Monitoring Pressure and Billing Practices: Evidence from Medicare Recovery Audits

Jianjing Lin*
University of Massachusetts Amherst

Juan Pantano†
University of Arizona and
University of Hong Kong

August 21, 2025

Abstract

Monitoring has been widely used as a tool to ensure accountability and proper use of resources. However, excessive oversight could result in substantial social costs, calling for a scale-back in the level of monitoring. We exploit a dramatic reduction in audit probability and examine how auditees respond to reduced monitoring in the context of Medicare hospital inpatient care. We develop a simple model characterizing hospitals' decision to code admissions intensely and examine how the level of hospitals' improper coding changes in response to variation in monitoring. To test the prediction empirically, we apply a difference-in-differences strategy that leverages differential reductions in audit pressure across diagnostic related groups. We find that hospitals code more aggressively admissions for which auditors lowered audit probability. Our results suggest that for every dollar Medicare saves on operating the program after the scale-back, it may have to suffer a loss of $25, among which 67% comes from hospitals' more intense coding due to less monitoring. It might require more thoughtful policy designs to balance the gain from lowering social costs of unnecessary audits, against potentially increased illicit activity arising from reduced monitoring.

JEL Codes: I13, I18, I11, L88, H51
Keywords: Government monitoring, improper payments, Medicare, hospitals

*Email: jianjinglin@umass.edu.
†Corresponding Author; email: juanpantano@gmail.com. We thank Paul Jacobs, Jeffrey Mccullough, Maggie Shi, Kevin Shih, Lucy Xiaolu Wang, April Wu, and seminar participants at several institutions for their valuable comments. We are very grateful to Maggie Shi for sharing the data on the Medicare Recovery Audit program used in this paper.

1 Introduction

Procurement of goods and services is often characterized by a purchaser contracting with a supplier who has private information about costs and the effort to contribute. The purchaser with limited ability to monitor the supplier usually hires third parties to assess the supplier's compliance with laws, regulations and policies. Typical examples include utilizing independent inspectors to monitor government procurement or hiring auditors to examine insurance claims. The literature evaluating public monitoring programs suggests that excessive oversight could result in substantial social cost and distortion, forcing the purchaser (or the government) to scale back the monitoring program. However, few studies are devoted to understanding how the supplier responds to the reduced scrutiny when the monitoring programs are scaled back, and what would be the potential social impacts. In this paper, we study the supplier's response to a dramatic reduction in audit probability in the context of health care.

Monitoring and auditing has been widely adopted in the healthcare sector to combat waste, fraud, and abuse in health spending. Medicare, the largest public health insurer in the U.S., implemented a variety of auditing strategies to protect the integrity of the program. We focus on hospital inpatient care, the largest contributor to Medicare expenditure, and the Medicare Recovery Audit Program (RAP) that conducts a wide range of post-payment reviews on hospital inpatient claims. Centers for Medicare and Medicaid Services (CMS), the government agency that administers the Medicare program, periodically contracts with recovery audit contractors (RACs), to review claims and ensure responsible spending. The threat of detection by auditors serves as a deterrent for hospitals and other medical providers who might otherwise engage in inappropriate billing practices.

During the first few years since the program was implemented, the RAP imposed substantial audit pressure to hospitals, who then decided to fight back by challenging a high volume of claims with overpayment determinations, resulting in a substantial backlog of appeals, and by conducting an aggressive media and legislative campaign, alleging that recovery audits were overzealous and resulted in substantial administrative burden.[1] To improve program efficiency and reduce hospitals' costs in the recovery audit process, CMS imposed a limit on the proportion of claims for which RACs can submit additional documentation requests (ADR) for review, among all the claims of the hospital paid by Medicare. We call it the *ADR limit*, which is regulated to be *0.5%* by CMS. This effectively reduced the proportion of

[1] Hospitals filed lawsuits against recovery audits. See here for a list of past litigation cases: https://www.aha.org/legal/past-litigation.

claims that could be reviewed by RACs for each hospital. Our empirical strategy, described more fully below, is built around this regulatory change.

Since the threat of detection by RACs mitigates the incentives hospitals have for improper billing, we can interpret this decline in audit rates as a substantial reduction in the detection risk faced by hospitals who consider to improperly bill some of their claims. We exploit the differential relative declines in audit rates across *diagnostic related groups* (*DRGs*)—grouping of similar patients on the basis of a shared disease or procedure—and auditors as exogenous variation in the incentives for inappropriate coding and seek to understand how hospitals respond to the dramatic change in audit probability in their decision to code a given admission intensely.

The results of our analysis have potentially important policy implications. First, if the RAP did pose substantial deadweight loss that mainly arose from the unnecessary interactions between RACs and hospitals (given that the reviewed claims were *legitimate*), we would expect that hospitals' coding intensity would not vary much after the cap given by the ADR limit was imposed. However, if we found evidence that hospitals responded to the change in the ADR limit by coding more aggressively, it might suggest that the scrutiny from RACs could play a role in inhibiting potentially fraudulent billing from healthcare providers.

Moreover, even if the RAP increased hospitals' already-high administrative burdens, considering its potential in curbing improper payments should help policy-makers find the optimal level of oversight on healthcare spending, striking the right balance between deterrence of improper billing and the costs of administrative burden. Second, we focus on the healthcare sector, an ideal and important setting for this topic. Given the nature of information asymmetry between healthcare providers and non-medical professionals, healthcare reimbursements could be a breeding ground for illicit activities (Becker et al., 2005). Monitoring and auditing is used as a common tool to ensure proper billing in health care. Medicare deals with hundreds of billions of dollars in reimbursements every year, and, on average, the reported improper payments account for approximately 8%-10% every year.[2] Even a small share of improper billing or healthcare fraud could result in a large amount of financial losses, which could be ultimately assumed by tax payers (Shekhar et al., 2023). Finally, our analysis and implications could be applied to a broader context, considering that the monitoring mechanism we consider has analogies to that of other settings, such as tax enforcement, government procurement, environmental regulation, and so on.

We develop a simple model where a hospital decides which severity level—represented

[2]See https://www.cms.gov/research-statistics-data-and-systems/monitoring-programs/medicare-ffs-compliance-programs/cert/additionaldata.

by a bill *code*–a patient is assigned to, among patients sharing a common primary disease or procedure. The higher the code a patient is reported with, the higher the reimbursement that the hospital will receive. In particular, the hospital engages in a practice called *upcoding*, if it codes the patient to a higher level than it is medically justified. The hospital decides whether to upcode a patient with a given primary diagnosis or procedure, after accounting for the extra revenue from upcoding and the outcomes of improper coding detection from audits, in the event that an audit on the admission takes place. We derive a comparative static, showing how the hospital's decision to upcode an admission will change as the audit probability varies.

The model predicts that hospitals tend to report more patients with the top code if the audit probability drops. While this prediction is relatively standard and in line with seminal work on the economic of illicit behavior (Becker, 1968; Allingham and Sandmo, 1972), the model is still worth laying out in some detail and eventually becomes more critical to guide some of our heterogeneity analysis where the predictions may not be as straightforward. For instance, the extent to which hospitals upcode in response to different levels of auditing could vary by the incremental financial return from upcoding. One might expect that hospitals code more aggressively among high-revenue codes given less monitoring. However, our model (which is essentially the Becker model with heterogeneity in risk and rewards) formalizes how the responsiveness depends on different competing forces. Fundamentally, hospitals need to tradeoff the financial gains from (improperly) billing more high-revenue codes against the higher probability of getting detected for this behavior due to greater audit pressure on these codes. We believe that this novel insight is of more general interest and applicable to other settings where the change in monitoring may have non-trivial differential impacts according to the stakes at play.

To test the theoretical prediction empirically, we examine the proportion of Medicare hospital discharges reported with the top codes before and after the adjustment of the ADR limit, using a 5% random sample of Medicare hospital inpatient claims and the administrative data from the Medicare RAP.[3] We use a difference-in-differences (DiD) strategy, comparing the share of discharges reported with top codes before and after the regulatory change.

