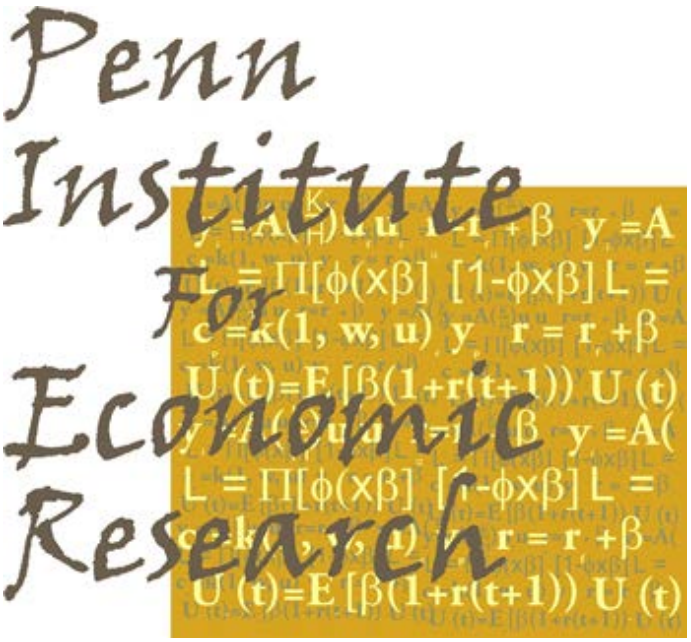


Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 16-006

"Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked"

By

Hanming Fang and Qing Gong

http://ssrn.com/abstract=2747152

Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked*

Hanming Fang† Qing Gong‡

March 7, 2016

Abstract

Medicare overbilling refers to the phenomenon that providers report more and/or higher-intensity service codes than actually delivered to receive higher Medicare reimbursement. We propose a novel and easy-to-implement approach to detect potential overbilling based on the hours worked implied by the service codes physicians submit to Medicare. Using the Medicare Part B Fee-for-Service (FFS) Physician Utilization and Payment Data in 2012 and 2013 released by the Centers for Medicare and Medicaid Services (CMS), we first construct estimates for physicians' hours spent on Medicare Part B FFS beneficiaries. Despite our deliberately conservative estimation procedure, we find that about 2,300 physicians, or 3% of those with a significant fraction of Medicare Part B FFS services, have billed Medicare over 100 hours per week. We consider this implausibly long hours. As a benchmark, the *maximum hours* spent on Medicare patients by physicians in National Ambulatory Medical Care Survey data are 50 hours in a week. Interestingly, we also find suggestive evidence that the coding patterns of the flagged physicians seem to be responsive to financial incentives: within code clusters with different levels of service intensity, they tend to submit more higher intensity service codes than unflagged physicians; moreover, they are more likely to do so if the marginal revenue gain from submitting mid- or high-intensity codes is relatively high.

Keywords: Medicare, Overbilling, Hours worked

JEL Classification Codes: H51, I13, I18

---

*We are grateful to Alex Li for helpful comments and suggestions. Fang gratefully acknowledges the generous financial support from NSF Grant SES-1122902. All remaining errors are our own.

†Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, and the NBER. E-mail: hanming.fang@econ.upenn.edu.

‡Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. E-mail: qinggong@sas.upenn.edu.

1 Introduction

Medicare benefit payments in 2014 totaled $597 billion, accounting for 14% of the United States federal budget (see Henry J. Kaiser Family Foundation (2015)); and Congressional Budget Office (2014) forecasts, as more baby boomers retire, government health care program expenditures will further increase and will account for 14% of U.S. GDP by 2039. From a public policy perspective, mitigating the inefficiencies in the Medicare system to ensure that every Medicare dollar is put to the best use is of first order importance. This paper is about a particular form of inefficiency that is broadly referred to as *overbilling,* where providers file improper claims in order to increase the reimbursement from Medicare or other insurance companies. The U.S. Department of Health and Human Services Office of Inspector General loosely defines two common types of overbilling (formally referred to as "improper claims"): *upcoding* refers to billing codes reflecting a more severe illness than actually existed or a more expensive treatment than was provided; *overcharging* refers to charging for more units of a service than was provided, or charging for services not provided at all (see Department of Health and Human Services (2015)).[1] Lorence and Spink (2002) estimated that overbilling costs the Federal government about $12 billion annually in the 1990s; and researchers continued to find evidence of overbilling ever since (e.g., Brunt (2011)).

Efficient and cost-effective detection of overbilling, preferably at the individual *provider* level, is crucial to reduce overbilling. However, this remains a challenging task. Most papers in the literature measure "overbilling" by the differential probability that higher-level codes are billed relative to lower level codes, or by the percentile of a provider's total reimbursement received in the distribution. But such measures could be confounded by factors such as selection on patient and provider characteristics. Medicare claims data, available in more recent years, enabled researchers to control for some, but not all, patient and provider heterogeneities. Rosenberg et al. (2000) developed a Bayesian model to adaptively detect questionable claims using previous hospital claims that insurers already selected for audit. However, new costly audits are required to apply the methods to any new claims data. The Comprehensive Error Rate Testing (CERT) program by the Centers for Medicare and Medicaid Services (CMS) faces a similar challenge, because the program needs to hire experts to review a large sample of claims every year (see Centers for Medicare and Medicaid Services (2015)). Geruso and Layton (2015) identified upcoding at the *market* level using risk scores and variations in financial incentives of physicians.

In this paper, we propose a novel approach to efficiently detect, or at least flag, potential

[1]There is a third type of improper claims that bill for services that lack medical necessity, sometimes known as *utilization abuse*. Detecting utilization abuse could be much harder and potentially controversial, so the approach we propose in this paper only targets detecting upcoding and overcharging.

Medicare overbilling using conservative estimates of the *hours worked* implied by service codes providers submit to Medicare for reimbursement. Our idea is very simple. Every provider has a *fixed* number of hours in any given period; and most of the service codes that are submitted for reimbursement require that the provider spends certain amount of time with the patient. If the hours worked implied from the service codes a provider submits to CMS are implausibly long, the provider is suspicious for overbilling. Our approach to flag potential Medicare overbilling has several key advantages. First, the existing physician-level billing data are sufficient to implement this approach, with no additional data collection needed. Second, by focusing on the implied hours worked within a given time period, our approach separates confounding factors such as selection on patient conditions. Third, our approach is flexible in the sense that it can be automated, and can be easily extended to a more general setting with augmented data, for example, by including other components of Medicare and/or more physician billing information such as billings for beneficiaries of other insurance programs. We should also note that our calculation of implied physician hours worked is deliberately conservative for the moment, and it is certainly not fail-proof especially given some well-noted data limitations (see, for example, O'Gara (2014); Jones et al. (2015)). Nonetheless, we believe it can serve as a useful first step for effective and more targeted auditing to reduce Medicare overbilling.

We apply our approach to detect potential Medicare overbilling using two waves of Medicare Part B Fee-for-Service (FFS) physician payment data. We construct conservative estimates for physicians' implied hours worked treating Medicare Part B FFS beneficiaries in 2012 and 2013.[2, 3] We find that about 2,300 physicians in our sample billed for more than 100 hours per week for Medicare Part B FFS patients alone. We consider such long hours of work highly implausible and refer to these physicians as "flagged physicians." A comparison with the unflagged physicians shows that flagged physicians are more likely to work in smaller group practices, more likely to be a specialist rather than a primary care physician, and provide both more and higher-intensity services. Results from simple regression analysis also suggest that the coding patterns of the flagged physicians are sensitive to variations in the marginal revenue of choosing a higher intensity code. Interestingly, the revenues from these higher-intensity services are not enough to offset the "longer" hours needed to furnish them, resulting in substantially lower reported hourly revenues than the unflagged physicians. Moreover, this large gap in hourly revenues is hard to reconcile using observable physician characteristics and geographical variations.

Our research is related to the literature on the prevalence and consequences of overbilling.

[2]CMS released the data to the public in April 2014.

[3]Details on how we construct the estimates for physicians' hours worked based on the service codes submitted to Medicare are described in Section 3.

Lorence and Spink (2002) surveyed organizational providers and found significant "coding optimization," despite serious penalties if the fraudulent billing practices were found out. Angeles and Park (2009) showed that upcoding imposed unnecessary cost to the already expensive Medicare program, was especially serious for Medicare Advantage, and this problem may become worse as the 76 million baby boomers age. The Department of Health and Human Services, alarmed at the rapid increase in Medicare spending from 2001 to 2010, conducted an in-depth study on the coding trends of evaluation and management (E/M) services in 2012, and found some physicians consistently billed higher-level codes. It is also related to the literature on possible determinants of overbilling. Adams et al. (2002) noted that the long documented difficulty of billing may lead to more erroneous coding, and provide room for fraudulent coding at the same time. Other factors examined include pressure from the management teams (Lorence and Spink (2002); Dafny and Dranove (2009)), hospital ownership (Silverman and Skinner (2004)), anti-fraud enforcement effort (Becker et al. (2005); Bastani et al. (2015)), fee differentials across codes (Brunt (2011); Bowblis and Brunt (2014)), and information technology such as electronic health records (EHR) (Adler-Milstein and Jha (2014)).

Finally, our paper is related to the recent growing list of papers that used the newly released CMS Physician Utilization and Payment data. Most studies that use this dataset look at utilization and/or payment patterns of a particular specialty or procedure (Bergman et al. (2014); Harewood et al. (2014); Clair and Goyal (2015); Dusetzina et al. (2015); Ip et al. (2015); Ko et al. (2015); Lapps et al. (2016); Menger et al. (2015); Schmajuk et al. (2014); Sutphin et al. (2014); Skolarus et al. (2015); Skolasky and Riley III (2015)). For example, Bergman et al. (2014) studied physician payments in general and found that high physician earnings were mainly driven by more services furnished per patient instead of a larger number of patients.

The remainder of the paper is structured as follows. In Section 2, we describe the data and the construction of our sample. In Section 3, we discuss our approach to estimate physician hours worked. In Section 4, we present our empirical results. In Section 5, we corroborate some of our conclusions using two external datasets, the National Ambulatory Medical Care Survey and the CMS Comprehensive Error Rate Testing (CERT) results. Finally, in Section 6, we conclude.

2 Data and Sample Construction

Our main data source is the Medicare Part B FFS Physician Utilization and Payment data released annually to the public by the CMS since April 2014. The two waves of data available now

are derived from all Medicare Part B FFS claims made in 2012 and 2013, respectively.[4] Each wave of data has about 9 million records at the provider-place-service level. Providers are uniquely identified by their National Provider Identifier (NPI) and characterized by a limited set of basic information (e.g. address, individual or organization indicator, gender and specialty). Places are categorized into office settings and facility (such as hospitals) settings, and reflect where the provider furnished a service. Services are identified by a 5-digit alpha-numeric code specified in the Healthcare Common Procedure Coding System (HCPCS). Hence each observation is a summary of a provider's rendition of a service at a place within the calendar year, as well as the payment the provider received for these services.

We supplement the Physician Utilization and Payment dataset with three other publicly available datasets. First we use the CMS Physician Compare database to get more detailed physician characteristics such as education background and group practice affiliations. Both datasets identify physicians by their NPI, and we are able to match 91 percent of the records. Then we use the National Physician Fee Schedule to get the Relative Value Units (RVUs) that quantify the amount of work required to furnish each service, which we will use to estimate the time needed for the services. Finally, we use a CMS on-site survey (Zuckerman et al. (2014)) that objectively measured the time needed for a subset of services to corroborate our estimates of physician hours worked.

Next, we construct our sample by including only providers that are individuals ("physicians") instead of organizations, who work in the continental U.S., completed professional medical training between 1946 and 2011, and have valid basic information (practice location, gender, and specialty). The first selection criterion discards about 5 percent of observations and the rest lead to negligible reduction of the sample size.[5, 6] We further restrict our sample to include only HCPCS codes that are actually services; exclude codes that are drugs, equipment or medical supplies, are only for quality administration purposes and not paid for, or are temporary codes for new services.[7, 8]

Finally, we aggregate the physician-place-service level data to the physician-service level. For each physician-service combination, we observe the physician's characteristics, workload to furnish the service (RVU and/or time needed), volume of the service billed each year (the number of times that service is furnished and the number of Medicare Part B FFS beneficiaries receiving that

[4]For every physician, the HCPCS codes claimed for fewer than 10 times in a calendar year are excluded from the datasets to protect patient privacy. Claims for durable medical equipments are also excluded.

[5]For expositional simplicity, we will refer to all individual providers as "physicians" even if a small fraction of them are nurses or physician assistants.

[6]We exclude those graduated in or after 2012 because they are likely to be residents, who are known for extremely long working hours. We discuss this in greater detail in the Online Appendix and show that our results are not affected when more possible residents are excluded from the sample.

[7]Drug codes are excluded when we estimate total physician hours worked but not when calculating total revenues.

[8]Temporary codes have no RVU information that we can use to reliably estimate the time needed to furnish them.

service), and total Medicare payments for these services. In the final sample we have 7.9 million observations on 623,959 physicians and 4,480 HCPCS service codes.

3 Measuring Physician Hours Worked

We define *physician hours worked* to be the total time a physician spent with patients to furnish the service codes submitted to Medicare for reimbursement. The time needed per service is readily available for some codes (referred to below as the "timed codes"). We use these time codes to estimate the time needed for other service codes that do not have a time requirement based on their Work Relative Value Units (Work RVUs) which are measures of workloads that CMS assigns to all services.

3.1 Timed Codes

The timed codes, which are the cornerstone of our estimation of physician hours worked, fall into two categories. The first category of timed codes have a suggested or required amount of time in their definition. These are mostly from the "evaluation and management" (E/M) code group, which include office or home visits. An important feature of these services is that there are usually multiple codes with different levels of intensity or complexity to furnish even for a narrowly-defined service, and the physician has discretion over which one to file. The American Medical Association (AMA) publishes guidelines on choosing the most appropriate code, and usually includes typical time needed for E/M codes (Gabbert et al. (2012)). As an example, Table 1 shows a typical cluster of E/M codes where multiple codes are available for the *same* service but have varying workload requirement and fees. All five HCPCS codes, 99201 through 99205, are for "office or other outpatient visit for new patients." But the lowest intensity code, 99201, only needs 10 minutes to furnish per the AMA guidelines, and generates $31.09 of revenue, whereas the highest intensity code, 99205, needs 60 minutes and generates $145.81.[9] Note that, incentive issues aside, if a physician were to overstate the service intensity by one level, revenue would increase by at least $20.

The second category of timed codes are those selected in a 2014 CMS survey that directly measures the time needed for certain services (Zuckerman et al. (2014)). The survey targets 112 HCPCS codes that are judged to be growing fast, frequently billed, or often billed together. These codes make up 18 percent of total Medicare physician fee schedule expenditures. Survey staff are sent on site to document the time used to furnish the interested services at several participating

[9]These fees are the baseline reimbursement amounts in the 2012 Physician Fee Schedule. Actual Medicare payments will vary slightly across geographic regions and specific settings in which the services are furnished.

HCPCS code	Typical time needed	Work RVU	2012 price ($)
99201	10 minutes	0.48	31.09
99202	20 minutes	0.93	53.54
99203	30 minutes	1.42	77.47
99204	45 minutes	2.43	118.18
99205	60 minutes	3.17	145.81

Table 1: Example of codes with varying intensity and time needed for the same service
NOTES: All five codes are for "office or other outpatient visit for new patient." The 2012 prices are for services furnished in office settings prior to the adjustment using Geographic Practice Cost Indices (GPCI).

institutions with large volumes of these service.

Our idea is to use the time requirement for timed codes described above to estimate the time requirement for all other codes. In order to do this, we construct the *expected time needed* for each code based on the "typical time needed" suggested by the AMA guideline.[10] This is important because the actual time to furnish a service code may vary both across and within physicians. We construct the "expected time needed" from the "typical time needed" as in AMA guideline as follows. Assuming the time needed follows a uniform distribution, we take the simple average of the minimum and maximum time allowed for each code to get the expected time. Specifically, some codes may have an explicit range of time needed, such as "5-10 minutes of medical discussion." For such codes, the expected time needed is simply the average of the lower and upper bounds. For codes that do not have such a range, physicians are supposed to file the code whose typical time needed is closest to the actual time spent. For example, between codes 99202 and 99203 as described in Table 1, a physician who spent 23 minutes should file the code 99202 instead of 99203. Following this logic, the expected time needed we will assign to HCPCS codes 99201 through 99205 are 7.5, 20, 31.25, 45, and 60 minutes, respectively. To see this, consider HCPCS code 99201 for example. Note that physicians who spends 0 to 15 minutes with a new patient is supposed to file HCPCS code 99201 if they follow the AMA guideline. Thus, under the plausible assumption that the actual time spent with patients follows a uniform distribution, the simple average of the minimum (0 minute) and maximum (15 minutes) time allowed for filing 99201 is 7.5 minutes. For the highest intensity codes within a cluster of codes, e.g. code 99205 in Table 1, we set the expected time to be the same as the typical time as there is no upper bound specified in the AMA guideline. In order to err on the conservative side, we moreover choose the smaller of the typical time for a service code and expected time we construct for the timed codes whenever the two differ. Finally,

[10] An exception is when the AMA guideline *requires* the physician to spend a certain amount of time when furnishing a service. For example, code 99360 is for "physician standby service, requiring prolonged physician attendance, each 30 minutes (eg, operation standby, standby for frozen section, for cesarean/high risk delivery, for monitoring EEG)," and explicitly prohibits filing this code for services less than 30 minutes.

we also exclude timed codes that do not require direct contact of the physician with the patient, such as intravenous drug infusions and phototherapies, again to be conservative in our estimate of the physicians' hours worked.

