Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked: Comment

By Brett Matsumoto*

Fang and Gong (2017) develop a procedure to detect potential overbilling of Medicare by physicians. In their empirical analysis, they use aggregated claims data that can overstate the number of services performed due to features Medicare billing. In this comment, I show how auditors can use detailed claims level data to better target improper overbilling.

Fang and Gong (2017) propose a simple test for detecting potential overbilling in Medicare. They estimate the total amount of time physicians work for all of the claims submitted to Medicare and flag individuals whose submitted claims imply implausible hours worked (in excess of 100 hours per week). For the test to be most useful, it requires an accurate measure of time per service as well as an accurate count of services. The data the authors use overstate the number of services performed for some physicians and the service time estimates used by the authors are not as conservative as claimed in the paper due to an error in the source of procedure time data as well as coding errors in the implementation of the service time estimation procedure.

In this comment, I detail the source of the data issues and use other sources of data to attempt to correct for these issues. In the case of the service time data, the authors mistakenly claim that their data source contains objective time measures from an Urban Institute study that was not completed at the time the authors conducted their analysis. The authors also make a number of coding errors implementing the estimation of procedure times that leads to higher procedure times being used in the calculation of physician working time. The authors' source of data for the number of procedures is the Medicare Provider Utilization data for physicians. For some physicians and services, the service counts in the utilization file are inflated for reasons that I outline in this comment. Finally, for certain procedures it is difficult to know how much of the procedure was actually performed by the billing physician due to features of Medicare billing. Overall, these data limitations may limit the practical usefulness of the test.

I re-conduct the authors' analysis using the limited objective time measures from the final Urban Institute study report and construct service counts from

* Bureau of Labor Statistics (BLS), 2 Massachusettes Ave NE, Suite 3105, Washington, DC 20212, (email: matsumoto.brett@bls.gov) Additionally, I would like to thank the anonymous referees, Hanming Fang, and Quin Gong for helpful guidance in revising this comment. The views and opinions expressed in this paper are my own and do not reflect the views of the BLS or the Federal Government. The author declares that he has no relevant or material financial interests that relate to the research described in this paper.

detailed claims level data for a 5% random sample of Medicare beneficiaries using a procedure that attempts to correct for common sources of overcounting in the utilization data. Using this alternative measure of service counts and the new service time estimates, I estimate that the number of flagged physicians declines by over 50%. The evidence from the sample of claims suggests that alternative methods of counting the number of services performed can have a large impact on the estimated working time for certain physicians. The alternative procedure I propose may improve the usefulness of the authors' procedure by reducing the number and share of false positives among the flagged physicians.

In section 2, I describe the sources of data and the methods I use to address the issues with the estimates of procedure times and the possible overcounting of services. Section 3 presents the results, and Section 4 concludes.

I. Data and Methods

In this section, I detail the sources of data and the methods used to estimate total physician working time. I detail issues related to the generation of service counts by physician in the data source used in Fang and Gong (2017) and describe situations where this can yield potentially misleading counts of services for the purposes of estimating total physician working time. I describe how the additional billing details available in disaggregated claims data can be used to generate a measure of services performed that is potentially more meaningful for the purposes of estimating physician working times.

A. Medicare Reimbursement for Physician Services

Medicare reimburses for physician services based on a fixed fee for a given procedure. Procedures are defined based on Healthcare Common Procedure Coding System (HCPCS) codes. Level I HCPCS codes are numeric and correspond to the American Medical Association's (AMA) Current Procedural Terminology (CPT) codes.[1] Surgical CPT codes involve a bundle of services covering all closely related treatments for a set time period. This includes the surgical procedure as well as pre-operative preparation, post-operative care, and supplies. The length of the post-operative period depends upon the specific code and can cover up to 90 days after the procedure. The time it takes to perform a procedure is called the intra-service time. The total time for a CPT code is the intra-service time plus the pre-service time, the post-service time, and the time of any follow-up care.

Medicare reimburses physicians for each service based on a formula that is intended to capture the relative resource cost of performing the service (defined at the hcpcs code level).[2] One component of the formula is the work relative

[1] Level II HCPCS codes are alpha-numeric and cover supplies and services not covered under the CPT codes.

