American Economic Review 2020, 110(12): 4004–4010
https://doi.org/10.1257/aer.20191970

Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Reply[†]

By Hanming Fang and Qing Gong[*]

Matsumoto (2020) pointed out data and coding errors in Fang and Gong (2017). We show that these errors have limited impacts: all qualitative findings remain after correcting them. Matsumoto also discussed potential service overcounting in the aggregated utilization data we used to illustrate our method, and then quantified the extent of overcounting with a sample of Medicare claims. We acknowledge the issue but discuss the noise and the bias in his quantification. Overall, our proposed method remains useful, as regulators who are interested in applying the method are unlikely to be subject to the data limitations. (*JEL* H51, I13, I18, J22, J44)

Fang and Gong (2017)—henceforth, FG—proposed using physicians' hours worked to detect potential overbilling in Medicare. The main insight was that all physicians were subject to the same time constraint, which circumvents problems such as unobserved patient heterogeneity that complicated the traditional auditing process. We argued that the main contribution of the paper was "a useful first step for effective and more targeted auditing," but *not* "definite evidence for fraudulent coding" or a "substitute for existing methods."[1]

With this caveat in place, we illustrated the method using the Medicare Part B Fee-for-Service (FFS) Utilization and Payment Data (henceforth, *the utilization data*). We first estimated the time per service using a subset of codes with known time requirements. We then multiplied these estimates with the number of times each service was claimed to have been rendered in a year to estimate the total hours worked by a physician. Notably, we chose to use the utilization data because (i) they provided as close to a full count of a physician's services as we could get, and (ii) they were publicly available and thus easy to use for illustration purposes. We did, however, acknowledge that the utilization data were quite coarse and had "well-noted limitations," that our time-per-service estimates were necessarily noisy, and that only Medicare Part B FFS services were accounted for (and subject to

[*] Fang: Department of Economics, University of Pennsylvania; School of Entrepreneurship and Management, ShanghaiTech University; and NBER (email: hanming.fang@econ.upenn.edu); Gong: Department of Economics, University of North Carolina at Chapel Hill (email: qinggong@email.unc.edu). Pinelopi Goldberg was the coeditor for this article. We are grateful to Brett Matsumoto for his comments on Fang and Gong (2017). All remaining errors are our own.

[†] Go to https://doi.org/10.1257/aer.20191970 to visit the article page for additional materials and author disclosure statements.

[1] Fang and Gong (2017, pp. 563 and 589).

truncation due to minimum reporting requirements).[2] The exact *level* of potential overbilling was neither a focus nor claimed to be a contribution of FG, as discussed multiple times throughout the paper.

Matsumoto (2020) raises two main issues. First, he pointed out the errors in the time-per-service imputation due to a data misinterpretation and typos in the Stata codes. We thank Matsumoto for spotting these inadvertent errors, which had limited impacts according to both his and our new calculations. We also show that the results of subsequent qualitative analyses in FG remain unchanged, including those on the characteristics and the coding patterns of flagged physicians.

Second, Matsumoto discussed how the utilization data could overcount services, mainly because they do not include *code modifiers* that inform the CMS of how a service is provided. He then tried to quantify the extent of overcounting with claims on a 5 percent sample of Medicare beneficiaries. We agree that using more granular information could improve the precision of service count estimates. But we are also concerned that using *one* 5 percent sample of claims will introduce noise and new biases, especially when the goal is to estimate the total hours of *each sampled physician*, rather than the distribution across physicians. We argue that, contrary to what Matsumoto suggested, our approach remains useful as it depends on the quality of data available to the regulators. Regulators are likely equipped with the *universe* of claims data and will not be bound by the data limitations of either FG or Matsumoto.

The remainder of this reply is organized as follows. In Section I, we address Matsumoto's comments on the estimated time per service. In Section II, we assess Matsumoto's quantification of the potential service overcounting in the utilization data we used and describe a major weakness of using the 5 percent claims data to infer each provider's total hours. In Section III, we conclude.

I. Time-per-Service Estimation: Similar Results after Correction

FG estimated the time needed to furnish each of the 4,480 services in the main sample in two steps: we first calculated the time per "work relative value unit" (work RVU, or wRVU[3]) for service codes with known time information ("timed codes"); we then estimated the time-per-service for other codes based on their wRVU, which was known for all codes.

The timed codes that we used to calculate the time-per-wRVU fell into two types. The first type is 145 codes whose definition includes a suggested or required service time; the second type is 112 codes that Zuckerman et al. (2014) selected to time on site. For the second type of timed codes, Matsumoto correctly pointed out that we misunderstood the values in column *Service Time* in Appendix B of Zuckerman et al. (2014). They were suggested service time by a committee of experts, instead of the objective time measures that Zuckerman et al. were planning to collect.