3.2 Work RVUs and Time Needed for Untimed Codes

Next we estimate the time needed for all other codes for which AMA guideline does not specify typical or required time. Our estimation is based on the Relative Value Units (RVUs) of the service codes. RVUs reflect the value of each HCPCS code along three dimensions. AMA appoints a special committee of experts from various specialties to assign and regularly update the RVUs; and the CMS uses them to determine Medicare reimbursements to physicians. The Physician Fee Schedule specifies the following formula for the baseline payment amount for a given HCPCS code:

$$\text{Payment} = \left[\begin{array}{c} (\text{Work RVU}) \times (\text{Work GPCI}) \\ +(\text{PE RVU}) \times (\text{PE GPCI}) \\ +(\text{MP RVU}) \times (\text{MP GPCI}) \end{array} \right] \times \text{CF},$$

where Work RVU captures the amount of work, primarily time, needed to furnish the service; PE RVU captures the practice expense (PE) of the service; MP RVU captures the malpractice insurance cost of the service; the Geographic Practice Cost Indices (GPCI) adjust for the geographic differences in the costs of practicing medicine, and vary slightly around 1 across 90 GPCI regions in the U.S.; and finally, the conversion factor (CF) translates the RVUs into dollar amounts, which is \$24.6712 per RVU in 2012, and \$34.023 per RVU in 2013 (Centers for Medicare and Medicaid Services (2013)).

We use *Work RVUs* to estimate the time needed for untimed codes. Though Work RVUs are imperfect measures of service time, time still plays the central role when Work RVUs are determined, making Work RVUs the best tool available for our purposes.[11] Our estimation takes two steps. First, we take the timed codes, for which we know the time needed and the Work RVUs, and estimate the *time needed per work RVU*. We use both the simple averages and regressions for robustness, and control for 15 service code groups to account for the difference in practice patterns across specialties. Second, we use the Work RVU of each code, whether it is a timed code or an untimed code, and the estimated time needed per Work RVU from the first step to calculate its time needed.

Thus, for every timed code, we will have three measures of time needed: its expected time needed per AMA guideline, and two estimated time needed using the two estimation methods in

[11] See Centers for Medicare and Medicaid Services (2014).

the first step; and for every untimed code, we will have the latter two measures of time needed. Again to err on the conservative side, we pick the minimum of the measures of the time needed for each service code. In the end, we get positive estimates of time needed for 75% of the HCPCS codes. Codes that do not have such estimates, which we will refer to as "*zero-time codes*," are drugs or supplies that do not require direct contact with the physician, or those that have negative time needed estimates according to our procedure.

Given the time needed estimates of the HCPCS codes, we calculate physician i's total hours worked in year t based on services i billed CMS for in calendar year t:

$$(\text{Hours worked})_{it} = \sum_{j \in J} \left[\text{Time needed per service } j \times (\text{Number of service } j \text{ billed})_{it}\right], \qquad (1)$$

where j is a HCPCS code in the set of codes, J, for which we have obtained positive time needed using the procedure described above.

3.3 Discussion of Estimated Physician Hours Worked

Our estimates of physician hours worked are likely to be a conservative *lower bound* of the actual hours, provided that the service codes truthfully reflect both the volume and the intensity of services the physicians actually furnished. First, as we described above we make every decision in the construction of the hours worked to err on the conservative side. Second, the Medicare Part B FFS Physician Utilization and Payment data that we use in our estimation only include Medicare Part B FFS claims, which on average account for less than 31% of a physician's services (see The Physicians Foundation (2012)). Third, as we mentioned in Footnote 4, for each physician the dataset excludes the HCPCS codes claimed for fewer than 10 times in a calendar year. Fourth, we only include the time needed for 75 percent of HCPCS codes that represent services requiring direct contact of the physician and the patient, have non-zero work RVUs, and end up with positive time estimates. Finally, some physicians bill under the NPI of organizational providers (e.g. a hospital or a group practice), which we exclude from our sample because it is impossible to identify an individual physician's contribution to the organizations' billing records.

4 Describing Physician Hours Worked

We convert the total hours worked in year t to *hours worked per week* in year t for easier interpretation in the analysis that follows, assuming physicians work 51 weeks (i.e., take only 9 days off each year). By doing so we essentially characterize physician hours worked *averaged* over

the entire year. Hence we are allowing the physicians to be possibly smoothing hours of work intertemporally during the year, which of course is another conservative choice that is likely to lead to under-detection of overbilling. This choice is necessitated by the data limitation that utilization and payment records are aggregated to the calendar year level, and not at a higher time frequency. However, if we are able to detect implausibly long hours worked per week under the lenient criterion permitting cross-week smoothing, it would serve as a stronger signal for potentially inappropriate coding.

Figure 1 graphs the distribution of average reported hours worked per week across all the physicians. Despite the conservative methods we used to estimate the physicians' hours worked, about 2,300 physicians submitted claims for service codes that would translate into over 100 hours per week on services for Medicare Part B FFS beneficiaries. Moreover, about 600 physicians submitted claims for service codes that would imply over 168 hours per week (that is, 24 hours a day, 7 days per week!). To put these numbers into perspective, the Accreditation Council for Graduate Medical Education (ACGME) restricts residency working hours to 80 hours per week since 2003 in light of the much studied sleep deprivation and performance deterioration of health care providers (Wolman et al. (2009)). Furthermore, as we will discuss in Section 5, the *maximum hours* spent on Medicare patients by physicians in National Ambulatory Medical Care Survey data are below 50 hours in a week.

In Table 2 we take a closer look at physicians who billed Medicare for long hours. We use different weekly hours as *flagging* thresholds, 80, 100, 112 (16 hours per day for 7 days), and 168 (24 hours per day for 7 days) respectively, and present the statistics by year. We will refer to physicians whose estimated weekly hours worked above the threshold as *flagged physicians* and those below as *unflagged physicians.* For example, under the 100 weekly hours threshold, we flag 2,292 physicians in 2012 and 2,120 physicians in 2013 as having submitted claims with implied hours worked exceeding that threshold. They account for 2.71% and 2.55% of all physicians in our data that have submitted claims implying at least 20 hours of service per week in at least one year, and 0.367% and 0.340% among all physicians in 2012 and 2013, respectively.

Physicians with very few implied hours worked in our sample could have few Medicare patients, or could have just as many Medicare patients but they specialize in the 25 percent zero-time service codes where information on time needed is unavailable. If it is the latter, one might be concerned that our results overlook physicians who are only overbilling on the zero-time codes. Table 2 indeed shows that, for example, the number of distinct zero-time HCPCS codes as a fraction of all 4,480 distinct HCPCS codes filed by flagged physicians ranges from about 9% using the 80-hour threshold to about 4% using the 168-hour threshold, while the corresponding fraction for unflagged physicians

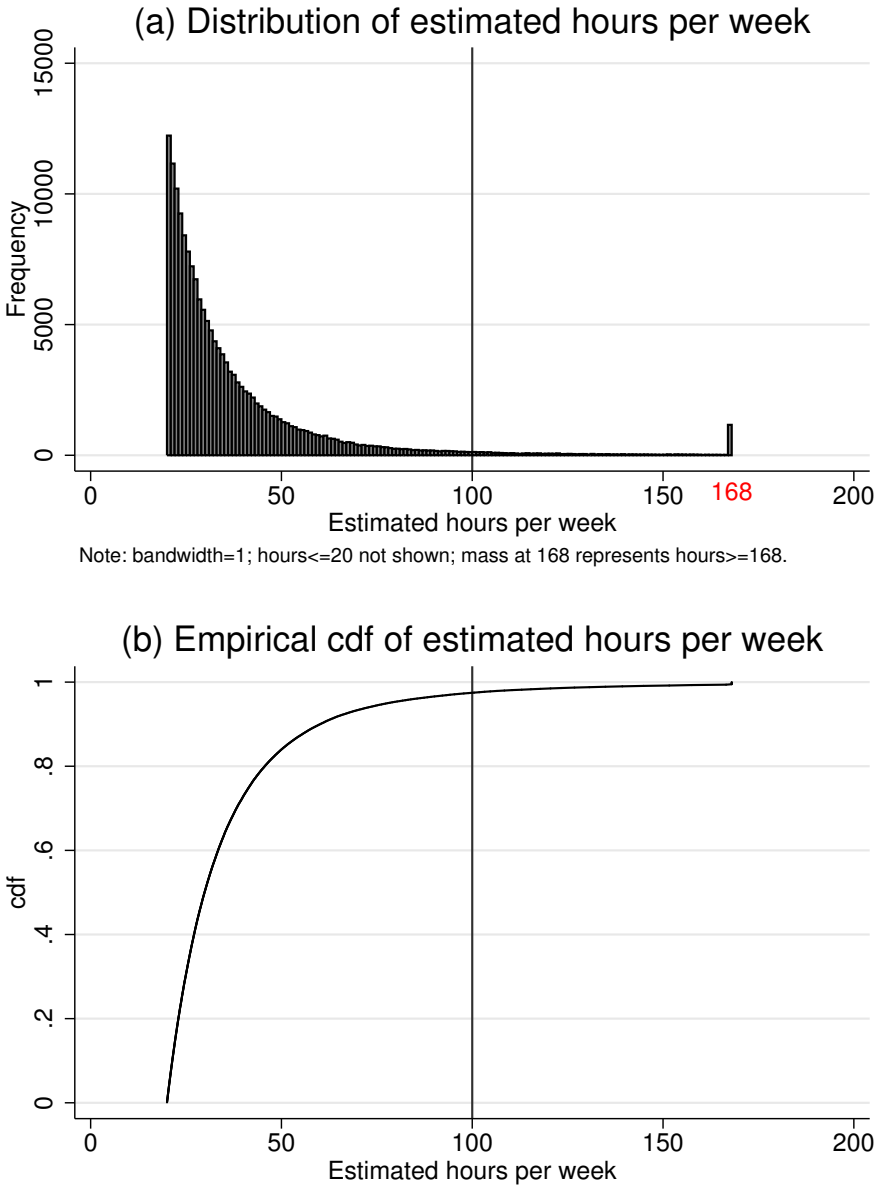


Figure 1: Distribution of estimated hours per week

NOTES: The horizontal axis shows physicians' hours worked per week estimated from the claims they submitted to Medicare for reimbursement. We restrict the sample to physicians whose claims imply at least 20 hours per week in at least one year. The mass at 168 per week in the top panel represents hours larger than or equal to 168 per week.

Hours threshold	80+		100+		112+		168+	
Year	2012	2013	2012	2013	2012	2013	2012	2013
Number of physicians flagged	4125	3838	2292	2120	1689	1546	615	530
% in physicians working 20+ hr/week	4.879	4.618	2.711	2.551	1.998	1.860	0.727	0.63
% in all physicians	0.661	0.615	0.367	0.340	0.271	0.248	0.099	0.085
% zero-time codes (flagged)	9.008	9.147	7.352	7.377	6.430	6.683	4.141	4.250
% zero-time codes (unflagged*)	13.17	13.30	13.16	13.30	13.15	13.30	13.14	13.30
% wRVU of zero-time codes (flagged)	0.171	0.172	0.115	0.108	0.107	0.098	0.033	0.054
% wRVU of zero-time codes (unflagged*)	0.591	0.611	0.590	0.611	0.590	0.611	0.590	0.611
% volume of zero-time codes (flagged)	11.43	11.85	9.173	9.391	8.152	8.134	4.104	4.421
% volume of zero-time codes (unflagged*)	15.70	15.83	15.69	15.83	15.68	15.83	15.67	15.83
% revenue from zero-time codes (flagged)	8.112	9.306	7.106	8.025	6.372	7.018	3.557	3.772
% revenue from zero-time codes (unflagged*)	7.442	7.743	7.435	7.743	7.432	7.743	7.426	7.743
Total number of physicians working 20+ hr/week	96,033							
Total number of physicians	623,959							

Table 2: Number and fraction of physicians flagged

NOTES: The table reports the number and fraction of flagged physicians in calendar years 2012 and 2013. "Hours threshold" shows the cutoff number of hours billed per week above which a provider is flagged. "% 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. "% 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 time needed estimates are not available and account for 25% 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.

Hours threshold	80+		100+		112+		168+	
Year(s) flagged	count	share (%)	count	share (%)	count	share (%)	count	share (%)
2012 only	1,135	27.52	704	30.72	539	31.91	233	37.89
2012 and 2013	2,990		1,588		1,150		382	
2013 only	848	22.09	532	25.09	396	25.61	148	27.93

Table 3: Flag patterns across time

NOTES: The table the flagged pattern across time from calendar year 2012 to 2013. "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.

is over 13% using any threshold. However, when weighted by Work RVUs of the service codes, the differences are much smaller. This is consistent with the fact that many of the zero-time codes have low Work RVUs. Similarly, the differences between flagged and unflagged physicians are big when we compare the unweighted number of claims for zero-timed codes, but shrink significantly once weighted by revenues. These suggest that specialization in different types of services is unlikely to result in large under-flagging of physicians, though having differential fractions of Medicare patients still is.[12]

Table 3 decomposes the flagged physicians into those flagged in 2012 only, in 2013 only, and in both years. For example, 1,135 physicians are only flagged in 2012 using the 80-hour threshold, making up 27.52% of all 4,125 physicians flagged in 2012.

4.1 Who Reported Implausibly Long Hours?

In Tables 2-3, we used four different weekly hours thresholds to flag physicians. From now on, we focus on physicians flagged using the 100-hour threshold, although all results hold under alternative flagging criteria.[13] In addition, we only focus on the subsample of 96,033 physicians with more than 20 hours worked per week treating Medicare Part B FFS patients, as we believe this is the more relevant group to be compared with the flagged physicians for reasons discussed earlier.

In Table 4 we compare the characteristics of physicians across the following groups, according to the column headings: (1) all physicians, (2) never flagged, (3) flagged in any year, (4) flagged in 2012, (5) flagged in 2013, (6) flagged only in 2012, (7) flagged in both 2012 and 2013, and (8) flagged only in 2013. Table 4 shows that flagged physicians are slightly more likely to be male, non-MD, more experienced, and provide fewer E/M services. Importantly, they work in substantially

[12] We can not address this issue given the fact that our data only contains Medicare claims. However, our method can be easily extended to a more general setting with augmented data from physician billing information for beneficiaries of other insurance programs.

[13] We present the results under the 112 and 168 weekly hours flagging thresholds in the Online Appendix.