[2] The basic formula is Payment = (Work RVU + PE RVU + MP RVU) $*CF$. It's the sum of the work

value unit (wRVU), which captures the time (and intensity) of the physician's work. Adjustments can be made to the payment based on facility type, partial procedures, multiple procedures, as well as additional factors that can be procedure specific.

B. Estimating Physician Hours Worked

Fang and Gong (2017) use two sources of objective time measures that cover a subset of procedures ("timed codes") and use the average time per wRVU to estimate the time needed for the other procedures ("untimed codes").[3] The first source of data for timed codes includes codes that include a time requirement in the description of the service. These codes primarily cover evaluation and management (E/M) procedures. The second source of timed codes are presented as being objective time measures from an "on-site survey (Zuckerman, et al. 2014) that objectively measured the time needed for a subset of services."[4] The Urban Institute was contracted to conduct the survey by the Centers for Medicare and Medicaid (CMS). This second set of timed codes is the only source of objective time measures covering surgical procedures.

The authors make a number of coding errors in estimating the time required for the untimed codes. These errors make the time estimates used less conservative than what is described in the paper and are detailed in the Online Appendix A.[5] In addition, the authors use an incorrect source of data for the second set of timed codes. Zuckerman et al. (2014) is the interim report for the Urban Institute study and no data from the study were available at that time.[6]

The Urban Institute study has been completed and they issued a final report in December 2016 (Zuckerman et al., 2016). The effort to collect data turned out to be much more difficult than they originally expected. A major issue was a general unwillingness among providers to participate. In the end, they were only able to collect data from 3 sites. They also found that the CPT service descriptions tended to be out of date and included tasks in the pre- and post-service periods

relative value unit (RVU), the practice expense (PE) RVU, and malpractice insurance (MP) RVU. Each RVU is adjusted by a geographical factor to control for differences in cost across areas. The total RVU for the procedure is multiplied by a conversion factor (CF) to obtain a dollar payment.

[3]See Online Appendix A or Fang and Gong (2017) for a description of the estimation method.

[4]Fang and Gong (2017) page 565.

[5]The work RVUs are set based on the time it takes the typical physician, so this method recovers the average time for a procedure. There can be a lot of heterogeneity in times across physicians for the same procedure. It is important for the procedure to use conservative procedure time estimates in order to decrease the likelihood of falsely flagging doctors for being efficient.

[6]The data the authors use come from the table in Appendix B of Zuckerman et al. (2014), which lists the hcpcs codes the study intends to collect objective time measures for as well as the current non-objectively measured time values for those codes. When setting the work RVU for a procedure, CMS relies on recommendations from the American Medical Association's Relative Value Scale Update Committee (AMA RUC) which provides service time estimates. These estimates are obtained through a survey of physicians. Many stakeholders believe the time estimates are generally overstated as physicians have an incentive to report higher times to justify a higher RVU (e.g., Government Accountability Office (2015) and Crosson (2017)). The Urban Institute study was commissioned to determine the feasibility of replacing the survey time values with objectively measured times for setting work RVUs.

that are commonly performed by nurses or physician assistants. Since they were unable to separate the amount of time spent by the physician on the pre- and post-service tasks, they limited the data collection to the intra-service period. They were able to collect intra-service time data for 60 hcpcs codes out of the originally planned 112.

I perform the analysis using two sets of procedure time estimates, which I denote as "original" and "new" procedure times. The original procedure times are the procedure time estimates used in Fang and Gong (2017). The new procedure times correct the coding errors and use the objectively measured times from Zuckerman et al. (2016).[7] Since the Urban institute study only collected data on intra-service time, I use the American Medical Association's Relative Value Scale Update Committee (AMA RUC) time values for pre- and post- service time when calculating total procedure time for these codes.

C. Claims Data

The service counts the authors use come from the Medicare Part B FFS Physician Utilization and Payment data (Utilization data). These data are aggregated up to the level of provider, procedure, and provider setting level from the physicians claims data (i.e., the Carrier Claims file, which contains the claims for non-institutional Medicare providers). Any provider / procedure / setting combination with fewer than 11 beneficiaries in a year is omitted due to privacy concerns. The minimum reporting threshold for a procedure causes the Utilization data to represent less than 100% of claims. In addition to the service level data, the Utilization data contains a summary table with the total payments by provider for 100% of claims (with no censoring). Among physicians in the analysis sample, the Utilization data on average represent a little over 75% of total claims (as measured by Medicare payments).