[2] See, for example, FG (p. 563), where we referred the readers to existing papers on such limitations, including O'Gara (2014), where several of the points in Matsumoto (2020) were mentioned.

[3] The wRVU is one of the three components of the total RVU of a HCPCS code. It quantifies the amount of work, primarily time, needed to furnish the service. The other two components are practice expense RVU and malpractice insurance RVU. The total RVU is the weighted sum of these three components, where the weights are geographical practice cost indices. The physician fee of a service is determined by multiplying its total RVU with a constant conversion factor.

We thank Matsumoto for pointing out this error; fortunately, it had a minor impact. The number of flagged physicians would *increase* slightly had we removed this error *alone*: 2,302 and 2,186 physicians were flagged at the 100-hours/week threshold in 2012 and 2013, respectively.[4]

Matsumoto also suggested other changes to FG's time-per-service estimation procedure, most of which we gratefully accept. As noted in Matsumoto (2020), these changes had limited impacts on the number of flagged physicians. Our new calculation confirms this: the number of flagged physicians decreased to 1,845 and 1,683 at the 100-hours/week threshold in 2012 and 2013, respectively, or a 20 percent reduction from the previous numbers of flagged physicians in FG.[5]

Furthermore, we re-did the subsequent analysis in FG and the qualitative findings remain unchanged, including those on the characteristics and the coding patterns of flagged physicians, the representativeness of specialties among flagged physicians, etc. We refer the readers to online Appendix Section A for details. In particular, we still find optometry, dermatology, and ophthalmology to be the most overrepresented among flagged physicians.[6] This result is consistent with not only the original findings of FG but also the Office of Inspector General (OIG) reports on the vulnerability of ophthalmology to inappropriate billing in 2012. With superior data and closer scrutiny, the OIG also found claims that were not allowed, not physically possible, or potentially abusing certain code modifiers for higher payments.[7]

II. Service Counts in the Medicare Utilization Data

A. *Potential Overcounting of Services in the Utilization Data*

The other main issue Matsumoto (2020) raised is that the utilization data could overcount the number of services. In his Section ID, Matsumoto described three sources of overcounting: multiple physicians providing the same service, synergy in providing services on multiple sites of the same patient, and the number of days *within a "bundle" service* coded as the number of bundles. All three can be identified with claims data, particularly the *code modifiers* that inform the payer (CMS) of details on how a service is provided.

These are valid concerns that complement FG's discussion of their data limitations. The absence of claim-level information does pose a threat to the precision of FG's hours estimation. But two notes are in order concerning its scope. First, regulators have all the claims and thus do not face this problem. Second, it is likely overshadowed by the other data limitations that made FG *underestimate* the weekly hours: only Medicare Part B FFS services were included, from which services

[4] FG originally flagged 2,292 and 2,120 physicians for the two years, respectively.

[5] Matsumoto (in his online Appendix Table 12) flagged 1,664 and 1,518, a roughly 27–28 percent reduction from the numbers in FG. Our new calculation differs from Matsumoto's because he generated time estimates for the *timed* codes, whereas we didn't. We don't find it a compelling choice to impute a time for the already timed codes, especially when the imputation is based on the time information on those timed codes.

[6] That is, these specialties have a "specialty flag index (SFI, defined in equation (2) of FG)" above 50.

[7] See paragraph 3 on page 3 and Table 3 on page 7 of Office of Inspector General (2015), as well as Table 2 on page 8 of Office of Inspector General (2014).

provided on 10 or fewer patients were excluded;[8] the data were also aggregated at the annual level, so FG only flagged physicians with long *average* weekly hours, assuming 51 *work* weeks per year.

Moreover, FG chose 100 hours/week as the threshold to flag physicians *in expectation of* noisy hours estimates. To put the number into perspective, the National Ambulatory Medical Care Survey (NAMCS) sampled physicians and recorded their work for a week; the *maximum* weekly hours on *Medicare patients* is 15.17 for physicians in solo practices (all of whose visits were sampled) and 49.82 for those in group practices (for whom the sampling rate is not reported). Even without the NAMCS numbers, it is not hard to see the abnormality in spending 100, or even 80, hours/week on Medicare Part B FFS patients only, for 51 out of 52 weeks during the year. So the 100 hours were never meant to be interpreted literally, as it already had built-in room for noise. Had one estimated the hours precisely, using the 100 hours threshold would simply be too lenient.