We define the treatment group for the DiD design as the DRG-RAC pairs in which the audit probability dropped after the ADR limit was implemented. While the extent to which audit rates vary after the imposition of ADR limit may be endogeneous, we argue that imposing the ADR limit has a binding effect on most of the RAC-DRG pairs, forcing

[3]We thank Maggie Shi for sharing these data, which is used in Shi (2024). She obtained these data via a Freedom of Information Act request.

a mechanical reduction in the audit rates of them. In principle, a RAC could audit more than 0.5% of claims for a particular DRG or set of DRGs within a hospital as long as it compensated with low audit rates in other DRGs so as to remain under the cap, in total, for that hospital. We expect the DRGs that got most attention from RACs in a hospital, i.e., those with relatively higher audit rates, before the ADR limit was imposed, would experience the most substantial declines in audit rates after the policy change. If a hospital had high audit rates in some DRG(s) but low audit rates in the others, then the audit rates of the DRGs with relatively high audit rates would have to decline substantially post-policy change so that the overall average audit rate for that hospital would be below the cap. Assuming this mechanism works for all the hospitals, the reduction in audit rates at the RAC-DRG level post-policy change relative to pre-policy change—our treatment definition—is plausibly exogenous. To explore whether pre-existing trends in the treatment group are inherently different from those in the control group, we use event studies to assess the evolution of the top-coding share in the two groups of DRG-RAC pairs leading up to the change in the ADR limit. Moreover, to reduce the concerns on the potential bias in the estimated treatment effect from the conventional two-way fixed effects (TWFEs) specification (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; De Chaisemartin and d'Haultfoeuille, 2024; Borusyak et al., 2024), we use a recently-developed DiD method (De Chaisemartin and d'Haultfoeuille, 2024) for our estimation.

We find that hospitals report more patients with top codes among the DRG-RAC pairs that experienced lower audit rates after the policy change, with an average increase of 7.2 percent relative to the control group. The increase in top-coding rate is statistically significant and persists for years during the post-treatment period. In contrast, we do not observe a differential pre-existing trend in the top-coding share between the two groups during the years leading up to the ADR limit. We conduct additional heterogeneity analyses to explore whether the effect we have found concentrates on certain types of disease codes or hospitals. We find that the relative increase in percent top codes post policy change mainly comes from the disease codes for which the extra revenues from top coding is smaller. A possible explanation is that hospitals tend to behave more cautiously in coding cases under greater surveillance. According to our model, the extent to which a hospital improperly codes a patient is determined by multiple forces. Since high-revenue disease codes could be subject to heightened monitoring, which is shown in our data, the empirical finding here suggests that the concern about greater detection risk for high-revenue disease codes dominates and thus, hospitals are more careful in assigning top codes among patients with these diseases.

Prior studies have found that for-profit hospitals or hospitals undergoing greater financial distress tend to code more aggressively to enhance revenues. Our heterogeneity analyses regarding hospital types suggest that hospitals respond to reduced monitoring by increasing their top-coding rate, regardless of their ownership type, location, or financial health status.

Overall, we find an uptick in the rate of top-coding after the imposition of the cap on the number of claims for which RACs can request additional documents for review. A back-of-the-envelope calculation shows that for every dollar Medicare saves on paying RACs after scaling back the program, Medicare might have to suffer a loss of $25, among which 67% arises from hospitals' more aggressive coding due to reduced monitoring. Our finding suggests that the RAP could be effective in combating potential fraud and waste in healthcare spending. Armed with these findings, policy-makers may benefit from exploring alternative policy designs to strike the right balance between lowering social costs from scaling back monitoring against these increases in improper billing due to lax oversight.

Our paper complements to two strands of literature. First, our paper contributes to the broad literature on the role of monitoring as a tool in improving accountability and management of resources. Most of these studies investigate the context in government procurement (Olken, 2007), environmental regulations (Telle, 2013), political corruption (Avis et al., 2018; Finan and Mazzocco, 2021), and tax enforcement (Kleven et al., 2011; DeBacker et al., 2018).[4] Some recent studies started to explore the effect of monitoring in the context of health care (Angerer et al., 2021; Groß et al., 2021; Shi, 2024). Similar to Shi (2024), we complement the literature by evaluating a national audit program in the healthcare setting, a sector that involves large dollar value at stake and in which fraudulent or abusive behavior could be hard to detect.

Many prior studies examining the effect of monitoring programs have found positive effects of these mechanisms, but increased oversight could also result in social costs, such as heavy administrative burdens (Shi, 2024), and agency problems (Duflo et al., 2013; Vannutelli, 2024). Our paper studies how hospitals' coding intensity changes following a large scale-back in the monitoring program due to these concerns and has found that hospitals code relatively more aggressively for bill codes experiencing declines in audit probability. Our findings suggest that policy-makers might face a tradeoff between reduced inefficiency and potential upticks in wasteful spending by scaling back monitoring programs.

Second, our paper also complements the literature that investigates how healthcare providers respond to regulators' non-financial policy changes, such as stricter requirements

[4] Slemrod (2019) provides a comprehensive review of the literature on tax compliance and enforcement.

for documenting care provision and increasing monitoring and detection of fraud and abuse. Sacarny (2018) has found that, after a payment reform in 2007, hospitals could have increased revenues by specifying a patient's heart failure condition but only capture about half of them due to substantial institution-level frictions. Shi (2024) finds that monitoring healthcare providers has reduced the rendering of unnecessary care and also resulted in compliance costs; in response, providers adopt new information technology to identify potentially medical unnecessary care. League (2024) examines the impacts of administrative burdens on provider behavior and market outcomes and finds that increased claim denial rates result in more healthcare consolidation and higher healthcare spending. Leder-Luis (forthcoming) and Howard and McCarthy (2021) have found significant deterrence effects of anti-fraud and abuse enforcement under the False Claims Act. Similar to these papers, we study how hospitals perform coding decisions given a change in policy-makers' oversight and fraud-detection efforts.

The rest of the paper proceeds as follows. Section 2 introduces the institutional background. Section 3 describes the theoretical model and the insight that underlies the empirical analysis. Section 4 presents the datasets and reports summary statistics. Section 5 discusses the empirical strategy. Section 6 shows the empirical results, and the last section concludes.

2 Background

2.1 Medicare Part A Hospital Inpatient Care

Medicare is the largest public health insurance program in the U.S., serving over sixty million beneficiaries and handling over $400 billion in reimbursements every year. Hospital inpatient care is the largest contributor to Medicare spending, and thus, we focus on the audits on this type of care. Since 1983, payments for hospital inpatient care are based on the Inpatient Prospective Payment System. Under this payment system, each patient admitted to a hospital is assigned a single diagnostic related group (DRG), depending on the diagnoses or procedures. For reimbursement purposes, each DRG corresponds to a weight. Medicare pays the hospital a flat amount for each admission, which is proportional to the DRG weight plus some adjustments. Specifically, the final amount of reimbursements the hospital receives equals the base payment (which is the product of the DRG weight and the base rate adjusted for geographic factors) plus additional adjustments associated with the hospital type and the specific care provided to the patient (Congressional Research Service, 2021). Thus, for the same DRG, the ultimate payments differ across hospitals, reflecting geographic heterogeneity

in the cost of hospital inputs, the hospital's performance on a set of quality measures that could lead to financial consequences, whether a hospital is a teaching hospital, whether the admission falls into a high cost outlier, etc. The DRG weights are periodically updated to reflect changes (increases or decreases) in the average cost of treating patients admitted under the same DRG.

Multiple DRGs are classified into one *base DRG* when all these DRGs share the same primary diagnosis or procedure. For instance, DRGs 693 – 694 designate admissions for the following diagnoses: "urinary stones w/ major complication or comorbidity (MCC)" and "urinary stones w/o MCC." Both belong to the base DRG—urinary stones. We refer to each DRG within a base DRG a tier, or a severity level intercheangeably, and refer to the most severe DRG as the *top code*. Hospital submitting claims to Medicare under the top code receive the highest reimbursement among all admissions within the same base DRG.

Here is an example of the base payments for a patient belonging to the base DRG "urinary stones." In Fiscal Year (FY) 2021, the overall average of DRG base payment rate is $5,961 for DRG weight 1 (Congressional Research Service, 2021).[5] Coding a patient from "urinary stones w/o MCC" to "urinary stones w/ MCC" could increase the reimbursement, on average, by $3,216 per discharge based on this base payment rate.[6]

All the clinical information collected during the course of treating the patient gets recorded into the patient's chart. Hospitals then employ coding staff that translate all the medical information from patients' charts into standardized, billable codes for reimbursement purposes, a process called "coding." "Upcoding" refers to the practice whereby DRG codes with higher reimbursement rates than medically justified are selected into claims forms. Up-coded claims submitted by hospitals could involve the claims that are incorrectly coded or lack sufficient documentation. Insufficient documentation of the claims could arise from the fact that the documentation justifying top-code billing does not exist or that documentation would facilitate an upcoding determination by auditors if the case is selected for review.

2.2 Medicare Recovery Audit Program

CMS adopt various auditing strategies to ensure providers bill Medicare appropriately. These include the Comprehensive Error Rate Testing (CERT) program, the Recovery Audit Pro-

[5]Fiscal year is the accounting period for the federal government, from the fourth quarter of the previous year to the third quarter of the current year.