The claims data I use is the 5% Carrier Standard Analytic File Limited Data Set.[8] The data include all claims for a 5% random sample of Medicare beneficiaries. The data contain a claims file with information on the total claim and a line file which lists the individual line charges for each claim. Each line item is associated with an hcpcs code. The Physician Utilization and Payment data are constructed from the line file for 100% of claims. The claims level data contain more detailed billing information than the Utilization data, which I use to generate an adjusted service count measure. Of the 623,959 physicians in the estimation sample in Fang and Gong (2017) from the Utilization data, 97.45% (608,050) are present in the 5% beneficiary sample.[9] The analysis sample is restricted to physicians who

[7]See Online Appendix H for the list of procedures included in Zuckerman et al. (2016), the times collected, and how they compare to the AMA RUC times.

[8]The Limited Data Sets remove beneficiary identifiable information and use a generated id variable to permit linking across services and years. Researchers can apply for use of these data sets at https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/Data-Disclosures-Data-Agreements/DUA_-_NewLDS.html

[9]I follow Fang and Gong (2017) and refer to individual providers as physicians even though some

appear in both the Utilization and claims data.

D. Sources of Service Count Bias

There are at least 3 potentially meaningful sources of service count bias which arise due to features of the Medicare billing process and generally lead to over-counting of services. First, multiple physicians can be involved in a procedure and all submit claims with a service count of 1 for the same hcpcs code. The service counts in the Utilization data assign the full treatment time to each physician involved in the treatment even though each physician is only performing part of the overall treatment. When a physician performs less than the full procedure, hcpcs modifier codes are used to indicate the specific responsibility of the billing physician, and the physicians are paid according to their specific contribution.

The second source of service count bias is that the "service count" variable in the utilization data does not always mean the number of times a particular treatment is performed.[10] For cataracts surgery (hcpcs code 66984), it is relatively common for one physician to perform the surgery (including day of surgery post service care) and another to handle the follow up care. The hcpcs code corresponding to cataract surgery has a 90 day global period for treatment.[11] Some doctors report the number of days the physician is responsible for follow-up care during the global period as the service count variable.[12] The authors' methodology treats the physician who only provides follow-up care for 90 days (approximately 1.7 work RVUs in 2013) as if that physician performed 90 full cataract surgeries (766.8 work RVUs).

The third main source of service count bias is that multiple procedures can be performed at the same time for less overall time than performing the procedures at separate visits as many of the pre- and post-service tasks are common to related procedures. Bilateral procedures are when a given procedure is performed on both sides of the body and can be thought of as a special case of multiple procedures.

These sources of service count bias are all identifiable in the claims data through the use of modifier codes. Table 1 compares the percentage of claim lines in the 5% beneficiary sample that contain one of these modifier codes for the physicians flagged at the 100 hour threshold in 2013 in the original paper, the unflagged physicians, and the unflagged physicians who have an estimated work time of

are non-physician practitioners (e.g., physical therapist, nurse, optometrist, or psychologist) who bill Medicare under their own National Provider Number (NPI).

[10]The documentation for the Utilization data cautions that the "service count" variable can mean different things for different procedures (Centers for Medicare and Medicaid Services, 2017).

[11]There is a fixed payment for follow-up care based on a percentage of the total work RVU, and if multiple doctors are responsible for care for different parts of the 90 day period then they are paid based on the fraction of days that they are responsible for care.

[12]The Medicare billing guidelines require that physicians that bill for follow-up care provide the range of dates that they are responsible for care in a separate field from the "service count" field. Some physicians who provide the follow-up care report a value of 1 in the service count field instead of the number of days. Coding the service count as days for follow-up care appears to be primarily driven by the preference of the individual claims processors (the Carriers or the private companies that are awarded the contract to process Medicare claims in a given area).

more than 20 hours per week. The sample is restricted to physicians who appear in both the 5% claims and the utilization data. Only codes that are eligible for a payment adjustment are counted (e.g., codes that are not eligible for the bilateral adjustment are not counted even if the claim includes the bilateral modifier code). Flagged physicians are more likely to submit claims indicating partial service, multiple procedures, and bilateral procedures. They also have a higher average service count per claim line.