	(1) All	(2) Never	(3) Ever	(4) 2012	(5) 2013	(6) 2012 only	(7) Both	(8) 2013 only
1(Male)	0.857	0.856	0.891	0.896	0.896	0.877	0.904	0.867
	(0.001)	(0.001)	(0.006)	(0.007)	(0.007)	(0.013)	(0.008)	(0.017)
1(MD)	0.838	0.844	0.660	0.667	0.714	0.500	0.741	0.624
	(0.001)	(0.001)	(0.010)	(0.010)	(0.011)	(0.020)	(0.012)	(0.024)
Experience (years)	24.14	24.12	24.69	25.14	24.20	26.15	24.70	22.71
	(0.034)	(0.034)	(0.191)	(0.208)	(0.218)	(0.393)	(0.243)	(0.470)
# providers in group	87.19	88.92	29.97	31.47	29.63	31.01	31.67	23.55
	(0.869)	(0.890)	(2.981)	(3.416)	(3.538)	(5.438)	(4.302)	(5.825)
# hospital affiliations	2.774	2.813	1.495	1.535	1.512	1.445	1.576	1.321
	(0.006)	(0.006)	(0.035)	(0.039)	(0.040)	(0.072)	(0.046)	(0.077)
1(in Medicare)	0.857	0.857	0.874	0.878	0.874	0.872	0.880	0.855
	(0.001)	(0.001)	(0.006)	(0.007)	(0.007)	(0.013)	(0.008)	(0.015)
1(in ERX)	0.463	0.463	0.466	0.483	0.500	0.365	0.535	0.395
	(0.002)	(0.002)	(0.009)	(0.010)	(0.011)	(0.018)	(0.013)	(0.021)
1(in PQRS)	0.396	0.396	0.399	0.404	0.424	0.327	0.439	0.378
	(0.002)	(0.002)	(0.009)	(0.010)	(0.011)	(0.018)	(0.012)	(0.021)
1(in EHR)	0.416	0.417	0.397	0.397	0.394	0.406	0.394	0.395
	(0.002)	(0.002)	(0.009)	(0.010)	(0.011)	(0.019)	(0.012)	(0.021)
Types of codes 2012	22.46	22.43	23.39	24.49	24.58	19.82	26.56	18.66
	(0.053)	(0.053)	(0.373)	(0.431)	(0.439)	(0.678)	(0.536)	(0.650)
Types of codes 2013	22.38	22.35	23.32	24.09	24.96	18.39	26.62	20.02
	(0.052)	(0.053)	(0.368)	(0.423)	(0.443)	(0.598)	(0.538)	(0.695)
Types of E/M codes 2012	6.179	6.228	4.573	4.678	4.551	4.639	4.695	4.120
	(0.014)	(0.015)	(0.076)	(0.086)	(0.085)	(0.166)	(0.100)	(0.163)
Types of E/M codes 2013	6.158	6.207	4.553	4.623	4.593	4.430	4.708	4.252
	(0.014)	(0.014)	(0.076)	(0.085)	(0.086)	(0.161)	(0.099)	(0.169)
N	96,033	93,209	2,824	2,292	2,120	704	1,588	532

Table 4: Characteristics of flagged physicians vs unflagged physicians

NOTES: The table compares the means of physician characteristics across subgroups (standard errors of the mean estimator are reported in parentheses). We restrict the sample to physicians billing at least 20 hours per week in at least one year. "All" refers to all physicians in this sample. "Never" refers to physicians never flagged in any year. "Ever" refers to those flagged in at least one year. "2012" and "2013" refer to those flagged in 2012 and 2013, respectively. "2012 (2013) only" refers to those only flagged in 2012 (2013) but not the other year. "Both" refers to those flagged in both years. Physician experience is imputed from the year of graduation. # providers in group refers to the number of providers in the group practice where the billing physician works at. It is 1 if the billing physician does not work in a group practice. The number of hospital affiliations are 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.

	(1) Ever	(2) 2012	(3) 2013	(4) 2012 only	(5) Both	(6) 2013 only	Mean of Never
1(Male)	0.032***	0.037***	0.037***	0.015	0.046***	0.007	0.856
	[0.007]	[0.007]	[0.008]	[0.011]	[0.009]	[0.018]	
1(MD)	-0.178***	-0.171***	-0.126***	-0.338***	-0.100***	-0.217***	0.844
	[0.034]	[0.037]	[0.032]	[0.053]	[0.036]	[0.038]	
Experience (years)	0.452*	0.799***	0.046	1.624***	0.429	-1.244**	24.124
	[0.245]	[0.278]	[0.252]	[0.602]	[0.279]	[0.593]	
# providers in group	-52.349***	-49.660***	-52.027***	-53.157***	-48.119***	-65.027***	88.919
	[5.991]	[6.139]	[5.981]	[10.174]	[5.969]	[10.346]	
# hospital affiliations	-1.392***	-1.376***	-1.338***	-1.565***	-1.300***	-1.476***	2.813
	[0.124]	[0.130]	[0.114]	[0.233]	[0.116]	[0.170]	
1(in Medicare)	0.004	0.003	0.001	0.014	0.000	0.009	0.857
	[0.007]	[0.008]	[0.008]	[0.012]	[0.009]	[0.016]	
1(in ERX)	-0.003	0.007	0.030*	-0.107***	0.057***	-0.061**	0.463
	[0.016]	[0.019]	[0.016]	[0.026]	[0.020]	[0.025]	
1(in PQRS)	-0.001	0.002	0.026	-0.086***	0.041**	-0.022	0.396
	[0.016]	[0.017]	[0.016]	[0.025]	[0.017]	[0.025]	
1(in EHR)	-0.028**	-0.023	-0.032**	-0.012	-0.028*	-0.045*	0.417
	[0.014]	[0.016]	[0.015]	[0.029]	[0.017]	[0.024]	
Types of codes 2012	0.220	0.966	1.460*	-3.611**	2.906***	-3.495***	22.428
	[0.905]	[1.027]	[0.889]	[1.494]	[1.045]	[0.934]	
Types of codes 2013	0.393	0.872	2.059**	-4.705***	3.248***	-2.034**	22.351
	[0.905]	[1.018]	[0.901]	[1.340]	[1.047]	[1.005]	
Types of E/M codes 2012	-1.826***	-1.814***	-1.793***	-1.945***	-1.764***	-1.900***	6.228
	[0.193]	[0.203]	[0.199]	[0.323]	[0.212]	[0.247]	
Types of E/M codes 2013	-1.780***	-1.792***	-1.683***	-2.087***	-1.671***	-1.737***	6.207
	[0.193]	[0.201]	[0.198]	[0.313]	[0.210]	[0.245]	
Num. of physicians in group	2,824	2,292	2,120	704	1,588	532	93,209

Table 5: 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. Standard errors clustered at the HRR level are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Definition of the physician characteristics variables are the same as specified in the notes of Table 4.

smaller group practices (if at all), and have fewer hospital affiliations.[14] These characteristics are similar to what Cutler et al. (2013) found about physicians who "consistently and unambiguously recommended intensive care beyond those indicated by current clinical guidelines."

To account for the heterogeneity in physicians' exposure to local Medicare markets, we compare the same characteristics controlling for Hospital Referral Region (HRR) fixed effects in Table 5. The 306 HRRs represent local health care markets and are commonly used as the unit of analysis for regional variations of health care in the U.S. (Wennberg and Cooper (1996)). The number in each cell in Table 5 is the estimated coefficient from an OLS regression on the subset of physicians

[14]The CMS Physician Compare data, from which we obtain the physicians' characteristics does not report their race and age.

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. For example, physicians ever flagged are 3.2 percentage points more likely to be male, 17.8 percentage points less likely to have a MD, and tend to practice in groups with 52.3 fewer providers, etc. As it is clear from the comparisons of Tables 4 and 5, the differences between flagged and unflagged physicians remain qualitatively unchanged even after taking into account the HRR fixed effects.

4.2 What are the Specialties of Flagged Physicians?

In addition to the individual characteristics of flagged physicians, we are also interested in whether some specialties are more likely to be associated with flagged physicians. For this purpose, we follow Fryer and Levitt (2004)'s approach toward quantifying the "blackness" of first names, and construct the *Specialty Flag Index* (SFI) for specialty s:

$$\text{SFI}_s = \frac{100 \times \Pr\left(s|\text{flagged}\right)}{\Pr\left(s|\text{flagged}\right) + \Pr\left(s|\text{unflagged}\right)} \tag{2}$$

where the conditional probability $\Pr\left(s|\text{flagged}\right)$ is defined as the fraction of flagged physicians in specialty s among all flagged physicians, and $\Pr\left(s|\text{unflagged}\right)$ is the fraction of unflagged physicians in specialty s among all unflagged physicians.[15] The index ranges from 0 to 100. If all physicians in specialty s are flagged, then SFI_s takes on a value of 100. If only unflagged physicians are specialty s, then SFI_s is 0. If flagged and unflagged physicians are equally likely to be in specialty s, then SFI_s is 50. If flagged physicians are four times as likely to be in specialty s than unflagged physicians, then $\text{SFI}_s = 100 \times 4/(4+1) = 80$. This measure is invariant to the fraction of the flagged physicians among all physicians, and to the overall popularity of the specialty among all physicians.

The SFI is a convenient summary of how a given specialty is represented among the flagged relative to its share in the entire physician population. A SFI of 50 indicates that specialty is "fairly represented" among the flagged, i.e. $\Pr\left(s|\text{flagged}\right)$ and $\Pr\left(s|\text{unflagged}\right)$ are both equal to the fraction of specialty s among all physicians. A SFI above 50 indicates that the specialty is over-represented among the flagged physicians.

Table 6 ranks the top specialties with at least 50 flagged physicians by their SFIs. For example, optometry is considerably over-represented among the flagged physicians, accounting for more than 20% of flagged physicians but less than 2% of all physicians, leading to SFIs over 90 in

[15] We use the self-reported primary specialty when a physician is in multiple specialties.

Specialty\Year	% in all	Num. unflagged		Num. flagged		SFI	
		2012	2013	2012	2013	2012	2013
Optometry	1.893	1,252	1,390	566	428	94.87	93.17
Dermatology	4.185	3,557	3,525	463	495	84.19	86.15
Ophthalmology	7.960	7,258	7,260	386	384	68.50	70.09
Pathology	2.746	2,567	2,578	71	60	53.08	50.76
Nephrology	4.900	4,607	4,615	99	91	46.78	46.62
Cardiology	11.120	10,543	10,579	136	100	34.54	29.51
Internal Medicine	11.089	10,573	10,567	77	83	22.95	25.81
All physicians		93,741	93,913	2,292	2,120		

Table 6: Physician specialties and flag status

NOTES: The table shows seven specialties with the highest SFI, defined in Equation (2), among the specialties with at least 50 flagged physicians. "% in all" shows the fraction 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.

both years. On the contrary, internal medicine physicians are much under-represented among the flagged physicians, with SFIs around 25. Moreover, note how the SFI as defined in (2) differs from the simple probability that physicians of a given specialty are flagged, namely $\Pr(\text{flagged}|s)$. For example, 386 (a mere 5%) of the 7,664 ophthalmologists are flagged in 2012, yet ophthalmology still gets a high SFI of 68.5 because it only makes up 7.96% of all physicians in our sample but contributes 16.8% to the 2,292 flagged physicians.

4.3 What Codes Do Flagged Physicians Tend to Bill?

Similarly to how we constructed specialty flag index, we can also construct the *Code Flag Index* (CFI) for each HCPCS code j as follows:

$$\text{CFI}_j = \frac{100 \times \Pr(j|\text{flagged})}{\Pr(j|\text{flagged}) + \Pr(j|\text{unflagged})}, \tag{3}$$

where the conditional probability $\Pr(j|\text{flagged})$ is defined as the number of claims for HCPCS code j filed by flagged physicians as a fraction of the total number of claims for all service codes filed by flagged physicians; and $\Pr(j|\text{unflagged})$ is the number of claims for HCPCS code j filed by unflagged physicians as a fraction of the total number of claims for all service codes filed by unflagged physicians. Like SFI, the CFI takes on values between 0 and 100; if a code is filed only by flagged physicians, then its CFI will be 100; and if a code is filed only by unflagged physicians, then its CFI is equal to 0. A code with a CFI of 50 indicates that it is filed by flagged and unflagged physicians at equal rates.

In Figure 2, we show that there is a nonlinear relationship between a code's CFI and the probability that it is filed by flagged physicians. For example, a HCPCS code with a 20% probability

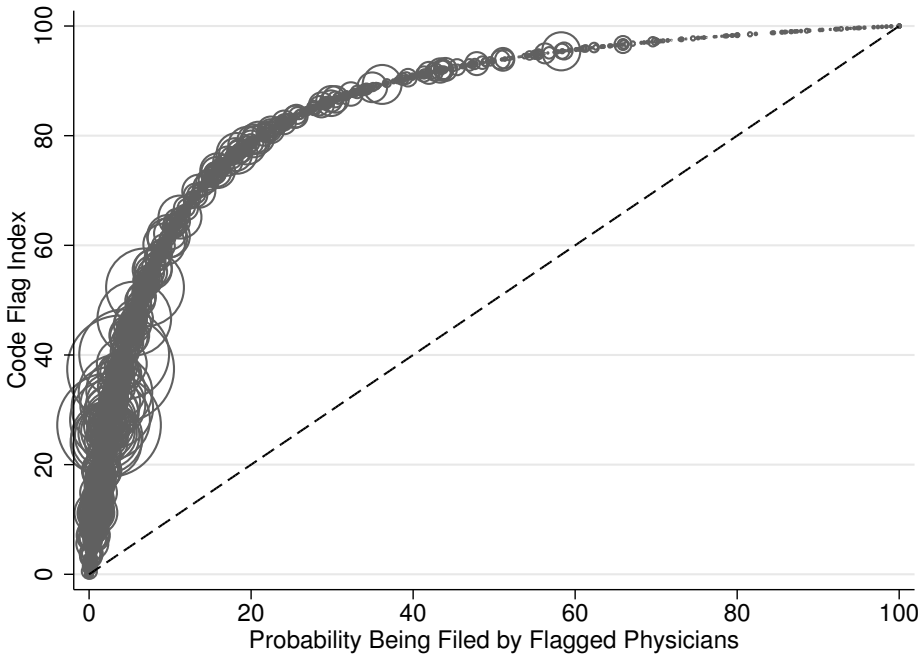


Figure 2: The Relationship between HCPCS Code Flag Index and its Probability of Being Filed by Flagged Physicians

NOTES: The horizontal axis shows the probability of the HCPCS code being filed by a flagged physician (in %). The vertical axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes filed by physicians billing at least 20 hours per week in at least one year. Each circle represents a HCPCS code, with the radius proportional to total revenue. The dashed line is the "45-degree" line.

of being filed by flagged physicians can have a CFI over 75 if unflagged physicians file the code at a much lower rate than flagged physicians.

In Figure 3 we plot the distribution of CFIs among all HCPCS codes in our data.[16] Panel (a) is the unweighted distribution, which is roughly uniform. Panel (b) weighs the codes by their corresponding service volume, i.e., the total number of times they are filed. Panel (c) weighs the codes by their total Medicare reimbursement. A comparison between Panels (b) and (c) shows that HCPCS codes with high CFIs do not necessarily have a lot of volume, but they do have disproportionate costs to Medicare relative to their volume.

In Figure 4 we compare the distributions of CFIs among flagged physicians and that of unflagged physicians, weighted by volume of service. By construction, flagged physicians do tend to report more high-CFI codes.

4.4 Decomposing the Long Hours and Quantifying Potential Overbilling

The long hours worked as implied by the flagged physicians' claims to Medicare can result from *high volumes* of services with a given distribution of service intensity (the "*extensive* margin") and/or a larger fraction of *higher-intensity* services (the "*intensive* margin"). In Table 7, we examine the composition of hours billed by flagged and unflagged physicians.

Table 7 shows some rather interesting differences between flagged and unflagged physicians. On average, flagged physicians submit more than twice as many service claims to Medicare in a year as unflagged physicians, have about twice as many distinct Medicare patients in total, treat about twice as many Medicare patients per day, and provide significantly more services per patient. However, flagged physicians tend to file service claims with longer time requirement, resulting in fewer services furnished per hour than unflagged physicians. Interestingly, the implied Medicare payment per reported hour worked for flagged physicians is significantly lower than that for unflagged physicians. Taking 2012 for example, flagged physicians on average furnish 176% more services per year than unflagged physicians (12,549 versus 4,540); they have 110% more Medicare Part B FFS patients (5,126 versus 2,430); they provide 72% more services on each patient (4.17 versus 2.43). They also tend to provide higher intensity services, which take longer to furnish (1.65 versus 2.88 services per hour, or 36 versus 21 minutes per service) and generate more revenue from Medicare payment ($80.21 versus $74.81). However, since the higher revenue services require longer time to furnish, the Medicare payment per hour for flagged physicians are substantially lower than their unflagged peers ($118.54 versus $162.01).

[16] Of the 4,480 HCPCS codes, about 1,800 have a CFI of 0 in either year, and about 220 have a CFI of 100. These codes are excluded from the figures so as not to distort the scales.