[6]This estimate is an overall average of the base payment across the country. The actual amount a hospital receives depends on the location of the hospital and the further adjustments. The weight in FY 2021 for the DRG "urinary stones w/ MCC" is 1.2688, and the weight for the DRG "urinary stones w/o MCC" is 0.7293.

gram (RAP), Medicare Administrative Contractors, Supplemental Medical Review Contractors, and the Zone Program Integrity Contractor audits. In this paper, we focus on the RAP. This program reviews claims from Medicare Parts A and B.[7] It is unique and distinct from other programs because of its ability to conduct widespread post-payment review. RACs can select for review whatever claim they deem appropriate and employ a variety of strategies to identify claims that are most likely to yield a determination of improper payments.

The program started as a demonstration program in six states, from March 2005 to March 2008. It was then rolled out nationally by the end of 2010, using four independent contractors, each responsible for claims in a geographically defined region that was about one-quarter of the country.[8] All the RACs are paid on a contingency fee basis, a percentage of the corrected amount, with the base ranging from 9-12.5 percent for all claim types other than durable medical equipment.[9] RACs get paid for both the finding of overpayments and underpayments but have to return the fee if the decision is overturned at all levels of appeal.

Over time, CMS implemented a series of program enhancements in the RAP, with the goal to increase program transparency, improve program efficiency and, more importantly, reduce the burden healthcare providers face in the auditing process. The most salient change among these improvements is imposing a baseline limit for additional documentation requests (ADR) at 0.5%, among Medicare institutional providers, i.e., hospitals, skill nursing facilities, etc.[10] Under this requirement, the number of additional documents RACs request to review for a particular institutional provider cannot exceed 0.5% of the providers' total number of claims paid by Medicare in the last 12 months. The ADR limit became effective, starting from January 1^{st}, 2016. Since the majority improper payments identified by RACs come from complex review that requires additional documents from providers, a direct outcome of this policy is a sharp decrease in annual audit probability, as shown in Figure 1. Also, the dramatic decline in audit rates immediately after the implementation of the limit is consistent across contractors. We exploit this discrete change as an important variation in our empirical analysis, which we will discuss in more detail in Section 5.

The imposition of ADR limits mainly arose from hospitals' complaints about the increas-

[7]Medicare Part A covers inpatient care in hospitals, skilled nursing facility care, hospice care, and home health care. Part B covers outpatient care, durable medical equipment, and so on.

[8]Appendix Figure A1 displays the areas operated by each of the four contractors, which was assigned in 2010.

[9]The base rate of the contingency fee is revised to be 10-14.4 percent in FY 2016 after the procurement and contract modification.

[10]The other program improvements that occurred at the similar time include using provider surveys to assess contractor performance, encouraging RACs to have specialist panels for consultation, setting up improved target accuracy and appeals overturn rates, and so on (Recovery Audit Program, 2015).

ing burden they face in dealing with RACs. Hospitals found the process of documentation and verification onerous if their claims were identified as improper by RACs.[11] This had become worse as RACs acted more aggressively over time, especially after they were allowed to scrutinize claims with short inpatient stays since 2011 (Shi, 2024). According to the American Hospital Association's RACTrac survey in Q1:2014, 48% hospitals spent over $25,000 per month on managing the recovery audit processes, and 11% hospitals spent over $100,000. To fight back, hospitals appealed a large number of claims with overpayment determinations, making the system heavily overloaded. For instance, the number of appealed claims increased by 506 percent between 2012 and 2013, resulting in a backlog of over 800,000 RAC appeals at the administrative law judge level as of 2014. Considering that only 72,000 appealed claims were processed per year at that point, it could take a decade to resolve some of the filed claims.[12] The American Medical Association issued a letter to CMS, requesting an overhaul of the recovery audit process, stating that RACs "are often wrong and their bounty-hunter like tactics have caused physician practices undue hardship and expense."[13] The strong push-back from healthcare providers is one of the important reasons why CMS decided to scale back the RAP and implemented a series of program enhancements.

Note that CMS procured a new round of recovery audit contracts, which became effective in early 2017. Three of the four original RACs continued in the new contracts. Most of the contract modifications were part of the program improvements mentioned above and became effective prior to the new round of contracts. To minimize the effect of contract renewal, our empirical analysis focuses on the regions that are consistently operated by the same RAC before and after the start of new contracts.

Claims selected for investigation can be reviewed by RACs in three ways: automated, semi-automated, and complex. The first two review approaches are essentially claims data analysis. The semi-automated approach sometimes entail supporting documents for substantiation. Complex reviews of claims involves qualified coders and/or medical professionals, who carefully examine all supporting medical records before claim adjudication. Most of the reviews are complex, and in fact, most improper payments from upcoding are detected during complex reviews with more human involvement. The most common reasons for improper payments include: (a) providers bill a service that is not covered by Medicare or does

[11] See https://www.cagw.org/thewastewatcher/legislation-puts-recovery-auditing-and-taxpayers-risk.

[12] See https://www.policymed.com/2016/04/courts-to-finally-take-up-cms-recovery-audit-contractors-appeals-backlog.html.

[13] See https://www.ama-assn.org/practice-management/medicare-medicaid/payment-recovery-audit-program-needs-overhaul-doctors-cms.

not meet the medical necessity criteria; (b) the service billed is not correctly coded; and (c) the documentation submitted does not support the service ordered. When an auditor detects an overpayment, the associated provider will receive a notification letter including the rational for the determination. After that, the provider either fulfills the payments or initiates a discussion or an appeal process.[14]

3 Model

3.1 Model setup and prediction

We develop a simple model characterizing a hospital's decision to top-code a patient within a particular base DRG. For ease of notation, our model conditions on base DRGs with two levels, although some of our empirical analyses extend to base DRGs with three levels.[15]

For each admission within a given base DRG, we use s to denote the severity of the patient's condition, which we assume follows some random distribution between 0 and 1. We let $\underline{s}^u$ denote the level of severity above which a patient qualifies as a top code case according to medical coding guidelines. The hospital will top-code the patient if her severity level $s \geq \underline{s}^u$. In this case, the hospital receives profits equal to a baseline payment rate, B, multiplied by the weight of the top tier DRG, ω^{top}, minus the cost, c^{top}, associated with treating the more severe patient.

For the remaining cases where $s < \underline{s}^u$, the hospital also needs to decide on their coding tier assignment. If the patient is coded to the bottom level, the profit from treating this patient equals $B\omega^{bot} - c^{bot}$, where ω^{bot} denotes the weight for the bottom DRG within the given base DRG and c^{bot} denotes the cost of treating a less severe patient. If the patient is coded to the top tier, it would be upcoding that leads to profits equal to $B\omega^{top} - c^{bot}$. However, by doing so, the hospital is also exposed to the risk of being detected in an audit for improper coding. If this is the case, the hospital has to return the exceeding amounts of reimbursements and could be subject to heightened scrutiny in future monitoring. Thus, the hospital has to decide the extent to which it inappropriately top-codes patients, balancing the gains from the extra revenue from top coding against the potential loss if it is found out in an audit.

Let $\tilde{s}^u$ denote the *least* severe patient that the hospital assigns to a top code but does *not*

[14] In the case of an underpayment, the letter specifies the process how the payment is returned to the provider.

[15] The key implications from the model can be easily extended to base DRGs with three levels.

meet the criteria for this level according to the medical coding requirements.[16] The hospital receives the following expected profit:

$$
\begin{aligned}
E[\pi(\tilde{s}^u)] = & \underbrace{\int_0^{\tilde{s}^u} \left[B\omega^{bot} - c^{bot}\right] ds}_{\text{Total profit from proper bottom coding}} \\
& + \underbrace{\int_{\tilde{s}^u}^{\underline{s}^u} \left[B\omega^{top} - c^{bot} - \phi^{\text{AUDIT}} \times \phi^u(|\underline{s}^u - s|) \times \xi^u\right] ds}_{\text{Total profit from upcoding}} \\
& + \underbrace{\int_{\underline{s}^u}^{1} \left[B\omega^{top} - c^{top}\right] ds}_{\text{Total profit from proper top coding}}
\end{aligned} \tag{1}
$$

where ϕ^{AUDIT} refers to the probability of an admission being audited by RACs.[17] ϕ^u denotes the probability that, in the event of an audit, auditors detect upcoding for the given admission. We assume that $\phi^u(\cdot)$ increases with $|\underline{s}^u - s|$ at an increasing rate. That is, the probability of an improperly top-coded admission getting detected and flagged as upcoding is greater if the severity level of the admission is further below the cutoff, whereas the ϕ^u for marginal cases that are close to the cutoff would be relatively small. Moreover, we also assume that RACs would find it *increasingly* difficult to identify upcoding errors as the actual condition of the case gets closer to the cutoff. ξ^u denotes the consequences associated with an audit determination of over-payments due to upcoding, including returned payments, reputation loss, and greater scrutiny by Medicare in the future.