Table 1—Characteristics of Claims for Flagged and Unflagged Physicians

Variable	Flagged	Unflagged	Unflagged & > 20 Hrs/week
Claim Lines	894,179	38,446,686	15,966,363
Multiple Procedure	5.65%	0.95%	1.23%
Bilateral Procedure	1.59%	0.37%	0.52%
Post-Service Care Only	0.36%	0.04%	0.03%
Intra-Service Care Only	0.53%	0.04%	0.09%
Technical Component Only	2.82%	0.75%	0.91%
Service Count	1.35 (3.53)	1.09 (1.10)	1.09 (0.98)
Major Surgical Procedure Claim Lines	98,162	1,071,755	544,812
Major Surgical Procedure Service Count	2.24 (9.95)	1.08 (2.51)	1.09 (2.53)

Note: Percent of claim lines with multiple procedures, bilateral procedures, post-service care only, intra-service care only, and technical component only based on hcpcs modifier codes. Other modifier codes that indicate partial size treatment (pre-service only or discontinued procedure) are rare (occur in less than 0.01% of claim lines). Mean service count with standard deviation in parentheses. Flagged physicians are those who are estimated to work more than 100 hours per week in Fang and Gong (2017). The sample includes HCPCS codes with a non-zero time in the original procedure time estimates for physicians who are in both the Utilization and the claims data. Major surgical procedures are defined as those with a 10 or 90 day global period.

The higher average service count among flagged physicians would not be a problem for the authors if the reported values are a count of the number of times the complete procedure is performed. Many procedures have an assumed quantity of 1 and claims that indicate a different service count will be processed and paid as if they have a quantity of 1. Surgical procedures that have a 10 or 90 day global period can not have a service quantity greater than 1.[13] Among the major surgical procedure codes, the flagged physicians have a significantly higher average service count.

For physicians billing for post-service care only for follow-up care to a surgical procedure, a service count greater than 1 indicates the number of days of post-service care rather than the number of times the entire procedure is performed. Three procedures are responsible for nearly all of the claims with a post-service care only billing modifier. The 3 hcpcs codes are all eye procedures and cover approximately 99% of claim lines that are reported with the post-service care

[13]Quantity for global surgical procedures is captured through the use of add-on codes. The add-on codes can have a quantity greater than 1, but the main procedure code has a quantity of 1. If the same procedure or closely related procedure needs to be performed during the global period for medical reasons, it would need to be submitted as a separate claim with the appropriate modifier.

only modifier. Table 2 lists the procedures and the average service counts for flagged and un-flagged physicians who work more than 20 hours per week. Also listed is the probability that the physicians report a service count of 1.[14] Flagged physicians have a higher average service count for these codes and less likely to report a service count of 1. The average Medicare payment and average allowed charge do not indicate that the higher service counts for the flagged physicians are associated with a much higher or more intensive level of service.

Table 2—Average Service Counts and Charges for Select Procedures, Post-Service Care Only

	hcpcs code					
	66984		66821		66982	
Variable	Not Flagged & > 20 hrs/week	Flagged	Not Flagged & > 20 hrs/week	Flagged	Not Flagged & > 20 hrs/week	Flagged
Claims	4507	2846	294	147	277	180
Service count	9.87 (24.72)	39.22 (40.65)	10.08 (25.19)	53.77 (40.10)	7.36 (21.25)	23.77 (35.18)
Proportion with 1 service count	0.841	0.411	0.857	0.245	0.874	0.572
Medicare Payment	50.65 (43.11)	55.93 (67.44)	33.60 (20.53)	37.15 (31.18)	49.82 (51.87)	42.73 (52.09)
Allowed Charge	64.67 (54.79)	71.37 (86.07)	42.66 (26.06)	48.17 (39.41)	63.26 (65.80)	54.29 (66.21)

Note: Medicare Payment denotes the total reimbursed by Medicare for the claim. HCPCS code 66984 is cataract removal with insertion of lens, code 66821 is YAG laser capusulotomy surgery, and code 66982 is complex cataract surgery. The unflagged physicians are restricted to those who work greater than 20 hours per week in Fang and Gong (2017) and flagged physicians are those who work more than 100 hours per week.