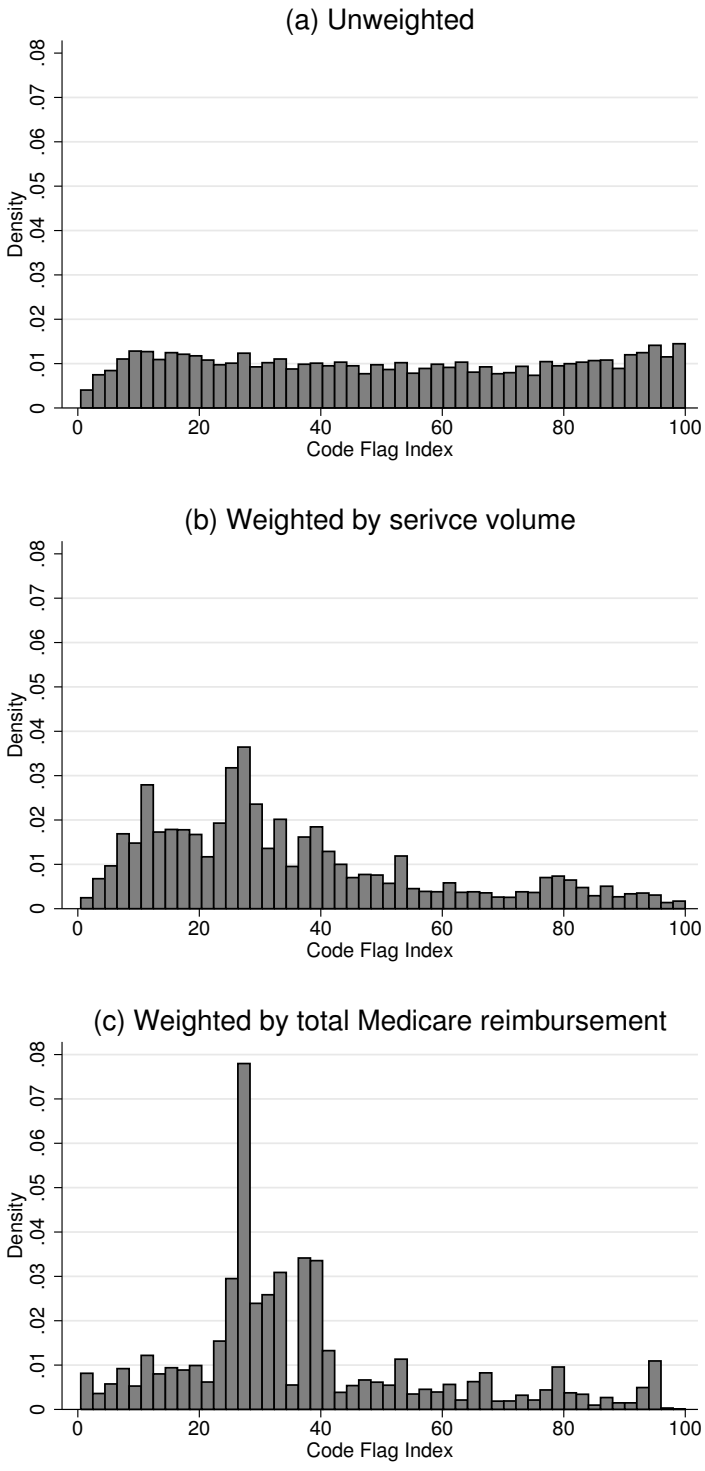


Figure 3: Distribution of HCPCS Code Flag Index

NOTES: The horizontal axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes with CFIs strictly larger than 0 and strictly smaller than 100. Bandwidth is 2 for all three histograms.

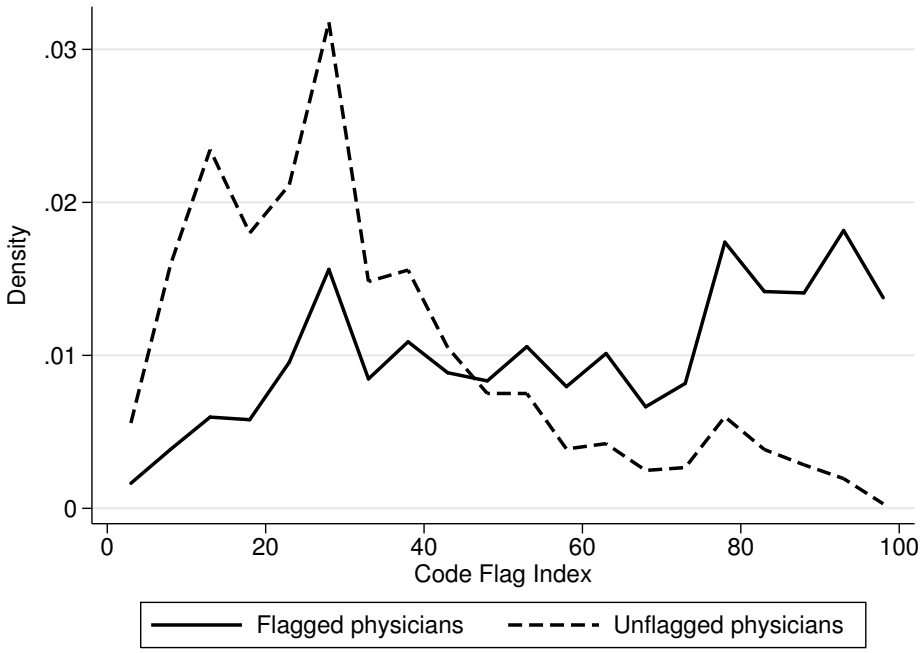


Figure 4: Distribution of Code Flag Index weighted by service volumes: flagged physicians vs. unflagged physicians.

NOTES: The horizontal axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes with CFIs strictly larger than 0 and strictly smaller than 100. The solid line shows the distribution of CFIs of codes billed by flagged physicians, and the dashed line shows that for unflagged physicians. Density is weighted by a HCPCS code's total service volume furnished by all physicians.

	Flagged		Unflagged	
Year	2012	2013	2012	2013
Num. of services provided	12,548.683	12,365.218	4,540.285	4,490.308
	(542.911)	(562.219)	(12.505)	(12.658)
Num. of services per patient	4.167	3.704	2.434	2.376
	(0.096)	(0.089)	(0.013)	(0.009)
Num. of services provided per hour	1.651	1.648	2.880	2.897
	(0.055)	(0.048)	(0.007)	(0.008)
Num. of patients	5,126.103	5,297.308	2,429.509	2,424.339
	(325.043)	(347.764)	(6.546)	(6.582)
Num. of patients per day	14.006	14.513	6.638	6.642
	(0.888)	(0.953)	(0.018)	(0.018)
Num. of patients per hour	0.705	0.725	1.577	1.591
	(0.033)	(0.029)	(0.004)	(0.004)
Medicare payment per service ($)	80.208	83.180	74.811	73.381
	(1.773)	(1.834)	(0.197)	(0.198)
Medicare payment per patient ($)	197.804	193.769	150.639	146.120
	(5.382)	(4.222)	(0.466)	(0.422)
Medicare payment per hour ($)	118.541	118.677	162.010	159.035
	(2.107)	(2.033)	(0.248)	(0.246)
N	2,292	2,120	93,741	93,913

Table 7: Volumes of services supplied: flagged vs. unflagged physicians

NOTES: The table compares the volumes of services furnished by physicians with different flag statuses. Standard errors of the mean estimator are reported in parentheses. We restrict the sample to physicians billing at least 20 hours per week in at least one year. "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) working days in year 2012 (2013). "Per hour" statistics are calculated using the estimated total hours worked of each physician.

	Flagged		Unflagged	
Year	2012	2013	2012	2013
Num. of services provided	7,707.660***	7,769.498***	4,540.285	4,490.308
	[801.128]	[785.011]		
Num. of services per patient	1.665***	1.203***	2.434	2.376
	[0.253]	[0.191]		
Num. of services provided per hour	-1.340***	-1.374***	2.880	2.897
	[0.089]	[0.081]		
Num. of patients	2,571.489***	2,846.790***	2,429.509	2,424.339
	[452.642]	[463.427]		
Num. of patients per day	7.026***	7.799***	6.638	6.642
	[1.237]	[1.270]		
Num. of patients per hour	-0.932***	-0.933***	1.577	1.591
	[0.057]	[0.052]		
Medicare payment per service ($)	8.897**	13.934***	74.811	73.381
	[3.758]	[3.346]		
Medicare payment per patient ($)	51.843***	52.467***	150.639	146.120
	[8.750]	[6.642]		
Medicare payment per hour ($)	-43.367***	-40.131***	162.010	159.035
	[5.619]	[4.808]		
N	2,292	2,120	93,741	93,913

Table 8: 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) working days in year 2012 (2013). "Per hour" statistics are calculated using the estimated total hours worked of each physician.

In Table 8 we show that the results persist when we control for the Hospital Referral Region fixed effects .

The sizable difference in Medicare payment per hour between flagged and unflagged physicians motivates our construction of the *Overbilling Potential Factor* (OPF), which quantifies the extent to which there may be overbilling. We provide two alternative ways to construct the OPF.

Our first measure of overbilling potential is:

$$\text{OPF1}_i \equiv \frac{\text{Total revenue}_i}{\text{Fair revenue}_i} = \frac{\text{Total revenue}_i}{(\text{Fair hourly revenue})_i \times (\text{Fair hours})}, \tag{4}$$

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

	Flagged Physicians	Unflagged Physicians
Reported hourly revenue ($)	116.325	157.434
	(1.439)	(0.172)
Predicted hourly revenue ($)	138.255	159.104
	(0.688)	(0.105)
Overbilling Potential Factor 1	1.907	0.575
	(0.033)	(0.001)
Overbilling Potential Factor 2	5.978	1.150
	(0.162)	(0.003)
N	4,412	187,654

Table 9: Hourly revenues and Overbilling Potential Factors (OPFs)
NOTES: The table compares the hourly revenues and OPFs (defined in Equations (4) and (5)) 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 of the mean estimator are reported in parentheses.

hours" is set to be 8 hours per day times 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.

Our second measure of overbilling potential is to compare the reported hours and the likely true hours, where the latter is the unknown number of hours physicians actually worked. Under the assumption that the goal of overbilling is to achieve the same revenue with fewer *actual* hours, we have, for each flagged physician i:

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

Thus,

$$\text{OPF2}_i \equiv \frac{\text{Reported hours}_i}{\text{True hours}_i} \equiv \frac{\text{Fair hourly revenue}_i}{\text{Reported hourly revenue}_i}, \tag{5}$$

where, as in (4), "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 "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.

Table 9 summarizes the reported and predicted hourly revenues and the two OPFs. For flagged physicians, the reported hourly revenues is $22, or 19%, less than the predicted revenue; but for unflagged physicians, reported and predicted hourly revenues are almost identical. Flagged physicians have an average OPF1 of 1.907, meaning that the total revenue from Medicare Part B

FFS beneficiaries is almost twice as high as that of an unflagged peer with identical observable characteristics, *assuming their actual hours worked are identical* (8 hours per day times 365 days). Similarly, the average OPF2 for flagged physicians is 5.978, suggesting that the reported hours could be six times as much as the likely true hours worked.

Panels (a) and (c) in Figure 5 plot the distribution of OPF1 and OPF2, respectively, among the flagged and unflagged physicians. Note that, despite the heterogeneity within flagged and unflagged physicians, the distributions under both OPFs for the flagged physicians represent a substantial rightward shift of those for the unflagged physicians. In Panel (b) we present the scatter plots of the predicted hourly revenue and the reported hourly revenue. It is clear that the majority of flagged physicians have lower reported hourly revenue than the predicted hourly revenue, whereas the opposite is true for unflagged physicians.

4.5 Coding Decisions and Fee Differentials

We now test whether coding decisions respond to financial incentives provided by different levels of service codes, and examine how they differ between flagged and unflagged physicians. In particular, we are interested in physicians' choice of code *intensities* conditional on filing a code from a given cluster. Using the code cluster presented in Table 1 as an example, we would like to know why a physician bills code 99203 more often than 99202 when a service in this code cluster - office or other outpatient visit for new patients - is furnished.

For this purpose, we analyze the physicians' coding decisions by K, the number of intensity levels in a code cluster, for $K = 3, 4$ or 5. For each $K \in \{3, 4, 5\}$, we use the following baseline regression specification:

$$\begin{aligned} Y_{ijt} \;=\; & \beta_0 + \beta_1 \text{Flagged}_{it} + \beta_{22} 1\{\text{Intensity} = 2\}_j + \cdots + \beta_{2K} 1\{\text{Intensity} = K\}_j \\ & + \beta_{32} \text{Flagged}_{it} \times 1\{\text{Intensity} = 2\}_j + \cdots + \beta_{3K} \text{Flagged}_{it} \times 1\{\text{Intensity} = K\}_j \\ & + \alpha_{\text{HRR}} + \eta_J + \phi_t + \varepsilon_{ijt} \end{aligned} \tag{6}$$

where Y_{ijt} is the number of times physician i filed code j in year t; Flagged_{it} is an indicator for whether physician i is flagged in year t; $1\{\text{Intensity} = 2\}_j$ is an indicator for code j having intensity 2 in its cluster; $1\{\text{Intensity} = K\}_j$ is defined likewise, with K, the highest intensity level, being 3, 4, or 5 depending on the cluster; α_{HRR} is the HRR fixed effect; η_J is the code cluster fixed effect, where J is the cluster that j belongs to; ϕ_t is the year fixed effect; and ε_{ijt} is the error which will be clustered at the physician level.

Our primary interest is on coefficients $(\beta_{32}, \ldots, \beta_{3K})$, which capture the *excess* tendency of

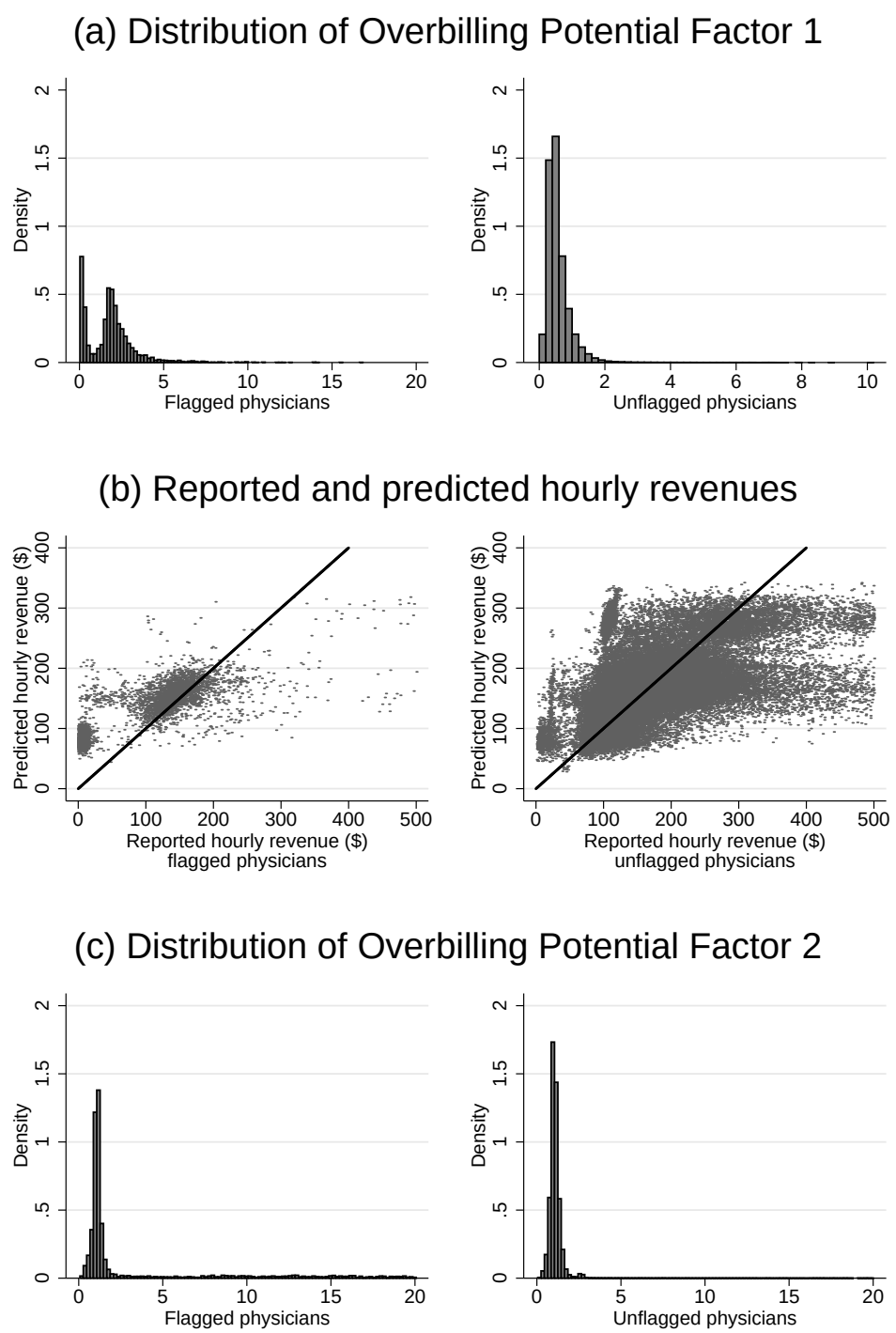


Figure 5: Overbilling Potential Factors (OPFs)

NOTES: Panel (a) shows the distributions of OPF1 for flagged (left) and unflagged (right) physicians. Panel (b) shows the predicted hourly revenues (on the vertical axis), based on OLS regression conditional on physician gender, credential (MD dummy), years of experience, as well as a full set of specialty, HRR, and year fixed effects, against reported hourly revenues (on the horizontal axis) for flagged and unflagged physicians. The thick solid line is the "45-degree" line. Panel (c) shows the distributions of OPF2 for flagged (left) and unflagged (right) physicians. The bin widths in all four histograms are 0.2.

