = \sum_{c \in \mathcal{C}_i} wRVU_c \times \hat{\alpha}_c \times N_{ic} \tag{A5}$$

   where $\mathcal{C}_i^s$ is the set of physician i's *sampled* services in the 5% claims data, n_{ic}^u is the *unadjusted* number of times service c was provided by i in the 5% claims data. $\mathcal{C}_i$ and N_{ic} are the counterparts of $\mathcal{C}_i^s$ and n_{ic}^u in the universe of claims

   (c) Use $\frac{A_i}{a_i}$ to estimate the inverse sampling rate *for each physician* i, where a_i is the sum of Medicare allowed amounts for physician i's *sampled* services, and A_i is the total Medicare allowed amount for all of i's services (disclosed separately by CMS). The allowed amounts roughly follows these formula:[1]

   $$a_i = \sum_{c \in \mathcal{C}_i^s} TotalRVU_c \times CF \times n_{ic}^u, \quad A_i = \sum_{c \in \mathcal{C}_i} TotalRVU_c \times CF \times N_{ic} \tag{A6}$$

   where $TotalRVU_c = wRVU_c + peRVU_c + mpRVU_c$: $peRVU_c$ reflects the practice expense component of the service, $mpRVU_c$ reflects the cost of malpractice insurance. CF is a common conversion factor that applies to all services and physicians ($34.023/RVU in 2013).

[1] Per the Physician Fee Schedule, each of the three RVUs in the formula is also multiplied by a potentially different geographical price index, which we have suppressed for simplicity.

(d) Estimate the time needed for *all* services of i:

$$\hat{T}_i^u = \hat{t}_i^u \times \frac{A_i}{a_i} \tag{A7}$$

2. *With service count adjustments,* estimate the total service time for each physician i, $\hat{T}_i^a$, as follows:

$$\hat{T}_i^a = \hat{t}_i^a \times \frac{A_i}{a_i} = \sum_{c \in \mathcal{C}_i^s} wRVU_c \times \hat{\alpha}_c \times n_{ic}^a \times \frac{A_i}{a_i} \tag{A8}$$

where n_{ic}^a is the adjusted number of times physician i provided service c in the 5% claims data, and $\hat{t}_i^a$ is the resulting estimated time needed for i's sampled services.

3. Matsumoto use the reduction in the number of flagged physicians to quantify the effect of potential service over-counting:

$$\Delta(\#\text{ flagged}) = \sum_{i \in \mathcal{I}_s} \mathbf{1}\left(\hat{T}_i^u > 100 > \hat{T}_i^a\right) \tag{A9}$$

where $\mathcal{I}_s$ is the set of physicians in the 5% claims data. This is equivalent to counting the number of physicians who would be flagged using unadjusted hours, but not using adjusted hours.

B.2 Noisy estimates due to sampling

There are several sources of bias in Matsumoto's estimation of the total hours. First, even if $\hat{T}_i$ (adjusted or unadjusted) is unbiased, it will be noisy due to sampling variations.

To see this, consider the following example. Also assume the best-case scenario (for Matsumoto's estimates) where (i) the researcher has a 5% truly random sample of each provider's services; (ii) $\frac{A_i}{a_i}$ is an unbiased estimator for the "true" inverse sampling ratio, $\frac{T_i^*}{t_i}$, for each physician; (iii) $\frac{A_i}{a_i}$ and $\hat{t}_i$ are uncorrelated. Then the estimated total hours, $\hat{T}_i$, which is a random variable itself due to sampling, will have a mean of T_i^*. But for a physician with a latent $T_i^* = 100$, this implies that $\Pr(\hat{T}_i > 100) = 0.5$, i.e. the physician who *should* be flagged only has a 50% chance of being flagged. *Assuming the sampling error is not positively correlated with a physician's true hours*, it is true that physicians with $T^* > 100$ would be falsely unflagged due to the sampling error with a lower probability. But if the sampling error increases with the true hours, then physicians with $T^* > 100$ could also be more likely to be falsely unflagged. Without a good estimate of the full distribution conditional on $T^* > 100$, it is hard to gauge the exact impact of the sampling error.