E. Method for Adjusting Service Counts

I use the detailed claims data to construct a service count variable that takes into account the appropriate units of the line service counts, doctors performing multiple procedures (including bilateral procedures), and doctors performing less than the full service. To correct the line service count variable, I set the line service count to 1 for all 10 day and 90 day global procedures. These surgical procedures by definition cannot be performed multiple times on the same day.[15] The quantity component of major surgical procedures is captured by add-on CPT codes (e.g., the first lesion treated is part of the main CPT code and a separate code is used to indicate additional lesions treated). Also, certain procedure codes

[14] A service count of one could mean that the physician provided one day of post-service care, or the physician only provides the range of dates in the separate field rather than a count of days in the service count field (the date range is required regardless of Carrier, but that field is not represented in the claims data).

[15] The exception is if it is performed as a bilateral procedure, which has it's own adjustment.

include the quantity in the CPT definition and are automatically billed as a quantity of 1 even though some physicians will report a different quantity.[16] This adjustment is sufficient to fully correct for this source of service count bias for these codes as long as they are not also affected by the remaining adjustments.

In other situations of service count bias the correction is less clear. In these other cases, Medicare adjusts the payment to reflect that less than the full procedure is performed. I make the assumption that there is a one-to-one relationship between the adjusted service count and the payment adjustment. Without objective time measures of the actual procedure times for these situations, there is no way to test this assumption, and it may be an over or an under correction.

When multiple procedures are performed, the main procedure is reimbursed at the full rate and additional procedures are reimbursed at a rate of 50%. So for these additional procedures, I reduce the quantity by half. I set the quantity for eligible bilateral procedures at 1.5.[17] For procedures with a technical and professional component, the physician fee schedule attributes the entire work RVU to the professional component, so I replace the technical component only with a zero quantity. For physicians that perform pre-, intra-, or post-service care only, a time is assigned based on a percentage of the total time. The percentage breakdown for the different components of a full procedure is in the physician fee schedule.[18] Then, the adjusted quantities are aggregated to the level of provider and hcpcs code (and modifier where appropriate) and used to calculate total time. Additional details of the service count adjustment procedure are provided in Online Appendix B.

F. Estimating Total Work Time from the 5% Beneficiary Claims Sample

The major limitation of the claims data is that it only includes a subset of each physician's total Medicare claims. In order to estimate the effect that the service count adjustment has on the propensity of physicians to be flagged, the hours estimate from the 5% sample needs to be adjusted to estimate total working hours. The Utilization data summary table includes the total allowed charges for each physician based on 100% of claims. I calculate the total allowed charges for each physician based on the claims in the 5% sample and divide it by the total allowed charges for that physician. The inverse of this ratio is multiplied by

[16]This list is not exhaustive, but the non-surgical codes where physicians will occasionally report service counts greater than 1 when a quantity greater than 1 is not possible include: 90960 dialysis 4 or more visits per month (some claims have a service count of 30 or 31 for the days in the month), 99232/99233 subsequent inpatient hospital visits, and 99214 office visit. The hospital and office visits are per diem codes and can only be billed once per day. It is rare for service counts greater than 1 to be reported for these services, however, it can have a large impact on the hours estimates for an individual physician.

[17]Some codes are defined as bilateral procedures so are not eligible for an adjustment. I set the quantity to 1 for these procedures. Also, for some codes, the bilateral procedure is treated as two separate procedures with no adjustment. The quantities are set to 2 for these codes.

[18]The physician fee schedule also indicates which codes are eligible for multiple procedure and bilateral adjustments. Adjustments are only made if the fee schedule indicates that the code is eligible for an adjustment.

the time estimates from the 5% sample to get an estimate of the total physician working time. Let t_5 and a_5 denote the total time and allowed charges in the 5% sample and A denote the 100% allowed charges. Then, the total time estimate is $T^* = t_5 * \frac{A}{a_5}$.