B. *Trade-Offs in Using Claims Data on* 5 *Percent Medicare Beneficiaries*

Matsumoto went on to quantify the extent of service overcounting in the utilization data. He adjusted service counts using the claims on a 5 percent sample of Medicare beneficiaries (henceforth, *the 5 percent claims data*) and re-estimated physician hours. We agree that adjusting the service counts with more granular data can be helpful. But we also want to point out the trade-offs in using the 5 percent claims data, which could result in unknown biases in Matsumoto's quantification.[9]

Advantages of the 5 Percent Claims Data.—The apparent advantage of the 5 percent claims data is the availability of code modifiers, which could help avoid overcounting in the ways Matsumoto pointed out. The other advantage is the availability of claim dates, which Matsumoto did not exploit. FG's utilization data were aggregated yearly, so they flagged physicians based on weekly hours *averaged* across a year, which was rather lenient. In fact, when discussing the potential of alternative data, FG (p. 589) wrote: "if higher-frequency data are available, [one could] flag physicians based on claims filed in a quarter, a month, or even a week, in which case there is less intertemporal smoothing than is permitted in our sample."

Drawbacks of the 5 Percent Claims Data.—The major drawback of the 5 percent claims data is it only includes a fraction of all the (Medicare Part B FFS) services of each sampled physician. Thus, one must estimate the total hours from the sampled hours. But the data are not a 5 percent random sample of each physician's claims. They include the claims for a 5 percent sample of Medicare beneficiaries. As a result, the sample sizes, the sampling rates, and the representativeness of the

[8] This minimum reporting requirement excluded 24 percent of all the services from FG's utilization data (measured in charge amounts). See Matsumoto (2020, Section IF).

[9] The 5 percent claims data had been used by many researchers, and was thus also a natural candidate when we sought data for illustrating the method in FG. But we decided against it after weighing the pros and cons (to be discussed below).

sampled services can all vary significantly across physicians.[10] Matsumoto needed to recover the 100 percent of service from a sample with an unknown sampling rate *for each physician*, some of whom only have a handful of claims sampled. This is a daunting task.

Matsumoto approached the problem in three steps. *For each physician i*, he first estimated the hours needed for the sampled claims as did FG $(\hat{t}_i)$; he then estimated the sampling rate associated with these claims $(\hat{r}_i)$; he finally estimated the hours needed for 100 percent of the physician's services $(\hat{T}_i = \hat{t}_i/\hat{r}_i)$. The sampling rate was inferred from charges for the sampled claims (a_i) and the total charges for the physician reported separately by CMS (A_i). That is,

$$\hat{T}_i = \hat{t}_i \times \frac{A_i}{a_i}. \tag{1}$$

We have two main caveats about this estimator of the total weekly hours, T_i^*. First, $\hat{T}_i$ is bound to be *biased* unless $\hat{t}_i$ and A_i/a_i are uncorrelated, even if A_i/a_i is an unbiased estimator of the "true" inverse sampling ratio.[11] Second, A_i/a_i cannot be shown to be unbiased, either. The "true" sampling rate is the ratio of T_i^* and $\hat{t}_i$, which are functions of wRVU; but a_i and A_i are functions of *total RVU*. The ratio of wRVU to total RVU varies greatly across groups of service codes. So the services in the 5 percent claims data need to be representative of the *distribution* of services across code groups *for each physician i*. This is neither verified for Matsumoto's case nor likely to be true for physicians with a small number of sampled services. We discuss the details in online Appendix Section B.

Implications on the Quantification of Potential Service Overcounting.—Matsumoto was aware of the above-noted issues and the potential bias in his quantification of service overcounting in the utilization data. He argued that the target of his estimation was not $\hat{T}_i$, but the *change* in the number of flagged physicians after he adjusted the service counts downward. Denote $\hat{T}_i^u$ and $\hat{T}_i^a$ as the estimated total weekly hours without and with his service count adjustments, respectively. Matsumoto's goal is then

$$\begin{aligned} E\big[\Delta(\#\text{flagged})\big] &= \sum_{i\in\mathcal{I}_s} \Pr\big(\hat{T}_i^u > 100 \geq \hat{T}_i^a\big) \\ &= \sum_{i\in\mathcal{I}_s} \Pr\left(\frac{\hat{t}_i^u}{a_i} > \frac{100}{A_i} \geq \frac{\hat{t}_i^a}{a_i}\right) \\ &= \sum_{i\in\mathcal{I}_s} E\left[\mathbf{1}\left(\frac{\hat{t}_i^u}{a_i} > \frac{100}{A_i} \geq \frac{\hat{t}_i^a}{a_i}\right)\right] \end{aligned} \tag{2}$$

[10] For example, in Matsumoto's 5 percent claims data, the number of claim lines sampled per physician ranges from 4 to 53,018 among the 2,120 physicians flagged by FG (median is 386). The *inferred* sampling rate in terms of dollar values of services ranges from 0.44 percent to 9.62 percent among these flagged physicians (median is 4.95 percent), whereas it ranges from 0.016 percent to 25.7 percent among the unflagged physicians (median is 4.9 percent). See his online Appendix Table 9.