	(1)	(2)	(3)	(4)	(5)	(6)
	$K = 3$	$K = 4$	$K = 5$	All K	All K & below average	All K & above average
Flagged	246.9***	542.9***	37.36**	194.5***	215.7***	156.8***
	[55.52]	[209.9]	[18.46]	[16.63]	[26.50]	[18.84]
Intensity=2	242.2***	169.7***	11.87***			
	[2.796]	[9.993]	[3.230]			
Intensity=3	130.0***	148.6***	240.4***			
	[2.430]	[10.54]	[3.365]			
Intensity=4		-80.39***	235.5***			
		[10.62]	[3.181]			
Intensity=5			33.92***			
			[3.036]			
Flagged × (intensity=2)	369.2***	155.4	91.08***			
	[69.39]	[274.5]	[22.25]			
Flagged × (intensity=3)	128.2**	90.05	143.4***			
	[63.72]	[267.4]	[25.23]			
Flagged × (intensity=4)		257.8	57.47**			
		[272.4]	[23.03]			
Flagged × (intensity=5)			0.115			
			[21.46]			
Mid-intensity				240.0***	21.48***	342.5***
				[1.761]	[1.302]	[2.837]
High-intensity				154.8***	36.74***	186.3***
				[1.509]	[1.277]	[2.499]
Flagged × Mid-intensity				91.72***	-114.4***	231.0***
				[23.25]	[26.77]	[32.79]
Flagged × High-intensity				-21.58	-112.4***	76.75***
				[20.17]	[26.33]	[28.22]
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	0.194	0.057	0.172	0.158	0.164	0.081
Observations	399,907	53,521	561,657	1,015,085	508,478	506,607

Table 10: 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. Furthermore, Columns (1) to (3) are only using the subsamples of code clusters with 3, 4, and 5 levels of intensities, respectively. Columns (4) pool codes in all clusters together, and re-classify intensities to low, middle, and high as specified in Table 11. 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. ** $p < 0.05$, *** $p < 0.01$.

flagged physicians to file codes at varying intensity levels relative to their unflagged peers. A positive β_{3K}, for example, indicates that flagged physicians bill more highest-intensity codes than unflagged physicians. And because higher code intensity translates into more work RVU and thus greater Medicare reimbursement, this is consistent with flagged physicians responding to financial incentives when choosing which code within a cluster to bill. Columns (1) to (3) in Table 10 report the estimation results on the subsample of codes in clusters with $K = 3, 4, 5$, respectively. Taking $K = 3$ clusters for example, a flagged physician files 369 more codes with intensity level 2, and 128 codes with intensity level 3 than an unflagged physician with identical observables. The same is true for other code clusters, although the estimates are much noisier for those in $K = 4$ clusters due to a small sample size. These results show that codes chosen by flagged physicians strongly tilt toward higher intensities to an extent that is hard to explain by service specialization or exposure to different markets.

In Column (4) of Table 10, we pool codes from all clusters and re-classify the intensities to three levels, *low*, *middle*, and *high* as specified in Table 11. The regression results show that flagged physicians tend to file more mid-intensity codes than their unflagged peers, but not so much for high-intensity codes. This seems to contradict the hypothesis that financial incentives affect the coding decisions of flagged physicians. However, we find very different patterns when we run the baseline regression separately on the two subsamples where the marginal increase in Work RVU between two adjacent levels is *below* average (Column (5)) and *above* average (Column (6)). For example, suppose that Work RVUs increase by 100% per intensity level on average, then codes 99201 through 99205 enter the regression in Column (5) instead of (6) because the Work RVU increase between any two adjacent intensities in that code cluster is less than 100%. We find that flagged physicians do not tend to file more higher-intensity codes with below average Work RVU increments; but they do so for codes with above average Work RVU increments. This shows that flagged physicians do not simply over-file all codes with higher intensities. Instead, data suggest that the coding patterns are consistent with a hypothesis that flagged physicians respond to financial incentives – recall that work RVUs are closely related to Medicare reimbursements. In particular, for codes where the marginal gain in revenue from "upcoding" is relatively low, flagged physicians actually file fewer mid- and high-intensity codes than their peers; but they do file more mid- and high-intensity codes when the marginal gain in revenue from "upcoding" is relatively high.

Note that the regression analysis above focuses on upcoding within service clusters. There could potentially be other ways of overbilling that are not captured in these results. Physicians could bill for more of a *given* service code than it was actually provided, regardless of its intensity; or could upcode *across* code clusters by misreporting the *type of service* provided (e.g., office visits of new

Original intensity			Reclassified
$K = 3$	$K = 4$	$K = 5$	intensity
1	1	1 or 2	Low
2	2 or 3	3	Middle
3	4	4 or 5	High

Table 11: Reclassification of code intensities

NOTES: This table shows how the original code intensities (shown in columns "$K = 3$," "$K = 4$," and "$K = 5$") are reclassified into three levels, *low*, *middle*, and *high*.

patients, which are paid more vs. office visits of established patients, which are paid less). There is no shortage of such overbilling practices according to Department of Health and Human Services (2015). For this reason, the regression results above are likely to be lower-bound estimates of the extent to which flagged physician are potentially overbilling.

5 Supplemental Results from External Data

Before concluding this paper, we corroborate our findings using two external datasets. Doing so both serves as a sanity check for our approach toward detecting potential Medicare overbilling, and at the same time points to possible directions in which our approach may improve existing ones.

5.1 Physician Working Patterns in the National Ambulatory Medical Care Survey (NAMCS) Data

The NAMCS by the National Center for Health Statistics, Centers for Disease Control and Prevention, provides a nationally representative sample of office-based physicians. Each sampled physician is randomly assigned a week for which detailed visit-level data are collected. Given that the CMS Medicare Part B FFS sample covers the vast majority of physicians in the U.S., the two should be fairly comparable.[17]

We first examine the self-reported fraction of Medicare services by physicians sampled in NAMCS. In Table 12, "Medicare patients" are those whose primary payer is Medicare. "% Medicare patients" is the weighted average of the fractions of such patients among all the sampled patients of the reporting physicians. "% Medicare services" and "% time spent with Medicare patients" are defined similarly using the number of services and recorded time physicians spent with Medicare patients. "% revenue from Medicare" is imputed from a categorical variable describing the fraction of the reporting physician's revenue from Medicare payments, with the categories being 0-25, 26-50, 51-75, and 76-100 percents. The figure reported in this table is calculated under the assumption

[17] We discuss the comparability of the two samples in more detail in the Appendix.

that the actual fractions are uniformly distributed within each bin.

	NAMCS
% Medicare patients	0.259
	(0.006)
% Medicare services	0.260
	(0.006)
% time spent with Medicare patients	0.261
	(0.006)
% revenue from Medicare	0.297
	(0.005)
No. of Unique Physicians	3,583

Table 12: Share of Medicare services
NOTES: Standard errors of the mean estimator are reported in parentheses.

Next we look into physician hours worked. Ideally, we would like to use NAMCS to calculate the self-reported total hours worked in the sampled week for each surveyed physician, who are asked to document the time they spent on patients during each visit. However, NAMCS does not sample all the visits within the chosen week, and the sampling rate varies from 100% for very small practices and 10% for very large practices. Because we don't have information on practice sizes (except for whether the physicians work in a solo or group practice), we cannot infer the total number of visits from the NAMCS sample. That said, the *maximum sampled number of hours* spent on Medicare patients per week is 15.17 for physicians working in a solo practice (all of whose visits are sampled), and 49.82 for those working in group practices (the size of which is unknown).

The above comparison shows that Medicare services typically account for about one-third of a physician's entire workload, and take far fewer hours than the 100-hour threshold we used to flag potential overbilling (at least for physicians working in solo practices). This again supports our view that the approach we develop to flag physicians for potential overbilling is likely to be conservative.

5.2 Comparison with the Comprehensive Error Rate Testing (CERT) Program

Finally, we relate the HCPCS code CFIs that we constructed in Section 4.3 with the findings from the CMS CERT program. CERT draws a "statistically valid random sample of claims" (about 50,000) every year, requests documentation from the filing providers, and hires medical review professionals to determine whether payments to these claims are proper or not based on their documentation. The reviewers can disapprove improper payments to claims that have insufficient documentation, questionable medical necessity, incorrect coding, or for other reasons. The disapproval rate for Medicare Part B claims, calculated as the percentage disapproved in all sampled Part B claims, is 18.93% in 2012, and CMS reports Part B improper payment rate - the percentage

of Medicare *dollars* paid incorrectly - for 2012 to be 12.1% (see Centers for Medicare and Medicaid Services (2015)).[18]

Applying the same idea as the Code Flag Index described in Section 4.3, we can use the prevalence of a code j among the disapproved claims and its prevalence among the approved claims to construct a *Code Disapproval Index* (CDI) for the 1,621 HCPCS codes reviewed by CERT:

$$\text{CDI}_j = \frac{100 \times \Pr\left(j|\text{disapproved claims}\right)}{\Pr\left(j|\text{disapproved claims}\right) + \Pr\left(j|\text{approved claims}\right)}, \tag{7}$$

where $\Pr\left(j|\text{disapproved claims}\right)$ is the fraction of claims for HCPCS code j among all disapproved claims; and $\Pr\left(j|\text{approved claims}\right)$ is the faction of claims for HCPCS code j among all approved claims. The CDI also ranges from 0 to 100. If HCPCS code j appears only among disapproved claims, then its CDI takes on a value of 100. If a code only appears among approved claims, then its CDI is equal to 0.

We should note that the CDI as calculated from the pools of approved and disapproved claims in CERT data is *not directly comparable* to the CFI we calculated in Section 4.3. CDI is based on the prevalence of a code j among the disapproved claims relative to its prevalence among the approved claims, while CFI is based on the prevalence of a code j among the flagged physicians relative to its prevalence among the unflagged physicians. Note that we flag physicians based on whether the hours worked implied by their claims are implausibly long, but we do *not* take a stand on whether or not any particular claim is suspicious. On the other hand, CERT program is examining whether particular claims are legitimate, but does not take into account of the overall billings of the physicians. Thus we believe that the CFI we construct and CDI calculated from the CERT data are complementary.

In Figure 6 we compare CDIs and CFIs for the 1,621 HCPCS codes that appear in the CERT data, where each HCPCS code is represented by a circle, with the radius proportional to its total Medicare reimbursement in our CMS sample.[19] Codes falling into the southwest quarter of the plane have both CFIs and CDIs below 50. On the other hand, codes in the northeast quarter are those with both indices above 50, and are thus more frequently associated with (potentially) inappropriate billing practice. These two quarters are cases where our flagging approach and CERT review results agree. On the other hand, codes in the southeast quarter are those with high CDIs from CERT but low CFIs in our sample. Similarly, codes in the northwest quarter are those with

[18] "Improper payment" defined by CMS includes both overpayment and underpayment. The latter contributes only 0.2 percentage points to the 12.1% improper payment rate as reported by Centers for Medicare and Medicaid Services (2015).

[19] The HCPCS codes not reviewed by CERT make up 6.2% of total reimbursement in our CMS sample.

high CFIs but low CERT CDIs. The overall unweighted correlation between the two indices is 0.1257.

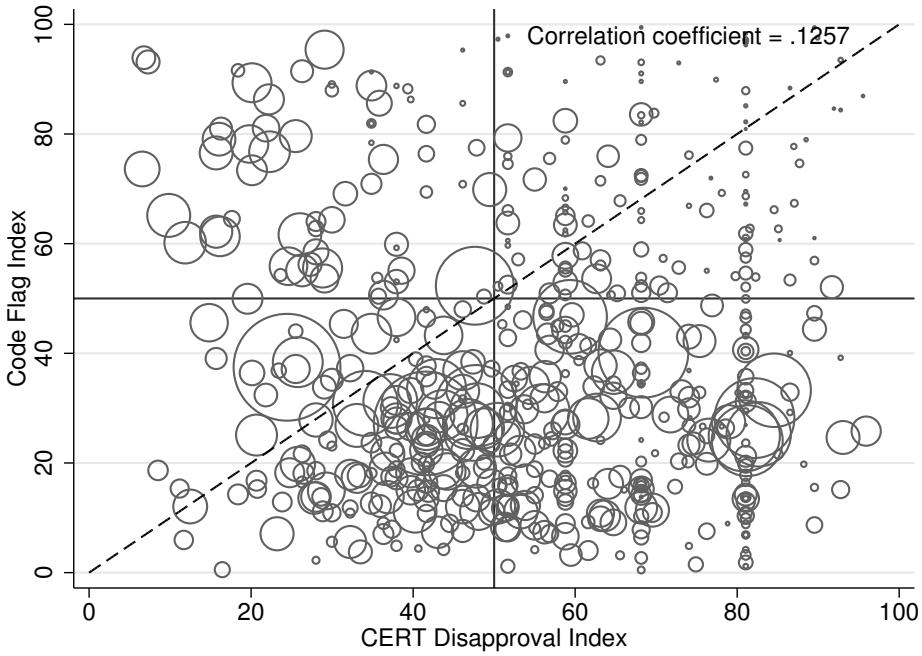


Figure 6: HCPCS Code Flag Index (CFI) and CERT Code Disapproval Index

NOTES: The horizontal axis shows the CERT Code Disapproval Index, defined in Equation (7). The vertical axis shows the CFI, defined in Equation (3). We restrict the sample to HCPCS codes filed by physicians billing at least 20 hours per week in at least one year and sampled in CERT. The graph has 1621 codes in total. Each circle represents a code, with the radius proportional to total Medicare reimbursement. The dashed line represents cases where the two indices are equal (i.e. a "45-degree" line). The solid horizontal and vertical lines show indices of 50.

The comparison between CDI and CFI suggests that our approach to construct Code Flag Index based on flagged and unflagged physicians could potentially contribute to existing auditing methods. Re-formulating sampling strategies to focus more on HCPCS codes with high CFIs, especially those where CFI and CERT CDI differ substantially, may help better detect and deter inappropriate billing with limited regulatory resources. The CFI we construct can help to screen codes that are more likely to be associated with potential overbilling and overbilling.

6 Conclusion

In this paper, we propose and implement a novel approach to detect potential overbilling in Medicare reimbursement based on the simple idea that all physicians have a fixed time budget in a given period (a calendar year for example) and the services claimed for reimbursement require time to complete. We construct the implied hours worked at the individual physician level based on

service codes submitted to Medicare. We flag physicians as potentially overbilling based on whether the implied hours worked are implausibly long. Our method for detecting potential overbilling has at least three advantages relative to the existing methods. First, it imposes minimal data requirements, and is easy to implement, automate, and update over time. Second, it mitigates the impact of confounding factors in the detection of overbilling such as selection and physician heterogeneity, because all physicians face the same time constraint regardless of their patient pool or practice patterns. Finally, it allows users to freely choose the level of stringency when flagging potential overbilling. For example, one could use a different threshold of weekly hours worked or, if higher frequency data are available, 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. Interestingly, we also find suggestive evidence that the coding patterns of the flagged physicians seem to be responsive to financial incentives: within code clusters with different levels of service intensity, they tend to submit more higher-intensity service codes than unflagged physicians; moreover, they are more likely to do so if the marginal gain revenue from submitting mid- or high-intensity codes is relatively high.

Our approach provides a quick and easy tool for detecting potential overbilling, but we would like to emphasize that it does not provide definite evidence for fraudulent coding, nor does it substitute existing methods based on auditing. Rather, we view our approach as a useful screening tool to identify individual physicians, specialties, or HCPCS codes whose billing patterns are highly consistent with overbilling and are hard to reconcile using observables. For example, the HCPCS code CFIs suggest that certain codes are disproportionately associated with flagged physicians, and thus may need more auditing attention. This can help improve the efficiency in the allocation of limited regulatory resources.

References

Adams, Diane L, Helen Norman, and Valentine J Burroughs, "Addressing Medical Coding and Billing Part II: A Strategy For Achieving Compliance. A Risk Management Approach for Reducing Coding and Billing Errors," *Journal of the National Medical Association*, June 2002, *94* (6), 430–447.

Adler-Milstein, Julia and Ashish K. Jha, "No Evidence Found That Hospitals Are Using New Electronic Health Records To Increase Medicare Reimbursements," *Health Affairs*, 2014, *33* (7), 1271–1277.