On average, T^* is greater than the total hours estimate based on the Utilization data.[19] The total charges from the Utilization summary table represent 100% of allowed charges for the physicians. The procedure level Utilization data that is used to estimate total working time in Fang and Gong (2017) represents about 76% of total charges due to the minimum reporting threshold. Also, T^* will not be an unbiased estimate of the 100% claims working time (which is unobserved) because t_5 and a_5 are not perfectly correlated. The total time estimate using this procedure has a higher variance because of the sampling noise than the total time calculated from the procedure level utilization data. Simulation evidence (see Online Appendix C.3) suggests that the sampling variability that results from the inverse allowed charge ratio transformation increases the number of physicians flagged regardless of service count measure used compared to the true total time. The physicians most susceptible to the service count adjustment tend to be more dispersed in the total hours distribution calculated from a sample, which makes it less likely that the service count adjustment will move them below a given threshold. Therefore, the estimated decrease in the number of flagged physicians from a sample of claims will tend to understate the true magnitude of the decrease if the service count adjustment was performed on 100% of claims in both absolute and percentage terms.[20] The estimated decrease in flagging from the service count adjustment is main parameter of interest from the estimate of total time using the 5% beneficiary sample.

II. Results

The results are presented using the two methods for generating service counts (unadjusted and adjusted) and the two methods of estimating procedure times (original and new). I calculate the unadjusted service counts in the claims data following the procedure used in the Utilization data by summing the line service count variable by doctor and procedure. The original procedure time estimates are the same as those in Fang and Gong (2017) and the new procedure time estimates correct errors in the estimation code and the source of data for the objective time measures. The first set of results limits the analysis to the 5% sample to examine the magnitude of the service count adjustment on the estimated hours worked. Then, I estimate the total hours worked from the 5% beneficiary sample using the allowed charge ratio transformation to estimate the impact of both the service count adjustment and the new procedure time estimates on the number of flagged

[19] See Online Appendix C for a comparison of the total time estimate distributions for the claims and Utilization data.

[20] The percentage decline is understated to a larger degree due to the higher initial number of flagged physicians.

physicians.

A. *The Effect of the Service Count Adjustment in the 5% Beneficiary Sample*

Table 3—Total Hours Worked Per Week Summary Statistics in the 5% Beneficiary Sample

Hours Distribution	Obs	Mean	SD	Median	Min	Max
Unadjusted Service Count	608050	0.572	0.922	0.299	0	55.14
With Service Count Adjustment	608050	0.535	0.738	0.290	0	42.64
				Centiles:		
	75	90	95	99	99.5	99.9
Unadjusted Service Count	0.732	1.384	1.959	3.780	4.854	9.054
With Service Count Adjustment	0.705	1.324	1.860	3.407	4.226	6.519

Note: Time estimates based on the 5% beneficiary sample. The "unadjusted service count" method calculates service quantities by summing the service count field on the claim line, which is the method used to calculate quantities in the Utilization data. The"with service count adjustment" method adjusts for common sources of overcounting in the service count field. The sample includes physicians who are in both the Utilization and claims data.

In this section, I estimate the effect of potential service count bias on the total hours worked based on the claims in the 5% beneficiary sample. To isolate the effect of the service count adjustment, I use the original service time estimates with both the unadjusted and adjusted service counts.[21] Table 3 presents the summary statistics for the total hours worked using the unadjusted service counts and with the service count adjustments described above. The mean total hours worked per week in the 5% sample decreases from 0.572 hours to 0.534 hours, but there is minimal change in the median hours worked. When flagging physicians for overworking, the relevant part of the distribution is the right tail. The bottom panel presents the centiles in the upper tail of the distribution. At higher centiles, there is a larger difference in the hours worked between the two distributions in both absolute and percentage terms.

For an individual physician, the estimate of the service count bias is defined as the difference between the total hours worked using the unadjusted service counts and with the service count adjustment. Table 4 compares the average estimated service count bias in the 5% beneficiary sample of flagged, all unflagged, and unflagged physicians who work more than 20 hours per week in the Utilization data (using the original service time estimates). The service count adjustment leads to an average decrease of approximately 3.3 hours per week among flagged physicians. This corresponds to a decrease of approximately 26.8%. Unflagged physicians and unflagged physicians who are estimated to work greater than 20 hours per week in the utilization data have a smaller decrease (in both absolute

[21]The results are similar using the new time estimates.