Thus, the total profit arises from (i) the expected profit from proper coding (both top and bottom coding) that is legitimate and (ii) the expected profit from upcoded claims that involve improper payments. In particular, the last term in total profit from upcoding $\int_{\tilde{s}^u}^{\underline{s}^u} \left[\phi^{\text{AUDIT}} \times \phi^u(|\underline{s}^u - s|) \times \xi^u\right] ds$, accounts for the possible financial costs from a potential audit. Hospitals seek to find out the optimal $\tilde{s}^u$, so as to maximize the expected profit. The hospital's tier coding decision is characterized as a tradeoff between the extra revenue from

[16] Our model can capture an alternative mechanism—reduced "undercoding" when monitoring becomes less. For instance, hospitals might undercode certain cases to minimize costly audits and then reduced undercoding when audit pressure became smaller. To model this case, $\tilde{s}^u$ (with $\tilde{s}^u > \underline{s}^u$) denotes the *most* severe patient that qualifies for the top tier but hospitals assign a bottom code. Note that our model can only capture one of the two mechanisms: increased upcoding with less monitoring and decreased undercoding with less monitoring. Here, we focus on the former case and provide more discussion on the latter case in Section 3.2.

[17] ϕ^{AUDIT} could also refer to the audit probability by other Medicare contractors, such as the CERT auditors, the Medicare Administrative Contractors, and so on. Assuming no significant changes in Medicare's other auditing strategies and given our focus on RACs, we let ϕ^{AUDIT} denote the audit probability by RACs.

coding (improperly) higher and the costs of being found to engage in improper coding, if any, in audits. We obtain the first order condition for profit maximization by taking the derivative of the expected profit with respect to $\tilde{s}^u$:

$$-B(\omega^{top} - \omega^{bot}) + \phi^{\text{AUDIT}} \times \phi^u(|\underline{s}^u - \tilde{s}^u|) \times \xi^u = 0. \tag{2}$$

Imposing the ADR limit in the RAP leads to a sharp decrease in audit rates, i.e., ϕ^{AUDIT}—the probability of a hospital getting audited—going down. We seek to understand how hospitals' coding decisions change in response to reduced audit pressure. To analyze this relationship in the model, we consider to what extent the level of upcoding varies given a change in ϕ^{AUDIT}. To achieve this, we take the derivative w.r.t. ϕ^{AUDIT} on both sides of Equation (2):

$$\phi^u(|\underline{s}^u - \tilde{s}^u|) \times \xi^u - \phi^{\text{AUDIT}} \times \frac{\partial \phi^u(|\underline{s}^u - \tilde{s}^u|)}{\partial(|\underline{s}^u - \tilde{s}^u|)} \times \frac{\partial \tilde{s}^u}{\partial \phi^{\text{AUDIT}}} \times \xi^u = 0.$$

Rearranging terms, we have

$$\frac{\partial \tilde{s}^u}{\partial \phi^{\text{AUDIT}}} = \frac{\phi^u(|\underline{s}^u - \tilde{s}^u|)}{\phi^{\text{AUDIT}} \times \frac{\partial \phi^u(|\underline{s}^u - \tilde{s}^u|)}{\partial(|\underline{s}^u - \tilde{s}^u|)}} > 0. \tag{3}$$

The partial derivative is positive as $\phi^u(\cdot) \geq 0$, $\phi^{\text{AUDIT}} > 0$, and $\frac{\partial \phi^u(|\underline{s}^u - \tilde{s}^u|)}{\partial(|\underline{s}^u - \tilde{s}^u|)} > 0$. The last inequality holds, arising from our assumption that the probability of getting detected for improper coding increases with the extent to which the patient's actual severity level falls below the threshold for the top code. The results in Equation (3) suggest that hospitals tend to engage in more upcoding practices as the audit probability drops.

3.2 Discussion on the model

Several points are worth noting. First, our model focuses on post-payment review of claims, abstracting from the steps of claim processing and denials that happened before payments were made. In other words, our analysis is conditional on claims passing the claims processing systems (managed by Medicare Administrative Contractors) and being approved for reimbursements. This part of the reimbursement process is studied in detail by League (2024). It is likely that hospitals' strategic response in coding due to the imposition of ADR limits may have changed the composition of claims submitted, which could affect the claim denial risk and hence the probability of facing a post-payment audit. Thus, our results may

also pick up the effect of the change in pre-payment claim denial risk.

Second, an alternative mechanism that is not captured in our model but could lead to similar empirical implications is the potential of undercoding legitimate top-tier claims prior to the policy change. Hospitals might previously choose to undercode these claims to avoid costly audits, even when they know they would ultimately prevail in those audits. Then, hospitals might reduce this type of defensive coding after the ADR limit was imposed. As a result, a reduction in the down-coding rate after the policy change may show a similar pattern to the mechanism we have discussed. We believe that this alternative mechanism is not very likely, according to the findings from our placebo test that we discuss in Section 6. As a result, our model here abstracts away from the undercoding mechanism.

Third, similar to most studies in the hospital upcoding literature, we consider the coding decision at the base DRG level, allowing for the incentive of revenues and costs varying across base DRGs.[18] Finally, we assume that hospitals only consider marginal cases—those close to $\underline{s}^u$—for upcoding given that the cost of coding cases far away from the cutoff could be exceptionally high.

Our theoretical implication is fairly standard and in line with that in the literature on crime and public economics (Becker, 1968; Allingham and Sandmo, 1972). However, we believe that it is worthwhile laying out in some detail as it becomes important for guiding one of our heterogeneity analyses — how the responsiveness to the policy change varies by the financial return from top coding. A natural prediction is that hospitals tend to code more aggressively among high-revenue cases given the same change in monitoring. However, the theoretical prediction turns out to be less straightforward according to our model, because the optimal level of coding depends on the tradeoff between the gains from coding higher-revenue cases and and the concerns on greater upcoding detection risk in these cases.

4 Data and Descriptive Statistics

4.1 Datasets

Our first dataset is the 100% of audited claims in the Medicare Recovery Audit Program, including all the Medicare hospital inpatient claims that have been reviewed by RACs. The audit microdata includes claim characteristics, such as service dates, service providers, DRGs, payments, and recorded diagnoses and procedures. It also includes information on

[18] Our current setting does not account for coding decision across base DRGs, though the main insights here could be carried over to the more complicated setting.

the audits, such as the date selected for review, the reason for review, final determinations, and the amounts of recovered payments. We calculate the total number of audits for each DRG by each RAC in each year, which is the numerator of our key variable of interest, the audit rate at the RAC-DRG-year level.

We obtain the denominator for the audit rate from our second dataset—the Inpatient Utilization and Payment Public Use File (Inpatient PUF) published by CMS.[19] We extract the hospital inpatient discharges per DRG per state and aggregate themn to the RAC-DRG level.[20] Combining the audit microdata and discharge counts, we construct the key variable of interest—the audit rate—defined as the total number of claims selected for review by RACs among all the discharges within the RAC-DRG-year cell.

Our analysis focuses on the years between 2014 and 2019 for the following reasons. First, we exclude the years before 2014 to minimize the incidence of any structural breaks in audit rates during our pre-treatment period. CMS prohibited RACs from reviewing claims regarding inpatient hospital patient status from October 2013 and this type of reviews accounted for a substantial portion of claims investigated by RACs before 2014 (Recovery Audit Program, 2014).[21] Second, we drop the data after the year 2019 because of the audit pause due to the pandemic of coronavirus disease 2019.

We divide the entire sample into the treatment and control groups, using the sharp decrease in audit rates due to the imposition of the ADR limit that became effective in January 2016. Specifically, for each RAC-DRG pair, we calculate the pre-treatment average audit rate by averaging the recovery audit rate across the years 2014 and 2015. Likewise, we compute the average audit rate during the post-treatment period, from 2017 to 2019. Then, we assign a RAC-DRG pair to the treatment group if the average post-treatment audit rate is smaller than the average pre-treatment audit rate. We drop the year 2016 in defining the treatment group because the audit rates for most DRGs in this year are close to zero due to the phase-out of recovery audits in the transition to the new contracts. Such a division results in 87% of observations in the treatment group and 13% in the control group.

Our third dataset comes from the 5% Inpatient Standard Analytic File from the Limited

[19]The data is available at https://data.cms.gov/provider-summary-by-type-of-service/medicare-inpatient-hospitals/medicare-inpatient-hospitals-by-provider-and-service.

[20]While the number of discharges is not equivalent to that of claims, less than 1% of bills come from multiple stays (https://resdac.org/articles/differences-between-inpatient-and-medpar-files).

[21]CMS adopted the "Two-midnight rule" in FY 2014, clarifying when inpatient hospital admissions are generally appropriate for Medicare Part A payments. In the meantime, CMS started the Probe and Educate process, providing education to providers in accordance with the rule, and prohibited RACs from conducting patient status reviews for inpatient claims. This type of reviews constituted a large portion of claims examined by RACs at that time.