Angeles, January and Edwin Park, "'Upcoding' Problem Exacerbates Overpayments to Medicare Advantage Plans," *Center on Budget and Policy Priorities*, 2009.

Bastani, Hamsa, Joel Goh, and Mohsen Bayati, "Evidence of Strategic Behavior in Medicare Claims Reporting," *Available at SSRN 2630454*, 2015.

Becker, David, Daniel Kessler, and Mark McClellan, "Detecting Medicare Abuse," *Journal of Health Economics*, 2005, *24* (1), 189 – 210.

Bergman, Jonathan, Christopher S Saigal, and Mark S Litwin, "Service Intensity and Physician Income: Conclusions From Medicare's Physician Data Release," *JAMA Internal Medicine*, 2014, *175(2)*, 297–299.

Bowblis, John R. and Christopher S. Brunt, "Medicare Skilled Nursing Facility Reimbursement and Upcoding," *Health Economics*, 2014, *23*, 821–840.

Brunt, Christopher S., "CPT Fee Differentials and Visit Upcoding under Medicare Part B," *Health Economics*, 2011, *20*, 831–841.

Centers for Medicare and Medicaid Services, "Medicare Carriers Manual Part 3 - Claims Process, Transmittal 1764," August 2002.

_ , "Medicare Learning Network Payment System Fact Sheet Series: Medicare Physician Fee Schedule," http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/medcrephysfeeschedfctsht.pdf April 2013.

_ , "Medicare Learning Network: How to Use the Searchable Medicare Physician Fee Schedule (MPFS)," http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/How_to_MPFS_Booklet_ICN901344.pdf April 2014.

_ , "Medicare Fee-for-Service 2014 Improper Payments Report," July 2015.

Clair, Brandon and Parul Goyal, "What Does Medicare Pay Rhinologists? An Analysis of Medicare Payment Data," *International Forum of Allergy & Rhinology*, 2015, *5* (6), 481–486.

Congressional Budget Office, "The 2014 Long-Term Budget Outlook," July 2014.

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," *NBER Working Paper No. w19320*, 2013.

Dafny, Leemore and David Dranove, "Regulatory Exploitation and Management Changes: Upcoding in the Hospital Industry," *Journal of Law and Economics*, 2009, *52* (2), pp. 223–250.

Department of Health and Human Services, "HIPPA Administrative Simplification: Standard Unique Health Identifier for Health Care Providers; Final Rule," 2004.

_ , "A Roadmap for New Physicians: Avoiding Medicare and Medicaid Fraud and Abuse," 2015.

Dusetzina, Stacie B., Ethan Basch, and Nancy L. Keating, "For Uninsured Cancer Patients, Outpatient Charges Can Be Costly, Putting Treatments Out of Reach," *Health Affairs*, 2015, *34* (4), 584–591.

Fryer, Roland G. and Steven D. Levitt, "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 2004, *119* (3), 767–805.

Gabbert, W, KH Kachur, and T Whitehead, *Current Procedural Coding Expert*, Salt Lake City, UT: Ingenix, 2012.

Geruso, Michael and Timothy Layton, "Upcoding: Evidence from Medicare on Squishy Risk Adjustment," *NBER Working Paper w21222*, 2015.

Gustafson, Shanna L, Gail Pfeiffer, and Charis Eng, "A Large Health System's Approach to Utilization of the Genetic Counselor CPT® 96040 code," *Genetics in Medicine*, December 2011, *13* (12), 1011–1014.

Harewood, Gavin C., Gary Foley, and Zarah Farnes, "Pricing Practices of Gastroenterologists in New York," *Clinical Gastroenterology and Hepatology*, 2014, *12* (11), 1953 – 1955.

Henry J. Kaiser Family Foundation, "The Facts on Medicare Spending and Financing," 2015.

Ip, Ivan K., Ali S. Raja, Steven E. Seltzer, Atul A. Gawande, Karen E. Joynt, and Ramin Khorasani, "Use of Public Data to Target Variation in Providers Use of CT and MR Imaging among Medicare Beneficiaries," *Radiology*, 2015, *275* (3), 718–724. PMID: 25658040.

Jones, Lyell K., Karolina Craft, and Joseph V. Fritz, "Medicare Payment Transparency: Implications for Neurologists," *Neurology: Clinical Practice*, 2015.

Ko, Joan S., Heather Chalfin, Bruce J. Trock, Zhaoyong Feng, Elizabeth Humphreys, Sung-Woo Park, H. Ballentine Carter, Kevin D. Frick, and Misop Han, "Variability in Medicare Utilization and Payment Among Urologists," *Urology*, 2015, *85* (5), 1045 – 1051.

Lapps, Joshua, Bradley Flansbaum, Luci Leykum, Josh Boswell, and Luigi Haines, "Updating Threshold-Based Identification of Hospitalists in 2012 Medicare Pay Data," *Journal of Hospital Medicine*, 2016, *11* (1), 45–47.

Lorence, Daniel P. and Amanda Spink, "Regional Variation in Medical Systems Data: Influences on Upcoding," *Journal of Medical Systems*, 2002, *26* (5), 369–381.

Menger, Richard P, Michael E Wolf, Sunil Kukreja, Anthony Sin, and Anil Nanda, "Medicare Payment Data for Spine Reimbursement: Important but Flawed Data for Evaluating Utilization of Resources," *Surgical Neurology International*, June 2015, *6* (Suppl 14), S391–S397.

O'Gara, Patrick T., "Caution Advised: Medicare's Physician-Payment Data Release," *New England Journal of Medicine*, 2014, *371* (2), 101–103. PMID: 24869596.

Pham, Hoangmai H., Deborah Schrag, Ann S. O'Malley, Beny Wu, and Peter B. Bach, "Care Patterns in Medicare and Their Implications for Pay for Performance," *New England Journal of Medicine*, 2007, *356* (11), 1130–1139. PMID: 17360991.

Rosenberg, Marjorie A., Dennis G. Fryback, and David A. Katz, "A Statistical Model to Detect DRG Upcoding," *Health Services and Outcomes Research Methodology*, 2000, *1* (3-4), 233–252.

Schmajuk, G, Bozic KJ, and Yazdany J, "Using Medicare Data to Understand Low-Value Health Care: The Case of Intra-Articular Hyaluronic Acid Injections," *JAMA Internal Medicine*, 2014, *174* (10), 1702–1704.

Silverman, Elaine and Jonathan Skinner, "Medicare Upcoding and Hospital Ownership," *Journal of Health Economics*, 2004, *23* (2), 369 – 389.

Skolarus, Lesli E., James F. Burke, Brian C. Callaghan, Amanda Becker, and Kevin A. Kerber, "Medicare Payments to the Neurology Workforce in 2012," *Neurology*, 2015, *84* (17), 1796–1802.

Skolasky, Richard L and Lee H Riley III, "Medicare Charges and Payments for Cervical Spine Surgery: Association with Hospital Characteristics.," *Spine*, 2015.

Sutphin, P. D., A. Ding, S. Toomay, S. P. Reis, A. K. Pillai, S. P. Kalva, and S. L. Hsu, "Interventional Radiologist Billing Practices to Medicare: Procedures Performed and Geographic Variation," *Journal of Vascular and Interventional Radiology*, 2014, *26* (2), S173–S174.

The Physicians Foundation, "A Survey of America's Physicians: Practice Patterns and Perspectives," 2012.

Welch, W Pete, Sally C Stearns, Alison E Cuellar, and Andrew B Bindman, "Use of Hospitalists by Medicare Beneficiaries: A National Picture," *Medicare & Medicaid Research Review*, 2014, *4* (2).

Wennberg, John E and Megan M Cooper, *The Dartmouth Atlas of Health Care*, American Hospital Publishing Chicago, IL, 1996.

Wolman, Dianne Miller, Michael ME Johns, Cheryl Ulmer et al., *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety*, National Academies Press, 2009.

Zuckerman, Stephen, 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," Centers for Medicare & Medicaid Services, 2014.

Appendix

A Multiple Physicians Billing Under the Same NPI

In our analysis, it is important that all claims under the same NPI are services furnished by the same individual. This should be the case per the request of NPI-related regulations. NPI was introduced in 2005 to improve the administration of Medicare, Medicaid, and other health programs, especially to facilitate electronic data transmission. According to the NPI Final Rule by the Department of Health and Human Services, NPIs are only assigned to "individuals and entities that are licensed and do furnish health care," and stay unchanged in most cases.[20] NPIs with "entity type code" of 1 are individual human beings ("individuals"), and those with "entity type code" of 2 are organizational providers ("organizations"), such as hospitals, clinics, and nursing homes. Individual providers who are members of an organization and the organization they are affiliated with need to have separate NPI numbers (Department of Health and Human Services (2004)). In addition, the NPI Final Rule also requires that "[providers], according to Federal statute and regulations, must be issued their own identification numbers in order to bill and receive payments from Medicare." Hence the providers have to bill for their own, or have a billing agency do it on their behalf, but cannot bill under other providers' NPIs. Because of its many advantages, NPIs are commonly used in scholarly articles to track physician activity (Gustafson et al. (2011); Welch et al. (2014)). One paper using the unique physician identification numbers (UPINs), which were established before NPIs, acknowledged that "in some cases, different physicians and loosely affiliated practices bill under the same identifier," and that the new NPIs would have avoided this problem (Pham et al. (2007)).

We are confident that in the vast majority of cases the claims filed under the same NPI are from the same provider. However there may be exceptions to the above rules. In cases where an auxiliary personnel furnished an "incident to" service following CMS guidelines,[21] the auxiliary personnel may bill under the NPI of the physician who sets the plan of care (POC). However, these exceptions have minimal influence. This is because (a) CMS guidelines for "incident to" services require that they must be furnished "under the [billing] physician's direct supervision," which means the billing physician must be in the same designated office area, and *immediately* available to provide assistance and direction. This indicates that the physician is spending almost

[20] An NPI is "a permanent identifier, assigned for life, unless circumstances justify deactivation, such as a health care provider who finds that his or her NPI has been used fraudulently by another entity" (see Department of Health and Human Services (2004)).

[21] CMS defines "incident to" services as "those services that are furnished incident to physician professional services in the physician's office (whether located in a separate office suite or within an institution) or in a patient's home" (see Centers for Medicare and Medicaid Services (2002)).

	NAMCS	CMS
1(MD)	0.941	0.940
	(0.005)	(000)
1(solo practice)	0.367	0.307
	(0.012)	(0.001)
1(in IT incentive program)	0.423	0.543
	(0.012)	(0.001)
No. of Unique Physicians	3,583	472,110

Table A1: NAMCS and CMS physician characteristics
NOTES: All NAMCS-related statistics are weighted. Standard errors of the mean estimator are reported in parentheses.

the same amount of time as the case where herself furnishes the service; (b) the place-of-service for these "incident to" services are restricted to non-facilities, which only account for part of our data.

B Comparability of the Main Sample and the NAMCS Sample

Table A1 shows the balancing test results between the NAMCS 2012 data and our main sample constructed from the CMS Medicare Part B FFS Physician Utilization and Payment Data. A few things are done to ensure the comparison between the two are sensible. NAMCS restricts its sample to Doctors of Medicine (MD) and Doctors of Osteopathy (DO). The CMS sample is thus also restricted to include only those with a MD or DO. A tiny fraction of physicians are both MD and DO (59 in total) and they are counted as DOs for calculations in this table. "Solo practice" in NAMCS questionnaires is not explicitly defined. Thus 1(solo practice) in the CMS sample is defined as having no more than 5 providers (including nurses and physician assistants, etc.). NAMCS only asks the sampled physician whether the (group) practice they belong to has *applied* for CMS incentive programs encouraging effective use of health IT. Thus 1(in IT incentive program) in CMS is defined accordingly as a dummy variable for participation in *any* of the incentive programs.

Online Appendix

Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked

Hanming Fang* Qing Gong†

March 7, 2016

In Section A of this online appendix, we show that our results are not driven by residents who may be working long hours. In Section B we show that all of our results are robust to using 112 or 168 weekly hours as the flagging thresholds.

A Are We Flagging Mostly Residents?

Residents are known to working long and continuous hours; and most regulations in recent years restrict resident working hours to no more than 80 hours per week averaged over four weeks (see Wolman et al. (2009)). Therefore extra caution must be exercised when "flagging" residents, for whom it can be perfectly normal (unfortunately) to have extremely long hours. For this reason, we only include in our main sample physicians at least one year out of medical school (i.e. those graduated in or before 2011). However, residency can range from one to seven years depending on the specialty, so it is still possible that some of the physicians graduated in more recent years are residents.

In order to check that we are not flagging mostly residents, we first identify possible residents in our sample, and see how many of them are flagged. We mark physician i of specialty s as a possible resident if i graduated in or after the year $(2012 - T_s^R)$, where T_s^R is one year plus the typical length of residency for specialty s. For example, the typical residency for family practice is 3 years, therefore we mark family practice physicians as possible residents if they graduated

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, and the NBER. E-mail: hanming.fang@econ.upenn.edu.

†Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. E-mail: qinggong@sas.upenn.edu.

Hours threshold	80+		100+		112+		168+	
Year	2012	2013	2012	2013	2012	2013	2012	2013
Number of physicians flagged	4125	3838	2292	2120	1689	1546	615	530
Number of possible residents	16	16	11	6	9	5	2	3
Possible residents/flagged (%)	0.388	0.417	0.480	0.283	0.533	0.323	0.325	0.566

Table A1: Number and fraction of possible residents flagged

NOTES: The table reports the number and fraction of possible residents flagged in years 2012 and 2013. Possible residents are identified by their year of graduation from medical school. Physician i of specialty s as a possible resident if i graduated in or after $2012 - T_s^R$, where T_s^R is one year plus the typical length of residency for specialty s.

in or after 2008. We are adding one year in order identify all possible residents - there could be variations in residency lengths, and some institutions require an extra year after residency to focus on research. Moreover, we use 2012 in the formula for finding possible residents so that the remainder of physicians are not residents in *any* year in our sample, which covers both 2012 and 2013.

Table A1 summarizes the number of possible residents flagged in 2012 and 2013 under varying thresholds. Only a handful of flagged physicians are possible residents, ranging from 16 under the 80-hour threshold to only 2 or 3 under the 168-hour threshold. These possible residents make up less than 0.6% of all flagged physicians. This shows that it is not the residents with long hours that are driving our results.

B Robustness of Results to Alternative Flagging Thresholds

In this section we show the robustness of our main results to the choice of flagging thresholds. One might be concerned, despite our deliberately conservative estimates of hours worked, that the 100-hour-per-week threshold might have caught physicians whose billing truthfully reflects the services they provide. Now we use the two higher thresholds, 112 and 168 hours per week, and show that the main results persist under these thresholds.

B.1 Who Reported Implausibly Long Hours?

Tables B2 and B3 are counterparts to Table 5 of in our paper, except that they use the two alternative thresholds, respectively. The flagged physicians are still more likely to be males, less likely to have a MD, slightly more experienced, work in much smaller group practices and have fewer hospital affiliations. These results are highly similar to those obtained using the 100-hour threshold in terms of sign, magnitude, and the level of statistical significance.