Table 4—Average Estimated Service Count Bias in the 5% Beneficiary Sample

Variable	Flagged Physicians	Unflagged Physicians	Unflagged and > 20 Hrs/week
Number of Physicians	2120	605930	80972
Hours Bias	3.288 (6.567)	0.025 (0.244)	0.096 (0.498)
Percent Hours Bias	26.78 (35.74)	2.558 (7.990)	3.860 (9.205)

Note: Standard deviations in parentheses. Flagged and unflagged are based on time estimates using the original data and methodology. For each physician total time per week is calculated using both unadjusted and adjusted service counts using the original service time estimates based on the 5% beneficiary sample. Service count bias is defined as total time using the unadjusted service counts minus total time using the adjusted service counts. Percent hours bias is calculated by dividing the service time bias by the unadjusted total time. The groups of physicians (flagged, unflagged, and unflagged and > 20 hours per week) are defined based on the hours estimates in Fang and Gong (2017). Flagged physicians are those who were estimated to work more than 100 hours per week.

and percentage terms) in total hours worked per week as a result of using adjusted service counts.

B. The Effect of the Service Count Adjustment on Total Physician Time

Total time from the 5% sample is multiplied by the inverse of the share of allowed charges in the 5% sample to get an estimate of total time for 100% of claims. There are 4 sets of total time estimates for unadjusted and adjusted service counts with both the original and new service time estimates.[22] Table 5 shows the number of physicians flagged using the different methods. The top panel includes the full sample of physicians who appear in both the claims and utilization data, while the bottom panel only includes the physicians that were flagged in the original paper (greater than 100 hours per week based on the Utilization data with the original service time estimates).

Looking at the unadjusted results for all physicians (top panel), the number of physicians flagged at each threshold is higher than what was found using the Utilization data, which is not surprising given the censoring of the Utilization data and the higher variance of the claims data estimate from the sampling and transformation. Of the physicians that were flagged in the Utilization data (bottom panel), approximately 88.3% (1871/2120) are flagged at the same threshold (100 hrs) when using the original time estimates and approximately 88.5% (1343/1518) are flagged at the same threshold using the new time estimates. The overlap is not perfect because some physicians are not well represented in the 5% sample.[23] The

[22]In Online Appendix E, I present the results of using the different service time estimates on the number of physicians flagged in the Utilization data. The effect of going from the original to the new time estimates is similar in the Utilization data and claims data.

[23]Of the 249 physicians that are flagged in the Utilization data but not the claims data, 199 bill one of the 3 common post-service care procedures (see Table 2) in the Utilization data. Of the 50 that do not bill one of these 3 codes in the Utilization data, 40 would be flagged in the claims data at the 90 hour threshold. The 199 doctors who perform the 3 procedures have an average service count of 2738 in the Utilization data on an average of 70 unique beneficiaries. Only 5 of the 199 doctors are flagged at the 100 hour threshold after recalculating hours worked in the Utilization data assuming a quantity of 1

sampling variability will also cause some physicians who would be flagged based on 100% of claims to not be flagged (false negative), although the sampling and transformation is more likely to produce the opposite outcome (false positive). The service count adjustment leads to an approximately 49% reduction in the number of flagged physicians at the 100 hour per week threshold using the new time estimates (approximately 42% using the original time estimates). As discussed above, this is likely to be a lower bound of the reduction of flagged physicians that would occur if the service count adjustment was performed on the basis of 100% of claims.