Data Sets provided by CMS. It is a 5% random sample of all the hospital inpatient claims among Medicare beneficiaries. We construct our key outcome measure, percent top codes, from this dataset, defined as the fraction of patients assigned the top DRG within a base DRG per hospital per year.[22] This is similar to the outcome measure used by most of the literature that studies hospital upcoding behavior (Silverman and Skinner, 2004; **?**; Li, 2014; Ganju et al., 2021; Gowrisankaran et al., 2023). Finally, we complement these datasets with DRG information released by CMS, where we obtain DRG type and DRG weights. Given our interest in the tier coding assignment within a base DRG, we focus on base DRGs with multiple levels. On average, there are 266 base DRGs with more than one tiers among the total of 334 base DRGs.

Sample construction. There are, on average, 4,661 hospitals (264 base DRGs) per year after we merge the audit micro data with the Inpatient PUF and the 5% Inpatient Standard Analytic File. We drop hospitals whose operating RACs changed after the recovery contract renewal in 2017, which leaves us an average number of 3,150 hospitals (264 base DRGs) each year. To avoid our results driven by certain RAC-DRG pairs that are only present either before or after the change in ADR limits, we construct a balanced panel, including the RAC-DRG pairs that are present in every year during our sample period. This leave us an average number of 3,136 (256 base DRGs) hospitals per year.[23]

4.2 Summary statistics

Figure 1 displays the average audit rates among the top-tier DRGs for the entire sample and separately for each RAC over time.[24] Specifically, the overall audit rate is a weighted average across the three RACs, with the share of claims to the total claims as the weight. Before 2016, it was 0.88% in 2014 and went up to 1.51% in 2015.[25] The average audit rate plummeted immediately after 2015, below the cap, 0.5%, in every year after the imposition of the limit.[26] This figure also shows the trend of average audit rates by RAC. While the general pattern is similar among all the three contractors, the relatively decline varies across

[22] Given that it is a random sample of the 100% Medicare hospital inpatient claims, we believe that the percent top codes constructed from this dataset is representative of the entire population.

[23] As a sensitivity test, we re-estimate our main specifications using the unbalanced panel or based on a balanced panel of hospitals, and the main findings hold in both cases. We provide more details in Section 6.6.

[24] Since one of the four original RACs left the program after the procurement, our analysis focuses on the three RACs who stayed.

[25] Appendix Table A1 provides more information.

[26] Note that the audit rate reached the bottom in 2016, which is also related to the partial phase-out of recovery audits due to contract renewal.

them. This constitutes important sources of variation for our identification—the variation in the decrease of audit rates before and after the imposition of the ADR limit across the RAC-DRG pairs.[27]

Figure 2 shows the frequency distribution of the average audit rates per DRG across time separately for the pre-treatment and post-treatment periods. We include a vertical line to indicate the ADR limit, 0.5%. Figure 2(a) shows the distribution for the entire sample. Prior to the imposition of this limit, the average audit rate per DRG was relatively evenly distributed from zero to approximately 2.5%. After the the ADR limit was imposed, most of the DRGs ended up having an audit probability below the cap.[28] Figures 2(b)-(d) show the distribution for each RAC, respectively, and all share a similar pattern. In principle, a RAC could audit more than 0.5% of claims for a particular DRG or set of DRGs within a hospital as long as it compensated with low audit rates in other DRGs so as to remain under the cap, in total, for that hospital. Indeed, as can be seen in Figure 2, the audit rate is above the cap in some cases post policy change, but the vast majority of audit rates at the DRG level (averaging across hospitals) are under the 0.5% cap. This suggests that our identification comes from the change in audit rates among a broad set of DRGs rather than a limited number of select DRGs.[29]

Table 1 shows the summary statistics for percent top codes and DRG weights for the overall sample and separately for the treatment and control groups. The average proportion of patients assigned the top DRG within a base DRG varies over time, ranging from 69.5% to 72.4%, with higher percent top codes among the control group. The average DRG weights are quite stable over time, with slightly greater DRG weights in the control group. The number of DRGs or RAC-DRG pairs is constant over time because our analysis is based on a balanced panel at the RAC-DRG level. There are more DRGs or hospitals in the treatment group because most of the DRGs experienced substantial changes in audit rates after the ADR limit was imposed.

[27]Note that our calculation of audit rates includes all the claims that were selected for review, regardless of the review type, even though the imposition of ADR limit is the most relevant to complex or semi-automated reviews, either of which requires ADR. We expect that the results of excluding automated reviews would be very similar because this type of reviews accounts for about 2.5% of the total investigated claims.

[28]There are approximately 9.4% of DRGs with audit rates below 0.5% before the policy change and 86.4% of DRGs with audit rates below 0.5% after the policy change.

[29]Appendix Figure A2 shows the frequency distribution of the average audit rates per hospital separately for the pre-treatment and post-treatment periods. Prior to the policy change, approximately 65% of hospitals with average audit rates below 0.5% among those with non-zero average audit rates, whereas this rate went up to almost 99% after the ADR limit was imposed. CMS granted an exception that RACs could request additional documents above the cap in certain cases, such as unusually high claim denial rates from the hospitals.

5 Empirical Strategy

The main goal of our empirical strategy is to explore whether claims in DRGs for which the recovery audit rates went down display any uptick in the rate at which they are coded into top reimbursement tiers within the given base DRG. The imposition of the ADR limit reduces the audit pressure facing hospitals and could lead to potentially more improper coding behavior, as predicted in Section 3. To examine empirically the extent to which hospitals vary the reporting of top-coded patients in response to the adjustment of ADR limit, we use a DiD strategy, considering the change in the fraction of top-coded patients within the RAC-DRG pairs that experienced declines in audit rates from before to after the policy change, relative to the change in this fraction among the comparison group. Specifically, we use the following specification:

$$\texttt{TopCodeRate}_{jdt} = \beta \times \mathbb{1}\{t > 2016\} \times \text{Treat}_{rd} + \alpha^h \omega_{dt}^{top} + \alpha^\ell \omega_{dt}^{bot} + \gamma_{jd} + \lambda_t + \varepsilon_{jdt}, \qquad (4)$$

where $\texttt{TopCodeRate}_{jdt}$ denotes the percent claims within base DRG d at hospital j in year t that are coded into the top reimbursement tier. Treat_{rd} indicates the treatment group, equal to 1 if RAC r lowered the audit probability for the top level DRG in base DRG d after the imposition of the ADR limit. We assume that it takes one year for the change in audit probability to affect hospitals' coding behavior. Since the ADR limit became effective from the year 2016, we expect the treatment to start to take effect in 2017.[30]

Moreover, we include the weights for the top-level DRG and the next lower level, ω_{dt}^{top} and ω_{dt}^{bot}, to control for the potential financial incentives for top coding. γ_{jd} denotes hospital-base DRG fixed effects, capturing the time-invariant unobserved heterogeneity at this level, such as the pre-existing differences in patient characteristics, coding patterns, healthcare practices, etc. Note that since each RAC operates a large geographic region that is equivalent to approximately one quarter of the entire country, we also control for the RAC fixed effects which are subsumed in γ_{jd}. λ_t denotes year fixed effects, capturing, for instance, the changes in patient population at the national level over time. ε_{jdt} denotes the regression residual.

We use β to capture the difference between the average change in top coding rate post-2016 relative to pre-2016 in the treated RAC-DRG pairs and the average change in this rate in the control group—all the other RAC-DRG pairs—over the same period. Our theoretical

[30] Hospitals could observe the change in audit probability through the ADR from RACs during the year 2016.

prediction is that $\beta > 0$, suggesting that hospitals report relatively more patients with top codes among the DRGs for which the likelihood of getting audited by their operating RAC became lower post-treatment relative to pre-treatment. This provides evidence that hospitals are more aggressive in their coding practices when the monitoring becomes less intense. Note that implicitly we use percent top codes as a proxy for the level of upcoding or $|\underline{s}^u - \tilde{s}^u|$ in our model. Similar to the prior literature, it includes two types of top coding: the part that is legitimate and the part that is inappropriate. The identifying assumption is that levels of sickness in the Medicare inpatient population or the threshold for the top code does not change over this period or that if it does, the changes at the DRG level are not systematically different for patients in DRGs experiencing declines in audit pressure relative to patients in other DRGs.