This error, inherent in using a sample, could not only explain Matsumoto's result that *some* *of the 2120 physicians flagged by FG are no longer flagged using the 5% claims data*, but also *his other result that 743 (28.4%) of the 2614 physicians flagged in his claims data were not flagged by FG.*[2] Moreover, it is precisely physicians whose hours are closer to the 100-hour threshold who are more prone to be falsely flagged or unflagged. Hence, after Matsumoto adjusted the service counts downward, these physicians' $\hat{T}_i^a$ would move down and away from the threshold, thereby losing the flag status if they were flagged based on $\hat{T}_i^u$.

Of course, we do not mean to say that this is the *only* reason that a fraction of the flagged physicians in Fang and Gong became unflagged by Matsumoto. We merely want to point out that Matsumoto's use of a 5% claims sample introduces sampling noise, which has the potential to mechanically *unflag some physicians that were flagged by FG, and flag some physicians that were not flagged by FG.*

B.3 Potential biases in estimated total hours

The second issue with Matsumoto's estimation of the total hours from the 5% claims data is that $\hat{T}_i$ (adjusted or unadjusted) is bound to be *biased* unless $\hat{t}_i$ and $\frac{A_i}{a_i}$ are uncorrelated. This bias will be present even if $\frac{A_i}{a_i}$ is an unbiased estimator of the "true" inverse sampling ratio. That is:

$$\mathbb{E}\left[\hat{T}_i\right] = \mathbb{E}\left[\hat{t}_i \times \frac{A_i}{a_i}\right] \neq \hat{t}_i \times \frac{T_i^*}{\hat{t}_i} = T_i^* \tag{A10}$$

Furthermore, there is no guarantee that $\frac{A_i}{a_i}$ will be unbiased, either. The "true" sampling rate is the $\frac{T_i^*}{\hat{t}_i}$, and both T_i^* and $\hat{t}_i$ are functions of wRVU. In contrast, the Medicare allowed amounts that Matsumoto used to estimate the inverse sampling ratio, a_i and A_i, are functions of *total RVU.* The relationship among $\hat{\alpha}_c$, $wRVU_c$, and $TotalRVU_c$ vary across codes, at least code specialties. For example, the lowest wRVU-to-total-RVU ratio is 0.319 for radiology codes, and the highest is 0.620 for E/M codes.

Put differently, Matsumoto's estimator for the sampling ratio captures the sampling rate of a physician's *total RVU*, *not wRVU or service time.* For this to be close to the true sampling ratio, the services sampled in the 5% claims data need to be representative of the *distribution* of services across code groups *for each physician i.* This is not verifiable, nor is it likely to be true, especially for physicians with a small number of sampled services.

[2] See Table 5 in Matsumoto (forthcoming).

B.4 Sampling variation in the simulation exercise

Matsumoto did a simulation exercise to show he could be under-estimating service over-counting. We pointed out in the Reply that what he computed was actually the sample counterpart of something other than his claimed parameter of interest. Moreover, a related concern is how much the estimated hours *for the same physician* varied across the 500 subsamples. Without Matsumoto's claims data, we are unable to guage this sampling variation, especially how the variance across subsamples for the same physician would compare with the size of reduction in estimated hours after the service count adjustment. The larger the former, then more one should be concerned about potential noise in Matsumoto's result based on *one* 5% sample of claims.

References

Cutler, David, Jonathan Skinner, Ariel Dora Stern, and David Wennberg, "Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending," *Harvard Business School Working Paper*, 2015, *No. 15-090.*

Fang, Hanming and Qing Gong, "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked," *American Economic Review*, 2017, *107* (2), 562–591.

Matsumoto, Brett, "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Comment," *American Economic Review*, forthcoming.

Zuckerman, Stephen, Katie Merrell, Robert Berenson, Susan Mitchell, Divvy Upadhyay, and Rebecca Lewis, "Collecting Empirical Physician Time Data: Piloting and Approach for Validating Work Relative Value Units," *Urban Institute Research Report*, 2016.

_ , Robert Berenson, Katie Merrell, Tyler Oberlander, Nancy McCall, Rebecca Lewis, Sue Mitchell, and Madhu Shrestha, "Development of a Model for the Valuation of Work Relative Value Units: Objective Service Time Task Status Report," *Urban Institute Research Report*, 2014.