Table 5—Number of Flagged Physicians

	Original Time Estimates		New Time Estimates	
	Unadjusted Quantities	Adjusted Quantities	Unadjusted Quantities	Adjusted Quantities
All physicians:				
# Hrs per Week > 20	102,027	96,163	80,031	75,877
# Hrs per Week > 80	4,697	3,220	3,332	2,098
# Hrs per Week > 100	2,614	1,514	1,903	975
# Hrs per Week > 112	1,962	986	1,432	653
# Hrs per Week > 168	748	219	557	150
Total # of physicians	608,050		608,050	
Flagged Physicians in the Utilization Data:				
# Hrs per Week > 20	1,995	1,629	1,437	1,128
# Hrs per Week > 80	1,957	1,502	1,407	1,041
# Hrs per Week > 100	1,871	1,224	1,343	841
# Hrs per Week > 112	1,526	872	1,101	599
# Hrs per Week > 168	603	209	444	142
Total # of physicians	2,120		1,518	

Note: Adjusted estimates calculate service counts from the claims data taking into account sources of overcounting. The top panel includes all physicians that appear in both the claims and Utilization data (data source for Fang and Gong (2017)). The bottom panel only includes the set of flagged physicians flagged in the Utilization data at the 100 hour threshold using the same procedure time estimates.

The cumulative effect of using both the new time estimates and the adjusted service counts can be found by comparing column 1 (original time estimates and no service count adjustment) to column 4 (new time estimates with adjusted service counts). At the 100 hour per week threshold, the number of flagged physicians declines by about 63% among all physicians and 55% among physicians flagged in Fang and Gong (2017) (based on the Utilization data). The remaining question is whether the remaining flagged physicians represent potential overbilling. In Online Appendix F, I describe exceptions to the authors' assumption that the billing physician is the person performing the entire service. These exceptions can be procedure specific or specific to a particular setting (billing in a group

per patient per day for these 3 procedures. The impact of these common post-service care only codes on the estimated hours worked in the claims data is examined in Online Appendix C.2.

practice). Among the commonly billed codes of the remaining flagged physicians, many are codes for which it is not possible to determine how much of the total time of the procedure was performed by the billing physician.

III. Conclusion

The authors propose a simple and transparent methodology to flag cases of potential Medicare overbilling by calculating the total hours worked implied by the submitted claims. The authors used an incorrect source of data for the Urban Institute study times and made a number of coding errors in implementing their procedure for estimating procedure times which causes their time estimates to be less conservative than suggested in the paper. Even when the correct data source is used, there is no source of objectively measured *total* times for surgical procedures. The Urban Institute study only objectively measured intra-service time since pre- and post-service care for surgical procedures are commonly performed by staff. Zuckerman et al. (2016) concludes the report by stating that, although feasible, an effort to collect total procedure times for a larger set of services would have to overcome many existing challenges and would require a substantial investment.

A larger issue is that the service count measure in the Utilization data tends to overstate the number of services performed for the purpose of calculating total physician working time. I propose a method for generating a measure of the quantity of services performed that adjusts the service counts in an attempt to correct for this issue. I estimate that using the adjusted service counts would decrease the number of flagged physicians by over 40%. I estimate that addressing both the service count and estimated service time issues would decrease the number of flagged physicians by over 50%, although the exact magnitude is imprecise due to the limitation of estimating total working time from a sample of claims. Finally, among the remaining flagged physicians, the most common and time intensive billed codes include many where it is impossible to know whether the billing physician provided the entire service. This final point, combined with the lack of objective time measures of total time for surgical procedures, limits the usefulness of this procedure as a general method to screen for potential overbilling. A modified version incorporating the adjustments outlined in this comment combined with the complete claims data could be useful to regulators, particularly in specialties where objective time measures for total procedure time are available and where the billed codes have minimal exceptions to the general requirement that the billing physician perform the entire procedure.

REFERENCES

Centers for Medicare and Medicaid Services. 2017. "Medicare Fee-For-Service Provider Utilization and Payment Data Physician and Other Supplier Public Use File: A Methodological Overview."

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf.

Crosson, Francis. 2017. "MedPAC comment on CMS's proposed rule on the physician fee schedule and other revisions to Part B for CY 2018." http://www.medpac.gov/docs/default-source/comment-letters/09082017_partb_2018_medpac_comment_sec.pdf?sfvrsn=0.

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

Government Accountability Office. 2015. *Medicare Physician Payment Rates: Better Data and Greater Transparency Could Improve Accuracy.*

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

Zuckerman, Stephen, Katie Merrell, Robert Berenson, Sue Mitchell, Divvy Upadhyay, and Rebecca Lewis. 2016. "Collecting Emprircal Physician Time Data: Piloting an Approach for Validating Work Relative Value Units." Urban Institute.