To investigate whether there were differential pre-treatment trends in percent top codes for the treatment group relative to the comparison group, we extend the specification in Equation (4) using an event study research design:

$$\texttt{TopCodeRate}_{jdt} = \sum_{\tau=2014,\, \tau\neq 2016}^{2019} \beta^{\tau} \times \mathbb{1}\{t = \tau\} \times \text{Treat}_{rd} + \alpha^{h}\omega_{dt} + \alpha^{\ell}\omega_{dt}^{bot} + \gamma_{jd} + \lambda_t + \varepsilon_{jdt}. \tag{5}$$

The specification above is almost the same as that in Equation (4) except that we replace the key variable of interest in Equation (4)—$\mathbb{1}\{t > 2016\} \times \text{Treat}_{rd}$—with the interaction terms between Treat_{rd} and a series of year dummies. We assume that it takes one year for hospitals to respond to the change in audit probability and thus, omit the year 2016.[31] Thus, the estimated β^{τ}'s measure the effect relative to this year. Interacting the treatment group indicator with year dummies also enables us to capture the shifting effect on percent top codes due to the change in ADR limit. Our theoretical prediction is that β^{τ} starts to be positive when $\tau \geq 2017$. This provides evidence that hospitals code more aggressively in response to reduced detection risk.

The identification arises from the relative variation in percent top codes at the hospital/base DRG level across time. The key identifying assumption is that the change in audit probability at the RAC-DRG level post-treatment to pre-treatment is exogenous to the unobservables that could affect hospitals' top-coding decision after controlling for the DRG weights and various fixed effects. A potential identification threat could occur if a RAC

[31] The ADR limit kicked in in 2016, which would start to have an impact on hospitals' coding behavior in 2017 by assumption.

maintained the same audit probability for a particular hospital because the RAC suspects that the hospital might be more likely to engage in improper coding practices due to the hospital's characteristics or for a particular condition because it is easier to upcode patients with this condition due to technological changes. We believe that the likelihood for the former is low because the variation we rely on for identification is the average change in audit probability across all hospitals in the region operated by the same RAC—each RAC operating in almost one fourth of the country—rather than the change in audit rates for specific (types of) hospitals. We also believe that the reduction of audit rates occurred across the board due to the imposition of ADR limit instead of targeting to specific DRGs and hence the latter case is not very likely.

We estimate our main specification in Equations (4) and (5) using a standard Two-Way Fixed Effects (TWFEs) approach. We cluster standard errors at the RAC and base DRG level. We weight our regressions by the number of discharges within a hospital/base DRG cell. A strand of recent studies points out potential problems with the TWFEs approach like the one above to implement DiD strategies and to test for absence of pre-existing differential trends (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; De Chaisemartin and d'Haultfoeuille, 2024; Borusyak et al., 2024). In our case, the treatment occurs once and at the same time for RAC-DRG pairs, and therefore some of the issues are not as important. However, to reduce these concerns, we also use one of the new approaches—the one developed by De Chaisemartin and d'Haultfoeuille (2024), referred as dCDH from now on—to estimate the treatment effects.

We also extend our main specifications to examine how hospitals' responsiveness to the change in audit probability varies by medical condition or hospital type. First, we examine in some detail the differences between high-spread and low-spread base DRGs. Here, we follow most of the related literature and use spread—the difference in DRG weights between the top and lower level DRGs within a base DRG—as a proxy for the extra revenue from top coding.

Another dimension we explore in heterogeneity analysis is studying how the results differ across various hospital types. The literature has found that for-profit hospitals tend to code more aggressively, compared to not-for-profit hospitals. Thus, we examine the differential effects of the change in audit rates separately for these two types of hospitals. Similarly, prior studies have also found that hospitals with different financial health status might engage in different levels of improper coding. For instance, hospitals that undergo greater financial distress might seize the opportunity of lower detection risk and upcode more patients to

improve financial performance. Thus, we estimate the treatment effect, for financially healthy hospitals and financially less healthy hospitals, respectively. Finally, we also investigate whether hospitals' responsiveness to reduced monitoring varies between those located in metropolitan areas and those that are not, given that rural hospitals are more susceptible to financial volatility and may behave differently from hospitals in non-rural areas.[32]

6 Results

6.1 Main results

We first discuss the results for the main specifications where we consider the effect of the ADR limit on hospitals' top coding decision. Figure 3 presents the estimated coefficients and their 95% confidence intervals for the key variables of interest in Equation (5), the interaction terms between the indicator for the treatment group—the DRGs for which the audit rates decline after CMS restricted the number of claims RACs can send out ADR—and the series of year dummies. We present the estimates from both the TWFEs model and the dCDH estimator. We include a dashed line before year 2017 to indicate when the treatment starts to have an impact. Columns (1) and (2) of Appendix Table A2 provide more details on the results.

Immediately after 2016, we observe a larger fraction of discharges reported with top codes among the DRGs for which the audit probability declined relative to the comparison group. The effect is significantly positive in every year post-treatment and lasts for several years. In contrast, there is visually no significant trend in the pre-treatment period, which is confirmed by the p-value of the test for joint significance of the pre period coefficients, equal to 0.594 (0.767) from the dCDH (TWFEs) estimator.

The treatment effect estimated from the dCDH estimator corresponds to an average increase of almost 5 percentage points in top coding, or an average increase by 7.2% (=0.0498/0.692 $\times$ 100%) among the treatment group following the policy change, relative to the control group.[33] The results suggest that hospitals top-code relatively more patients among the DRGs for which the operating RAC lowered the audit probability following the

[32]See https://www.kff.org/health-costs/issue-brief/rural-hospitals-face-renewed-financial-challenges-especially-in-states-that-have-not-expanded-medicaid/.

[33]We obtain the 7.2% by dividing the estimated treatment effect from the dCDH estimator (shown in Column (1) in Panel A of Appendix Table A2) by the average percent top codes among the treatment group during the years 2014-2016 (shown in Table 1).

implementation of the ADR limit.[34] It is consistent with our theoretical prediction. The small difference between the coefficients from both estimators suggests that the bias from the conventional TWFE model is limited, as expected.

6.2 Bottom coding rate due to reduced monitoring

We examine whether hospitals vary percent *bottom* codes in a similar way to percent *top* codes, in response to reduced monitoring.[35] We define percent bottom codes as the fraction of patients reported with the *lowest* DRG within a given base DRG. Similar to the main analysis, we divide all the bottom DRGs (from base DRGs with multiple levels) into the treatment and control groups, using the same criteria; that is, we assign a bottom DRG into the treatment group if the operating RAC lowered the audit probability for this bottom DRG post-treatment to pre-treatment. We then re-estimate the main specifications using percent bottom codes as the dependent variable. If financial returns are the primary driver for hospitals' coding decision, we expect little difference in percent bottom codes between the treatment and control groups after the ADR limit was imposed; moving patients to the lowest level would not generate extra revenue, regardless of the level of monitoring. In contrast, the presence of significant changes in percent bottom codes among the treatment group post-treatment to pre-treatment suggests that factors other than financial incentives may also play a role. For instance, transition to the 10^{th} Edition of the International Classification of Diseases codes was implemented in October of 2015, which could affect the coding patterns of inpatient diseases and procedures. Moreover, as mentioned in Section 3, an alternative mechanism that may also lead to increased percent top codes following the implementation of ADR limits is the practice of undercoding when there were no restrictions on ADR.

[34] Shi (2024) does not find any significant increase in admissions after reduced audits on inpatient hospital patient status by RACs. A potential explanation is that the decision-making of coding an admission to higher levels could be different from that of admitting (unnecessary) patients. Also, the way that RACs identify improperly coded claims could be different from the way of adjudicating unnecessary admissions. Thus, the extent to which hospitals respond to the respective policy change could vary. Moreover, the policy context is different between both papers. A main reason that CMS forced RACs to reduce the review of patient status among hospital inpatient claims—the policy related to Shi's paper—is because of the implementation of the Two-Midnight rule, a new policy that makes the cutoff of a proper admission clearer and could potentially reduce the headroom for admitting unnecessary patients. However, in the context of our paper, there was no additional guidance or instruction on coding along with the implementation of the ADR limit. Thus, the headroom for upcoding might remain similar after the ADR limit was imposed. In fact, our finding of more top coding and Shi's finding of no significant changes in admissions, both due to reduced monitoring, are somewhat similar to the results in **?**, where she found no significant impacts on admissions but more patients reported with higher codes in response to a change in reimbursement rates.

[35] An alternative placebo test is to examine percent top codes using non-Medicare claims. We cannot do it due to data limitation.

Particularly in the latter case, we expect a *decrease* in percent bottom codes among the treated DRGs following the policy change since hospitals face less audit pressure now and may code less conservatively to recover more billable services than before.

Figure 4 presents the estimates of the interaction terms between the indicator for the treatment group and year dummies separately for percent top codes and percent bottom codes, based on the dCDH estimator.[36] The fraction of patients reported with bottom codes is rather flat, evolving around the X-axis from pre- to post-treatment periods, which is consistent with financial incentives being an important factor in the coding decision and does not support the mechanism of previous undercoding.