[11] Matsumoto (2020) acknowledged the noise and the bias in his Section IF and online Appendix Section C.1.

where $\mathcal{I}_s$ is the set of sampled physicians, and $\hat{t}_i^u, \hat{t}_i^a, a_i$ are all random variables due to sampling. It is important to know that the very last expectation should be taken over *different draws of a subset of claims for the same physician i*.

With *one* draw of claims (i.e., the 5 percent claims data), what Matsumoto calculated was

$$(3) \qquad \sum_{i \in \mathcal{I}_5} \mathbf{1}(\text{loss of flag}_i) = \sum_{i \in \mathcal{I}_5} \mathbf{1}\left(\frac{\hat{t}_i^u}{a_i} > \frac{100}{A_i} \geq \frac{\hat{t}_i^a}{a_i}\right),$$

where $\mathcal{I}_5$ denotes the set of physicians covered in the 5 percent claims data. As such, Matsumoto's estimate of the changes in the number of flagged physicians due to service count adjustment using the 5 percent claims data is also noisy, and subject to sampling variations.

Matsumoto tried to corroborate the results with a simulation exercise. He drew 500 subsamples from the 5 percent claims data, each having 10 percent of the sampled beneficiaries and thus 0.5 percent of all claims on average. He then calculated the average change in the number of flagged physicians due to service count adjustments across the 500 subsamples. Comparing this average change with the "true" change, which he obtained from the hypothetical "population" (the 5 percent claims data), he found that the average change is smaller than the "true" change. Thus, he concluded that using a sample of claims would understate the impact of service overcounting. However, what he calculated in this exercise was

$$(4) \qquad \frac{1}{500} \sum_{k=1}^{500} \left[\sum_{i \in \mathcal{I}_{0.5}^k} \mathbf{1}\left(\frac{\hat{t}_i^u}{a_i} > \frac{100}{A_i} \geq \frac{\hat{t}_i^a}{a_i}\right)\right],$$

where $\mathcal{I}_{0.5}^k$ denotes the set of physicians covered in the kth 0.5 percent subsample ($k = 1, 2, \ldots, 500$). Note that the average subsample included 390,258 (or 64 percent) of the 608,050 physicians in Matsumoto's 5 percent claims data, implying that $\mathcal{I}_{0.5}^k$ varied drastically across k.[12] As a result, Matsumoto's numerical simulation confounds the impacts of both the sampling variation of *claims* for each physician and the sampling variation of *physicians*. The real object of interest, however, is (2), where the expectation should only be over different draws of claims *for the same physician*. What Matsumoto computed was really the sample counterpart of something other than his parameter of interest, thus it cannot inform us of whether his previous findings were overstating or understating the impact of service count adjustments.

III. Conclusion

In this reply, we address the two major issues raised by Matsumoto (2020). First, we thank Matsumoto for pointing out the inadvertent mistake in our use of the time values in Zuckerman et al. (2014) and the coding errors in our time-per-service estimation. We corrected these issues and our results remain qualitatively the same.

[12] See online Appendix Table 8 of Matsumoto (2020). The standard deviation is 540.8 physicians.

Second, we acknowledge that the lack of claim-level information in the utilization data could result in potentially overcounting services and that adjustments could improve the precision of our approach. We also discuss the noise and the bias in Matsumoto's quantification of the potential overcounting using a 5 percent sample of claims. Overall, we see the methodologies and the data used by FG and Matsumoto as complements: each has its strengths and weaknesses, and neither dominates the other. Regulators do not face the data limitations of either FG or Matsumoto (2020), though. They have the universe of more granular claims data and could still apply our approach, with the adjustments suggested by Matsumoto, as a first step to detecting potential overbilling.

REFERENCES

Fang, Hanming, and Qing Gong. 2017. "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked." *American Economic Review* 107 (2): 562–91.

Fang, Hanming, and Qing Gong. 2020. "Replication Data for: Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Reply." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E119192V1.

Matsumoto, Brett. 2020. "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Comment." *American Economic Review* 110 (12): 3991–4003.

Office of Inspector General. 2014. "Medicare Paid $22 Million in 2012 for Potentially Inappropriate Ophthalmology Claims." Washington, DC: Department of Health and Human Services.

Office of Inspector General. 2015. "Questionable Billing for Medicare Ophthalmology Services." Washington, DC: Department of Health and Human Services.

O'Gara, Patrick T. 2014. "Caution Advised: Medicare's Physician-Payment Data Release." *New England Journal of Medicine* 371 (2): 101–103.

Zuckerman, Stephen, Robert Berenson, Katie Merrell, Tyler Oberlander, Nancy McCall, Rebecca Lewis, Sue Mitchell, et al. 2014. *Development of a Model for the Valuation of Work Relative Value Units: Objective Service Time Task Status Report.* Washington, DC: Urban Institute, Social and Scientific Systems, Inc., and RTI International.