	(1) Ever	(2) 2012	(3) 2013	(4) 2012 only	(5) Both	(6) 2013 only	Mean of Never
1(male)	0.026***	0.031***	0.034***	0.000	0.045***	-0.001	0.856
	[0.008]	[0.009]	[0.009]	[0.014]	[0.010]	[0.020]	
1(MD)	-0.224***	-0.217***	-0.165***	-0.393***	-0.137***	-0.260***	0.843
	[0.038]	[0.042]	[0.038]	[0.052]	[0.042]	[0.041]	
Experience (years)	0.784***	1.013***	0.420	1.755**	0.648**	-0.299	24.124
	[0.298]	[0.347]	[0.279]	[0.719]	[0.315]	[0.588]	
# providers in group	-53.668***	-53.271***	-52.214***	-57.653***	-51.186***	-55.331***	88.515
	[5.991]	[6.341]	[6.797]	[10.527]	[7.239]	[10.102]	
# hospital affiliations	-1.622***	-1.596***	-1.563***	-1.803***	-1.504***	-1.759***	2.807
	[0.121]	[0.130]	[0.108]	[0.223]	[0.115]	[0.145]	
1(in Medicare)	0.008	0.009	0.006	0.013	0.007	0.001	0.857
	[0.008]	[0.009]	[0.009]	[0.014]	[0.011]	[0.017]	
1(in ERX)	-0.020	-0.011	0.011	-0.113***	0.035	-0.065**	0.463
	[0.018]	[0.020]	[0.020]	[0.026]	[0.022]	[0.029]	
1(in PQRS)	-0.006	-0.007	0.020	-0.081***	0.026	0.001	0.396
	[0.018]	[0.020]	[0.020]	[0.028]	[0.021]	[0.031]	
1(in EHR)	-0.030*	-0.030	-0.044**	0.008	-0.050**	-0.026	0.417
	[0.017]	[0.019]	[0.018]	[0.026]	[0.021]	[0.028]	
Types of codes 2012	-0.207	0.478	1.230	-4.419***	2.708**	-3.631***	22.443
	[1.086]	[1.225]	[1.074]	[1.583]	[1.236]	[1.231]	
Types of codes 2013	-0.101	0.348	1.794*	-5.590***	3.069**	-2.414*	22.366
	[1.086]	[1.214]	[1.086]	[1.418]	[1.240]	[1.277]	
Types of E/M codes 2012	-2.177***	-2.149***	-2.168***	-2.223***	-2.123***	-2.326***	6.223
	[0.183]	[0.192]	[0.181]	[0.314]	[0.190]	[0.233]	
Types of E/M codes 2013	-2.126***	-2.122***	-2.056***	-2.343***	-2.026***	-2.165***	6.201
	[0.183]	[0.191]	[0.182]	[0.303]	[0.190]	[0.243]	
N	2,085	1,689	1,546	539	1,150	396	93,948

Table B2: Characteristics of flagged physicians (threshold being 112 hours/week) 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 using the physician characteristic in the corresponding row as the dependent variable, and the flag status dummy (defined by the heading of the column) as the explanatory variable together with HRR fixed effects. Standard errors clustered at the HRR level are in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. "All" refers to all physicians in this sample. "Never" refers to physicians never flagged in any year. "Ever" refers to those flagged in at least one year. "2012" and "2013" refer to those flagged in 2012 and 2013, respectively. "2012 (2013) only" refers to those only flagged in 2012 (2013) but not the other year. "Both" refers to those flagged in both years. Physician experience is imputed from the year of graduation. # providers in group refers to the number of providers in the group practice where the billing physician works at. It is 1 if the billing physician does not work in a group practice. The number of hospital affiliations are 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.

	(1) Ever	(2) 2012	(3) 2013	(4) 2012 only	(5) Both	(6) 2013 only	Mean of Never
1(male)	0.031***	0.043***	0.040***	0.00900	0.064***	-0.0230	0.857
	[0.012]	[0.013]	[0.014]	[0.028]	[0.016]	[0.031]	
1(MD)	-0.380***	-0.366***	-0.324***	-0.512***	-0.281***	-0.447***	0.842
	[0.046]	[0.052]	[0.050]	[0.055]	[0.057]	[0.054]	
Experience (years)	1.081*	1.668***	0.151	3.213***	0.740	-1.410	24.13
	[0.573]	[0.635]	[0.490]	[1.190]	[0.559]	[0.902]	
# providers in group	-55.846***	-52.425***	-58.609***	-49.408*	-54.156***	-70.516***	87.67
	[9.297]	[10.052]	[7.906]	[25.371]	[7.355]	[16.062]	
# hospital affiliations	-2.128***	-2.108***	-1.987***	-2.455***	-1.901***	-2.227***	2.790
	[0.133]	[0.147]	[0.123]	[0.211]	[0.136]	[0.154]	
1(in Medicare)	0.00800	0.00400	0.00200	0.0200	-0.00400	0.0220	0.857
	[0.014]	[0.016]	[0.015]	[0.024]	[0.019]	[0.026]	
1(in ERX)	-0.058**	-0.0410	-0.0370	-0.106***	-0.00100	-0.135***	0.463
	[0.026]	[0.028]	[0.029]	[0.040]	[0.031]	[0.050]	
1(in PQRS)	-0.0140	-0.0150	0.00800	-0.0660	0.0140	-0.00800	0.396
	[0.030]	[0.033]	[0.033]	[0.048]	[0.037]	[0.049]	
1(in EHR)	-0.0310	-0.0480	-0.0310	-0.0300	-0.0590	0.0470	0.416
	[0.031]	[0.034]	[0.034]	[0.048]	[0.038]	[0.044]	
Types of codes 2012	-3.386**	-2.806	-1.088	-8.777***	0.718	-6.070***	22.47
	[1.533]	[1.783]	[1.825]	[1.316]	[2.280]	[1.419]	
Types of codes 2013	-2.865*	-2.532	-0.166	-9.167***	1.394	-4.490***	22.39
	[1.561]	[1.798]	[1.874]	[1.199]	[2.320]	[1.572]	
Types of E/M codes 2012	-2.989***	-3.034***	-2.793***	-3.447***	-2.788***	-2.822***	6.202
	[0.170]	[0.181]	[0.177]	[0.254]	[0.189]	[0.279]	
Types of E/M codes 2013	-2.893***	-2.943***	-2.638***	-3.483***	-2.620***	-2.703***	6.180
	[0.166]	[0.175]	[0.171]	[0.230]	[0.181]	[0.268]	
N	763	615	530	233	382	148	95270

Table B3: Characteristics of flagged physicians (threshold being 168 hours/week) vs. unflagged physicians, conditional on Hospital Referral Region (HRR)
NOTES: See notes to Table B2.

B.2 What are the Specialties of Flagged Physicians?

Tables B4 and B5 show that SFIs for the 7 specialties in Table 6 of our paper remain qualitatively unchanged. The 4 specialties that are over-represented among the flagged physicians, optometry, dermatology, ophthalmology and pathology, still have SFIs above 50; the 3 specialties that are under-represented, nephrology, cardiology and internal medicine, still have SFIs below 50, although with slight changes in their rankings. In fact, the discrepancies in SFIs become larger when we use a higher flagging threshold – over-represented specialties get even larger SFIs, and under-represented specialties get even smaller SFIs.

	% in all	Num. unflagged		Num. flagged		SFI	
Specialty\Year		2012	2013	2012	2013	2012	2013
Optometry	1.893	1323	1448	495	370	95.43	93.98
Dermatology	4.185	3661	3638	359	382	84.56	86.52
Ophthalmology	7.960	7379	7383	265	261	66.73	68.36
Pathology	2.746	2585	2587	53	51	53.38	54.65
Nephrology	4.900	4655	4661	51	45	37.96	37.11
Cardiology	11.12	10597	10617	82	62	30.18	26.30
Internal Medicine	11.09	10607	10610	43	40	18.46	18.73
All physicians		94344	94487	1689	1546		

Table B4: Physician specialties and flag status (threshold being 112 hours/week)
NOTES: The table shows seven specialties with the highest SFIs among those with at least 50 flagged physicians. "% in all" shows the fraction 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 (unflagged) physicians by year in our sample.

	% in all	Num. unflagged		Num. flagged		SFI	
Specialty\Year		2012	2013	2012	2013	2012	2013
Optometry	1.893	1,551	1,614	267	204	96.39	95.79
Dermatology	4.185	3,908	3,894	112	126	81.64	85.36
Ophthalmology	7.960	7,551	7,578	93	66	65.65	61.08
Pathology	2.746	2,616	2,617	22	21	56.61	59.12
Nephrology	4.900	4,700	4,698	6	8	16.53	23.48
Internal Medicine	11.09	10,639	10,640	11	10	13.82	14.48
Cardiology	11.12	10,670	10,672	9	7	11.57	10.57
All physicians		95,418	95,503	615	530		

Table B5: Physician specialties and flag status (threshold being 168 hours/week)
NOTES: See notes to Table B4.

B.3 What Codes Do Flagged Physicians Tend to Bill?

Figures B1 and B2 plot the relationship between HCPCS Code Flag Indices (CFIs) and the probability a code is filed by a flagged physician. Not surprisingly, the non-linearity is preserved under alternative flagging thresholds, and becomes stronger when the threshold is higher.

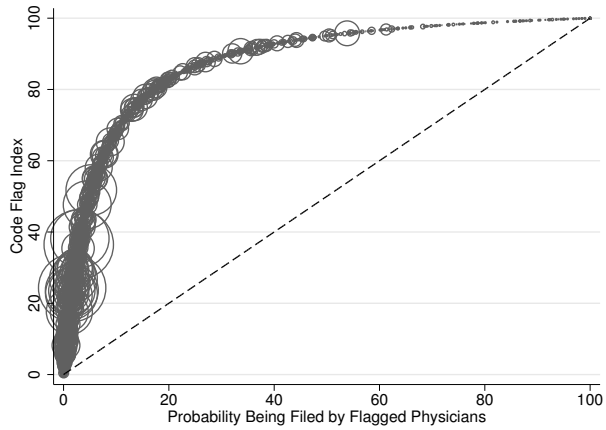


Figure B1: Threshold = 112 hours/week

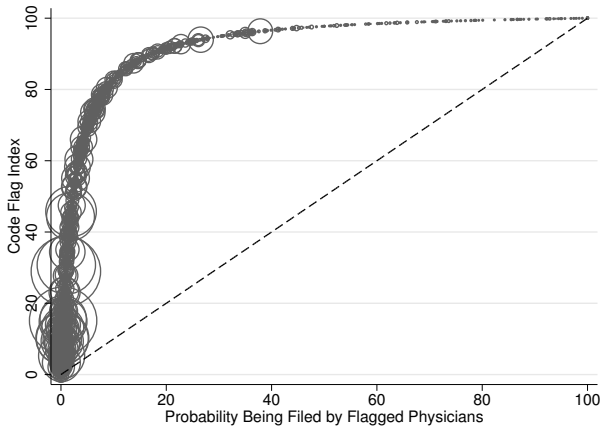


Figure B2: Threshold = 168 hours/week

The Relationship between HCPCS Code Flag Index and its Probability of Being Filed by Flagged Physicians

NOTES: The horizontal axis shows the probability of the HCPCS code being filed by a flagged physician (in %). The vertical axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes filed by physicians billing at least 20 hours per week in at least one year. Each circle represents a HCPCS code, with the radius proportional to total revenue. The dashed line is the "45-degree" line.

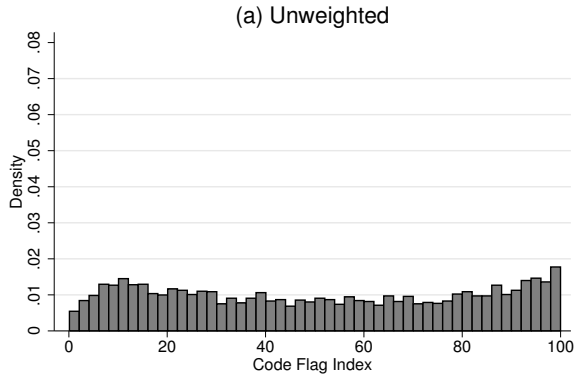
Figure B3 and B4 plot the distribution of CFIs under the two alternative thresholds. We still see that high-SFI codes have disproportionately high shares of reimbursement relative to their volumes.

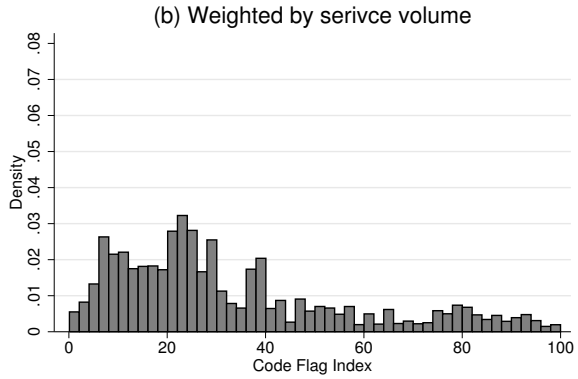
Figures B5 and B6 plot the CFI distributions for codes filed by flagged physicians (solid lines) and by unflagged physicians (dashed lines). Again by construction, flagged physicians file more high-CFI codes. The difference between flagged and unflagged physicians is more dramatic when the higher flagging threshold, 168 hours per week, is used.

B.4 Decomposing the Long Hours and Quantifying Potential Overbilling

Tables B6 and B7 show how the decomposition of services provided by flagged physicians differs from that of unflagged physicians. Just as Table 8 in our paper shows, flagged physicians provide more services and treat more Medicare Part B FFS patients in total; they also provide more services per patient, and tend to choose services of higher intensity; with average per-service revenues only slightly higher than those of unflagged physicians, they end up with substantially lower imputed hourly revenues. Again, the differences are larger under the 168-hour threshold.

Tables B8 and B9 compare the hourly revenues and Overbilling Potential Factors (OPFs) between flagged and unflagged physicians. The results are still highly similar to those in our paper using the 100-hour threshold, both qualitatively and quantitatively. In particular, flagged physicians have very large discrepancies between their reported and predicted hourly revenues; their





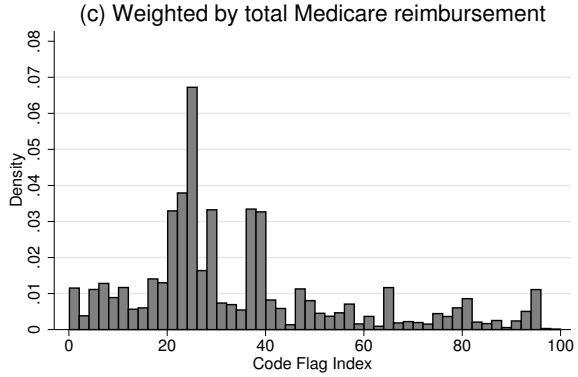
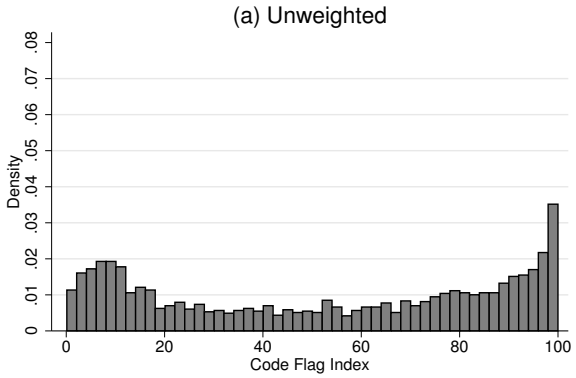
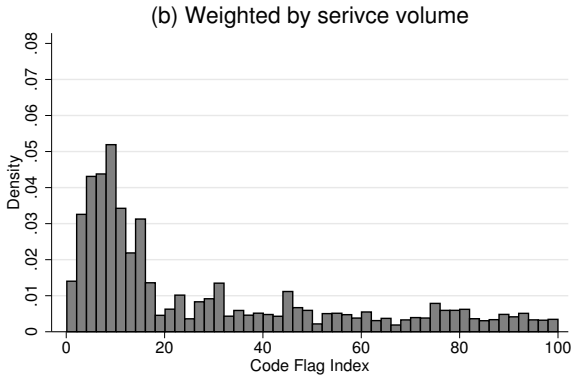


Figure B3: Threshold = 112 hours/week





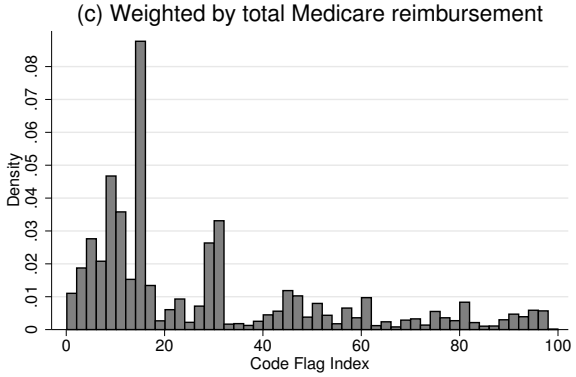


Figure B4: Threshold = 168 hours/week

Distribution of HCPCS Code Flag Index

NOTES: The horizontal axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes with CFIs strictly larger than 0 and strictly smaller than 100. Bandwidth is 2 for all three histograms.

	Flagged		Unflagged	
Year	2012	2013	2012	2013
Num. of services provided	8363.404***	8643.937***	4579	4523
	[1047.754]	[1054.903]		
Num. of services per patient	2.007***	1.459***	2.438	2.380
	[0.301]	[0.228]		
Num. of services provided per hour	-1.457***	-1.459***	2.874	2.891
	[0.106]	[0.099]		
Num. of patients	2714.590***	3128.023***	2444	2436
	[597.842]	[627.702]		
Num. of patients per day	7.417***	8.570***	6.677	6.675
	[1.633]	[1.720]		
Num. of patients per hour	-0.998***	-0.984***	1.572	1.586
	[0.069]	[0.063]		
Medicare payment per service ($)	6.011	11.512***	74.91	73.48
	[4.393]	[4.309]		
Medicare payment per patient ($)	44.109***	47.909***	151.1	146.5
	[8.354]	[8.081]		
Medicare payment per hour ($)	-51.834***	-49.186***	161.9	158.9
	[6.336]	[5.439]		
N	1689	1546	94344	94487

Table B6: Volume of services supplied conditional on Hospital Referral Regions: flagged vs. unflagged physicians (threshold being 112 hours/week)

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) working days in year 2012 (2013). "Per hour" statistics are calculated using the estimated total hours worked of each physician.