6.3 Treatment Effect Heterogeneity by Reimbursement Spread

Prior literature has found that financial incentives play an important role in hospitals' tier coding assignment (Silverman and Skinner, 2004; **?**; Cook and Averett, 2020). In this case, the extent to which hospitals respond to the change in the likelihood of being audited could vary according to the financial returns from top coding. To examine this, we re-estimate Equations (4) and (5) for base DRGs with different spread. We define spread as the difference in DRG weights between the top and next lower tier within a base DRG. Most of prior studies that investigate hospital DRG upcoding behavior use spread as a proxy for the extra revenue hospitals receive from coding more intensely.

The difference in responsiveness between high-spread and low-spread base DRGs can be understood through the lens of our theoretical model. According to Equation (3), the extent to which a hospital upcodes a patient depends on the following forces: (i) the baseline audit rate (ϕ^{AUDIT}), (ii) the probability of being flagged for upcoding in an audit ($\phi^u(\cdot)$), and (iii) the distance between the optimal upcoding threshold level and the appropriate threshold for the top tier ($|\underline{s}^u - \tilde{s}^u|$). We expect that $\phi_\ell^{\text{AUDIT}} < \phi_h^{\text{AUDIT}}$, where the subscript h (ℓ) denotes the high-spread (low-spread) base DRGs. This is confirmed by our data: the average pre-treatment audit rate for the top level DRGs is 1.74% (0.85%) among the high-spread (low-spread) base DRGs, whose pre-treatment average spread is above (below) the median. RACs might pay more attention to top-coded cases in base DRGs with larger spread, as they suspect that those top-coded claims are more likely to involve improper billing. Moreover, RACs might have more incentive to scrutinize those top-coded cases as they are paid on a contingency fee basis; the amount of the commission depends on the recovered payment

[36] Note that the estimates for percent top codes in this figure are the same as those in Figure 3. Appendix Figure A6 shows the estimates for the same coefficients using the TWFEs estimator. We provide more details on the estimates in Appendix Table A2.

amount, which is proportional to the spread. Given that $\phi_\ell^{\text{AUDIT}} < \phi_h^{\text{AUDIT}}$, we expect that $|\tilde{s}_\ell^u - \underline{s}_\ell^u| > |\tilde{s}_h^u - \underline{s}_h^u|$, i.e., the level of upcoding is more severe among base DRGs with smaller spread due to less audit pressure on these DRGs. Furthermore, with $|\tilde{s}_\ell^u - \underline{s}_\ell^u| > |\tilde{s}_h^u - \underline{s}_h^u|$, we have $\phi^u(|\tilde{s}_\ell^u - \underline{s}_\ell^u|) > \phi^u(|\tilde{s}_h^u - \underline{s}_h^u|)$ and $\frac{\partial \phi^u(|\underline{s}_\ell^u - \tilde{s}_\ell^u|)}{\partial \tilde{s}_\ell^u} < \frac{\partial \phi^u(|\underline{s}_h^u - \tilde{s}_h^u|)}{\partial \tilde{s}_h^u}$, by the assumption that $\phi^u(\cdot)$ is increasing and convex in $|\tilde{s}^u - \underline{s}^u|$. Taken the competing forces together, it remains theoretically ambiguous which is larger: $\frac{\partial \tilde{s}_\ell^u}{\partial \phi_\ell^{\text{AUDIT}}}$ or $\frac{\partial \tilde{s}_h^u}{\partial \phi_h^{\text{AUDIT}}}$.

It is then an empirical question which effect dominates. We examine whether the increased top coding in response to the reduced audit probability varies by the spread across base DRGs. To separate all the base DRGs into those with high spread and low spread, we first calculate the average spread during the pre-treatment period for each base DRG and then use the median as the cutoff to define the low- and high-spread groups.

Figure 5 reports the estimated coefficients and their 95% confidence intervals for the interaction terms with the indicator for the treatment group, separately for low-spread and high-spread base DRGs.[37] Following the policy change, we see that hospitals report more patients with top codes among the low-spread base DRGs that experienced declines in audit rates, relative to the comparison group. The effect is statistically significant in every year during this period, with an increasing trend over time. In contrast, the percent top codes among high-spread base DRGs during the post-treatment period was consistently flat around the X-axis. The results suggest that the response to the policy change is mainly driven by the base DRGs in which the extra revenue from top coding is lower. It may imply that, for hospitals, the concerns on greater audit pressure among the high-spread base DRGs dominates the potential gains from coding admissions more aggressively there.

6.4 Differential Effects by Hospital Type

From a policy perspective, it might also be important to understand how the responsiveness to the change in the level of monitoring varies by hospital type (Becker et al., 2005). We now discuss the results of examining differential treatment effects by hospital type. Figure 6(a) shows the main estimates separately for for-profit vs. not-for-profit hospitals.[38] Interestingly, both types of hospitals report more patients with top codes among the DRGs experiencing declines in audit probability post-treatment to pre-treatment, relatively to the comparison group. The treatment effect is also similar between both types of hospitals, as shown in

[37] Columns (3) and (4) of Appendix Table A2 provide more information. We present the results from TWFEs in a similar format in Appendix Figure A7(a).

[38] We report the estimates in Columns (1) and (2) of Appendix Table A3 and present the results from the TWFEs in Appendix Figure A7(b).

Appendix Table A3. It seems that hospitals, regardless of profit status, seize the opportunity to enhance revenues when monitoring becomes less intense.

We also compare the treatment effects between hospitals with different financial health status. Figure 6(b) displays the main results separately for financially-healthy and financially less-healthy hospitals.[39] We define a hospital to be the former (latter) if its average debt-asset ratio during the pre-treatment period is above (below) the median. Following the policy change, both types of hospitals saw an uptick in the fraction of top-coded patients among DRGs in the treatment group, relative to the control group. Visually, the relative increase in percent top codes seems to be greater for hospitals undergoing less financial distress. According to the estimated treatment effects, financially-healthy hospitals experienced an average increase of 6.1 percentage points in fraction top-coding, whereas the increase of this fraction is 3.9 percentage points for financially less-healthy hospitals.[40]

Finally, we estimate the differential effects of reduced monitoring between hospitals located in metropolitan areas and those in non-metropolitan areas. We define a hospital located in a metropolitan core if the county where the hospital is located belongs to an area that is densely populated and has a high degree of economic and social interaction.[41] Figure 6(c) presents the estimates separately for each type of hospitals. Similar to the above, the estimated treatment effect is similar between both types of hospitals.[42]

6.5 Economic Magnitude of More Intense Coding

We first estimate how the scale-back of the RAP affected the amount of reclaimed overpayments. To achieve this, we rerun the main specifications, replacing top-coding rates with the total amount of reclaimed overpayments at the hospital-DRG level, and present the results in Appnedix Figure A5.[43] Note that in this analysis we use the year 2015 as the baseline year and drop the year 2016 because the number of audits is very small in this year due to the partial phase-out of the old contracts. We see an immediate decline in corrected over-

[39] We present the results in Columns (3) and (4) of Appendix Table A3 and plot the estimates from TWFEs in Appendix Figure A7(c).

[40] Appendix Table A3 provides more details.

[41] We classify a county in a metropolitan core if its Rural-Urban Commuting Area Code is 1. See https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation/. We obtain the data on the designation of metropolitan status from the U.S. Census: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/.

[42] We include more information on the estimates in Appendix Table A3 and present the results using TWFEs in Appendix Figure A7(d).

[43] We focus on overpayments assuming that the policy change has little impact on underpayments (also consistent with the results in the placebo test in Section 6) and considering that most of the reclaimed improper payments arose from overpayments during our sample period—over 85%, on average.

payments following the imposition of ADR limits, corresponding to an average reduction of $7,839 per hospital-DRG pair. Thus, the direct savings from reclaimed payments dropped by $6,996, and Medicare paid RACs a smaller amount of contingency fee, by $843, at the same level.[44]

We next quantify the magnitude of additional reimbursements due to the more intense coding after the implementation of the ADR limit. The median discharges per hospital per base DRG among the treatment group is 43. From Column (1) of Appendix Table A2, hospitals assign more patients to the top-tier DRGs in the treatment group by 4.98 percentage points post-treatment to pre-treatment. The pre-treatment average spread of the base DRGs with treated DRGs is approximately 1.13. In 2021, the overall average base payment rate per DRG weight 1 is $5,961. Thus, the in-sample prediction of the extra payments made by Medicare per hospital-DRG pair due to more aggressive coding is $21,421.

Taken together, a back-of-the-envelope calculation suggests that for every dollar Medicare saves on paying RACs after scaling back the program, Medicare may have to bear a loss of $25, among which 67% comes from hospitals' more aggressive coding due to reduced monitoring. Note that our calculation here does not account for the potentially greater profits and lower administrative burdens due to less audit pressure facing hospitals, but rather shows the financial impacts on Medicare after the scale-back of the RAP.