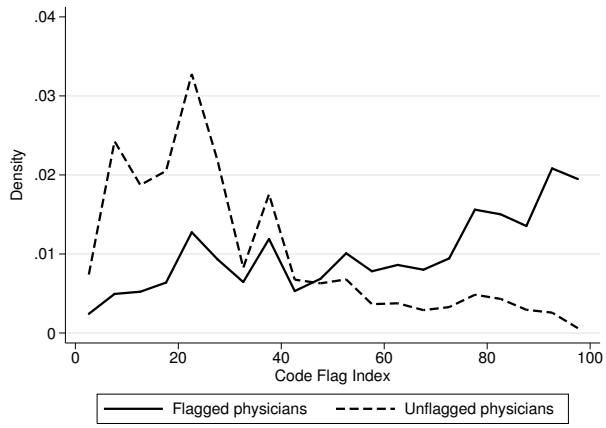


Figure B5: Threshold = 112 hours/week

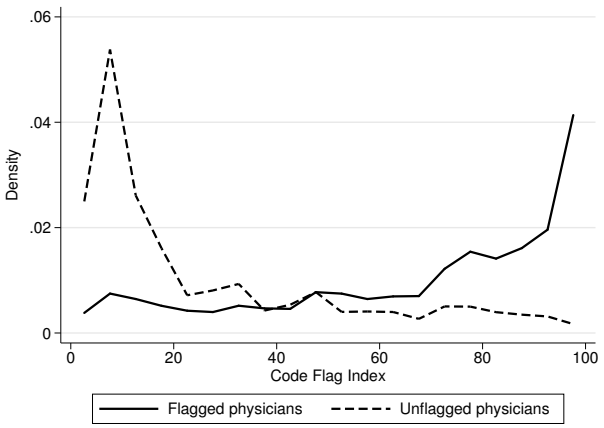


Figure B6: Threshold = 168 hours/week

Distribution of Code Flag Index weighted by service volumes: flagged vs. unflagged physicians

NOTES: The horizontal axis shows the Code Flag Index (CFI). We restrict the sample to HCPCS codes with CFIs strictly larger than 0 and strictly smaller than 100. The solid line shows the distribution of CFIs of codes billed by flagged physicians, and the dashed line shows that for unflagged physicians. Density is weighted by a HCPCS code's total service volume furnished by all physicians.

	Flagged		Unflagged	
Year	2012	2013	2012	2013
Num. of services provided	11979.801***	12488.739***	4654	4596
	[2593.733]	[2728.369]		
Num. of services per patient	3.443***	2.880***	2.453	2.389
	[0.458]	[0.352]		
Num. of services provided per hour	-1.583***	-1.614***	2.861	2.878
	[0.230]	[0.219]		
Num. of patients	3420.234**	4097.479**	2472	2465
	[1485.181]	[1653.190]		
Num. of patients per day	9.345**	11.226**	6.753	6.754
	[4.058]	[4.529]		
Num. of patients per hour	-1.127***	-1.127***	1.563	1.578
	[0.144]	[0.135]		
Medicare payment per service ($)	-13.288***	-7.560	75.06	73.66
	[5.125]	[4.943]		
Medicare payment per patient ($)	26.464**	35.998***	151.7	147.0
	[12.141]	[11.879]		
Medicare payment per hour ($)	-79.931***	-75.784***	161.5	158.6
	[7.233]	[6.974]		
N	615	530	95418	95503

Table B7: Volume of services supplied conditional on Hospital Referral Regions: flagged vs. unflagged physicians (threshold being 168 hours/week)

NOTES: See notes to Table B6.

OPF1, which captures the excess revenue they get relative to their unflagged peers (assuming identical hours worked), is still around 2; and their OPF2, which describes the extent to which they could be over-reporting hours worked (assuming the goal of overbilling is to achieve the same revenue with fewer hours), ranges between 6.178 and 9.805. The results for unflagged physicians also barely change from those reported in the paper.

	Flagged Physicians		Unflagged Physicians	
	2012	2013	2012	2013
Reported hourly revenue ($)	106.909	108.766	158.790	155.861
	(2.418)	(2.357)	(0.243)	(0.241)
Predicted hourly revenue ($)	131.204	134.543	160.605	157.387
	(1.231)	(1.156)	(0.149)	(0.148)
Overbilling Potential Factor 1	1.964	2.031	0.590	0.574
	(0.061)	(0.065)	(0.001)	(0.001)
Overbilling Potential Factor 2	7.347	6.178	1.165	1.143
	(0.277)	(0.270)	(0.005)	(0.004)
N	1,689	1,546	94,344	94,487

Table B8: Hourly revenues and Overbilling Potential Factors (OPFs) (threshold being 112 hours/week)

NOTES: The table compares the hourly revenues and OPFs 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 of the mean estimator are reported in parentheses.

	Flagged Physicians		Unflagged Physicians	
	2012	2013	2012	2013
Reported hourly revenue ($)	74.501	78.803	158.415	155.527
	(3.927)	(3.962)	(0.243)	(0.241)
Predicted hourly revenue ($)	111.203	117.775	160.179	157.028
	(2.140)	(2.083)	(0.150)	(0.148)
Overbilling Potential Factor 1	2.189	2.315	0.606	0.589
	(0.149)	(0.164)	(0.002)	(0.001)
Overbilling Potential Factor 2	9.805	9.155	1.190	1.163
	(0.449)	(0.505)	(0.005)	(0.004)
N	615	530	95,418	95,503

Table B9: Hourly revenues and Overbilling Potential Factors (OPFs) (threshold being 168 hours/week)

NOTES: See notes to Table B8.

Figures B7 and B8 plot the OPFs for both flagged and unflagged physicians using the 112-hour and the 168-hour thresholds, respectively. The distributions of flagged physicians' OPF1 and OPF2 are still shifted rightward relative to the distributions of unflagged physicians. Moreover, Panel (b)'s of both figures also show that many flagged physicians' reported revenues fall below their predicted "fair" hourly revenues, whereas the reverse is true for unflagged physicians.

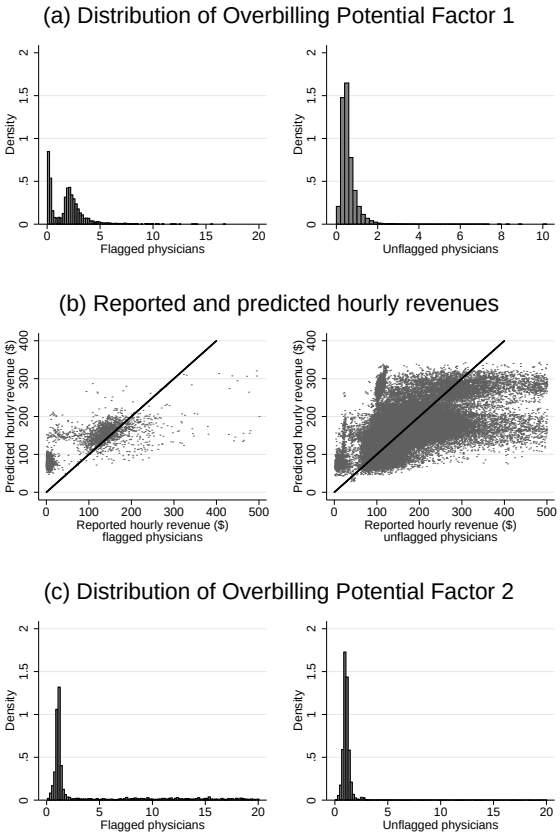


Figure B7: Threshold = 112 hours/week

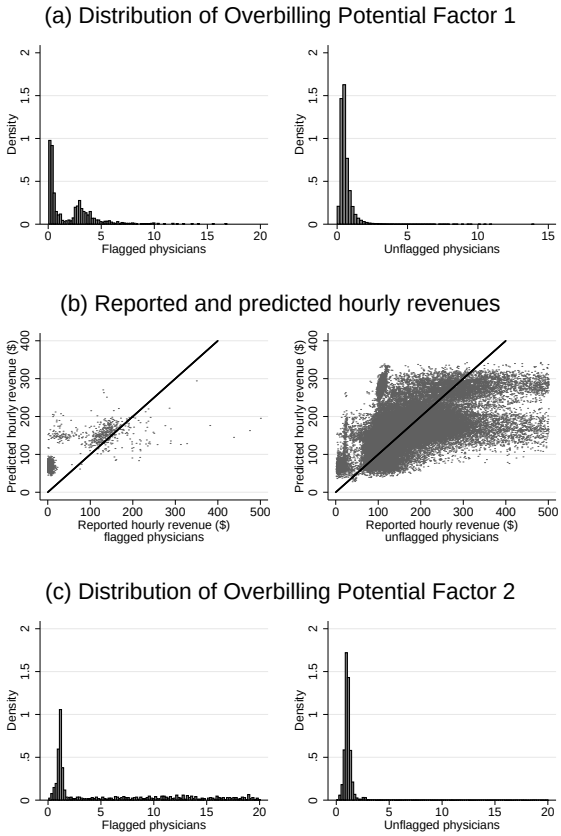


Figure B8: Threshold = 168 hours/week

Overbilling Potential Factors (OPFs)

NOTES: The two figures on the top show the distribution of OPF1 for flagged (left) and unflagged (right) physicians. The two scatter plots in the middle are showing predicted hourly revenues (on the vertical axis, based on OLS regression conditional on physician gender, credential (MD dummy), years of experience, as well as a full set of specialty, HRR, and year fixed effect) against reported hourly revenues (on the horizontal axis). The thick solid line is the "45-degree" line. The two figures on the bottom show the distribution of OPF2 for flagged (left) and unflagged (right) physicians. The bin widths in all four histograms are 0.2.

B.5 Coding Decisions and Fee Differentials

Tables B10 and B11, counterparts to Table 10 of our paper, present the regression results obtained under the two alternative flagging thresholds. All key findings are robust to the choice of thresholds, except that estimates for flag-related variables tend to have larger standard errors. This is because the thresholds used here lead to a much smaller group of flagged physicians, making estimates noisier.

B.6 Comparison with the Comprehensive Error Rate Testing (CERT) Program

Figures B9 and B10 plot the comparison between HCPCS Code Flag Indices (CFIs) that we constructed using the CMS data and Code Disapproval Indices (CDIs) we calculated using CERT auditing results. Under the higher thresholds (112-hour and 168-hour), CFIs become more extreme, which adds to the incomparability of CFIs and CDIs (see discussion in the paper). This naturally reduces the correlation between the two indices.

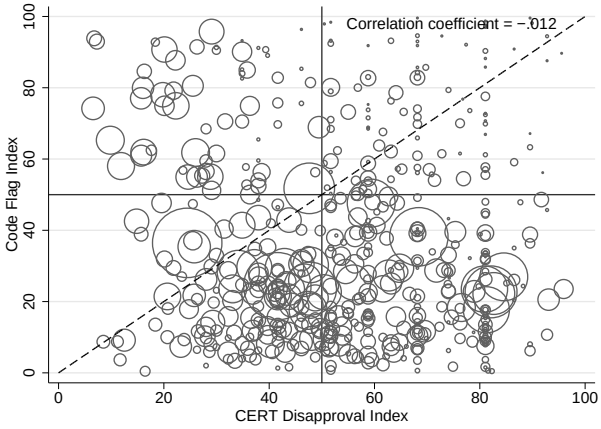


Figure B9: Threshold = 112 hours/week

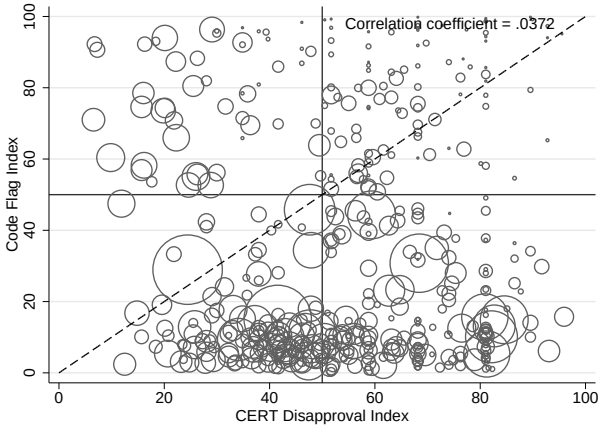


Figure B10: Threshold = 168 hours/week

HCPCS Code Flag Index (CFI) and CERT Code Disapproval Index

NOTES: The horizontal axis shows the CERT Code Disapproval Index. The vertical axis shows the CFI. We restrict the sample to HCPCS codes filed by physicians billing at least 20 hours per week in at least one year and sampled in CERT. The graph has 1621 codes in total. Each circle represents a code, with the radius proportional to total Medicare reimbursement. The dashed line represents cases where the two indices are equal (i.e. a "45-degree" line). The solid horizontal and vertical lines show indices of 50.

References

Wolman, Dianne Miller, Michael ME Johns, Cheryl Ulmer et al., *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety*, National Academies Press, 2009.

	(1)	(2)	(3)	(4)	(5)	(6)
	$K = 3$	$K = 4$	$K = 5$	All K	All K & below average	All K & above average
Flagged	261.5***	653.7*	33.56	178.2***	193.8***	145.9***
	[68.48]	[349.5]	[22.71]	[18.54]	[31.79]	[18.54]
Intensity=2	244.0***	171.9***	12.83***			
	[2.823]	[10.13]	[3.233]			
Intensity=3	130.4***	150.6***	241.2***			
	[2.448]	[10.67]	[3.365]			
Intensity=4		-77.70***	235.9***			
		[10.78]	[3.177]			
Intensity=5			33.69***			
			[3.027]			
Flagged × (intensity=2)	322.2***	-85.40	94.39***			
	[92.27]	[405.6]	[26.17]			
Flagged × (intensity=3)	155.2*	-61.57	151.8***			
	[82.93]	[401.1]	[30.12]			
Flagged × (intensity=4)		395.3	43.17			
		[461.5]	[28.50]			
Flagged × (intensity=5)			14.60			
			[27.22]			
Mid-intensity				240.2***	19.46***	343.9***
				[1.764]	[1.235]	[2.846]
High-intensity				154.0***	34.68***	186.3***
				[1.506]	[1.206]	[2.508]
Flagged × Mid-intensity				83.77***	-86.67***	201.6***
				[26.15]	[31.93]	[36.00]
Flagged × High-intensity				-10.76	-84.58***	71.07**
				[23.46]	[31.37]	[32.40]
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	0.190	0.052	0.171	0.157	0.163	0.079
Observations	399,907	53,521	561,657	1,015,085	508,478	506,607

Table B10: Billing patterns and code intensity level (threshold being 112 hours/week)

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. Furthermore, Columns (1) to (3) are only using the subsamples of code clusters with 3, 4, and 5 levels of intensities, respectively. Columns (4) pool codes in all clusters together, and re-classify intensities to low, middle, and high as specified in our paper. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	(1)	(2)	(3)	(4)	(5)	(6)
	$K = 3$	$K = 4$	$K = 5$	All K	All K & below average	All K & above average
Flagged	413.1	360.9*	-19.39	158.1***	172.7***	140.6***
	[257.9]	[217.4]	[25.07]	[32.40]	[60.80]	[33.86]
Intensity=2	245.5***	172.2***	15.05***			
	[2.876]	[10.32]	[3.233]			
Intensity=3	131.0***	151.4***	243.3***			
	[2.506]	[10.89]	[3.368]			
Intensity=4		-76.17***	236.3***			
		[10.99]	[3.176]			
Intensity=5			33.35***			
			[3.019]			
Flagged × (intensity=2)	372.8	234.7	133.2***			
	[349.1]	[597.3]	[33.06]			
Flagged × (intensity=3)	299.6	46.25	158.5***			
	[329.9]	[386.2]	[41.95]			
Flagged × (intensity=4)		2161.3**	19.62			
		[942.5]	[35.36]			
Flagged × (intensity=5)			73.31*			
			[42.08]			
Mid-intensity				240.5***	16.75***	346.0***
				[1.763]	[1.045]	[2.852]
High-intensity				152.6***	31.62***	186.2***
				[1.499]	[0.993]	[2.513]
Flagged × Mid-intensity				28.56	-43.80	64.26
				[45.57]	[64.89]	[57.12]
Flagged × High-intensity				-34.08	-40.24	-16.95
				[40.33]	[56.88]	[60.56]
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	0.186	0.049	0.170	0.155	0.161	0.077
Observations	399,907	53,521	561,657	1,015,085	508,478	506,607

Table B11: Billing patterns and code intensity level (threshold being 168 hours/week)
NOTES: See notes to Table B10.