6.6 Additional Analyses and Robustness Checks

First, we examine whether our finding of the relatively greater proportion of patients reported with top codes in the treatment group is driven by the relatively sicker population in that group over time. To achieve this, we replace percent top codes with the following outcome measures: average length of stay, average number of diagnoses reported and average Charlson Comorbidity Index (CCI).[45] If patients from the treatment group become sicker across time, we might expect longer hospital stays, more accompanying underlying medical conditions, or a higher score of CCI over time. Appendix Figure A3 shows the event study results separately for each of these measures. In none of them do we see a similar pattern to our main results.

Second, we conduct the following analyses to assess the robustness of our main results.

[44] We obtain the contingency fee by multiplying $7,839 by the contingency fee ratio, 10.75% (the average of 9% and 12.5%), similar to the approach in Shi (2024). The direct saving is derived from subtracting $843 from $7,839.

[45] CCI is a morbidity score reflecting mortality risk. It is also commonly used to assess patients' clinical situations (Charlson et al., 2022). We calculate a score for each claim based on the reported diagnosis codes and obtain the average at the hospital/DRG/year level.

we first re-estimate the main specifications using a balanced sample of hospitals. To achieve this, we further drop hospitals which only show up in parts of the sample periods. Appendix Figure A4(a) shows that the estimates are materially the same as our main results. Second, we redo the estimation on the unbalanced panel to see whether the main results are driven by select DRGs in the reduced (balanced) sample. Appendix Figure A4(b) presents the event study estimates. The similarity in the results between both samples suggests that the selection bias in the balanced panel is limited. Third, we re-estimate the main specifications by additionally including the following hospital characteristics: logged bed size, logged net patient revenues, logged non-medicare bad debt expense, logged uncompensated care costs, and year fixed effects interacted with for-profit and not-for-profit hospital indicators, respectively. Appendix Figure A4(c) shows the estimates. The main findings hold. Fourth, we redo the estimation without using sampling weights, and the main findings still hold, as shown in Appendix A4(d). To sum up, all these additional analyses here suggest that our main specification is rather robust to different (sub)samples or alternative specifications.

7 Conclusion

Given that information asymmetry is common between purchasers and suppliers, monitoring has been used as a common tool to ensure suppliers engage in responsible spending practices, in compliance with policies and regulations. However, increased oversight could be challenging, due to the potential social costs and agency problems. Prior studies have investigated the effectiveness and potential costs of ongoing monitoring programs. We complement the literature by examining how auditees respond to a large scale-back in a national audit program in the context of health care.

We first develop a simple model to characterize hospitals' decision to code admissions intensely in the presence of government audits. To test the hypothesis empirically, we exploit the declines in the audit rates across top-tier DRGs and RACs as exogenous variation in the incentives to upcode. We apply a difference-in-differences framework and find that hospitals top-code relatively more patients as their DRGs experience declines in audit probability, which is consistent with our theoretical prediction. For every dollar Medicare saves on the program after the scale-back, it may have to bear a loss of $25, with over two thirds coming from the increased payments to hospitals due to their more aggressive coding when monitoring becomes less. Our heterogeneity analyses suggest that the increase in top coding in response to reduced monitoring mainly comes from the DRGs generating less extra revenue,

perhaps because they might more easily escape the auditors' scrutiny. We also find that the responsiveness to the reduction in audit pressure is similar across hospitals, regardless of profit status, financial health status, or location.

Our results suggest that the Medicare Recovery Audit Program has potential in combating fraud, waste, and abuse in healthcare spending. While scaling back the program helps decrease social costs, such as lowering administrative burden for healthcare providers, reduced monitoring might lead to certain improper billing practices to flourish, undeterred by a lax auditing environment. Balancing the tradeoff between lowering social costs from oversight and the potentially increased improper billing that arises from reduced monitoring calls for thoughtful policy designs from regulators.

References

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of public economics 1*(3-4), 323–338. 1, 3.2

Angerer, S., D. Glätzle-Rützler, and C. Waibel (2021). Monitoring institutions in healthcare markets: Experimental evidence. *Health Economics 30*(5), 951–971. 1

Avis, E., C. Ferraz, and F. Finan (2018). Do government audits reduce corruption? estimating the impacts of exposing corrupt politicians. *Journal of Political Economy 126*(5), 1912–1964. 1

Becker, D., D. Kessler, and M. McClellan (2005). Detecting medicare abuse. *Journal of Health Economics 24*(1), 189–210. 1, 6.4

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy 76*(2), 169–217. 1, 3.2

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies 91*(6), 3253–3285. 1, 5

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230. 1, 5

Charlson, M. E., D. Carrozzino, J. Guidi, and C. Patierno (2022). Charlson comorbidity index: a critical review of clinimetric properties. *Psychotherapy and psychosomatics 91*(1), 8–35. 45

Congressional Research Service (2021). Medicare Hospital Payments: Adjusting for Variation in Geographic Area Wages. *CRS Reports*. 2.1

Cook, A. and S. Averett (2020). Do hospitals respond to changing incentive structures? Evidence from medicare's 2007 drg restructuring. *Journal of Health Economics 73*, 102319. 6.3

De Chaisemartin, C. and X. d'Haultfoeuille (2024). Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*, 1–45. 1, 5

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2018). Once bitten, twice shy? the lasting impact of enforcement on tax compliance. *The Journal of Law and Economics 61*(1), 1–35. 1

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013). Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from india. *The Quarterly Journal of Economics 128*(4), 1499–1545. 1

Finan, F. and M. Mazzocco (2021). Combating political corruption with policy bundles. Technical report, National Bureau of Economic Research. 1

Ganju, K. K., H. Atasoy, and P. A. Pavlou (2021). Do electronic health record systems increase medicare reimbursements? The moderating effect of the recovery audit program. *Management Science*. 4.1

Gowrisankaran, G., K. Joiner, and J. Lin (2023). How do hospitals respond to Medicare payment reforms? *NBER Working Paper No. 26455*. 4.1

Groß, M., H. Jürges, and D. Wiesen (2021). The effects of audits and fines on upcoding in neonatology. *Health economics 30*(8), 1978–1986. 1

Howard, D. H. and I. McCarthy (2021). Deterrence effects of antifraud and abuse enforcement in health care. *Journal of Health Economics 75*, 102405. 1

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica 79*(3), 651–692. 1

League, R. (2024). Administrative burden and consolidation in health care: Evidence from medicare contractor transitions. Technical report, Working Paper, 2023. https://rileyleague. github. io/files/MAC_transitions. pdf. 1, 3.2

Leder-Luis, J. (2020). Can whistleblowers root out public expenditure fraud? evidence from medicare. 1

Li, B. (2014). Cracking the codes: Do electronic medical records facilitate hospital revenue enhancement. *Working paper*. 4.1

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of Political Economy 115*(2), 200–249. 1

Recovery Audit Program (2011-2016). Recovery auditing in Medicare for Fiscal Year 2011 – 2016. *Centers for Medicare and Medicaid Services*. 10, 4.1

Sacarny, A. (2018). Adoption and learning across hospitals: The case of a revenue-generating practice. *Journal of Health Economics 60*, 142–164. 1

Shekhar, S., J. Leder-Luis, and L. Akoglu (2023). Unsupervised machine learning for explainable health care fraud detection. *NBER Working Paper No. 30946*. 1

Shi, M. (2024). Monitoring for waste: Evidence from medicare audits. *Quarterly Journal of Economics*. 3, 1, 2.2, 34, 44

Silverman, E. and J. Skinner (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics 23*(2), 369–389. 4.1, 6.3

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of economic literature 57*(4), 904–954. 4

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199. 1, 5

Telle, K. (2013). Monitoring and enforcement of environmental regulations: Lessons from a natural field experiment in norway. *Journal of Public Economics 99*, 24–34. 1

Vannutelli, S. (2024). From lapdogs to watchdogs: Random auditor assignment and municipal fiscal performance. Technical report, National Bureau of Economic Research. 1

Table 1: Summary Statistics for Key Variables

Year	2014	2015	2016	2017	2018	2019
All						
Fraction top coding (%)	69.5	70.2	70.5	71.6	72.4	72.4
Average DRG weights	2.90	2.90	2.90	2.91	2.89	2.94
# DRGs	256	256	256	256	256	256
# DRG-RAC pairs	749	749	749	749	749	749
# hospitals	3,252	3,231	3,121	3,111	3,082	3,019
Treatment group						
Fraction top coding (%)	68.6	69.4	69.7	71.0	71.8	71.9
Average DRG weights	2.91	2.91	2.91	2.92	2.90	2.95
# DRGs	253	253	253	253	253	253
# DRG-RAC pairs	627	627	627	627	627	627
# hospitals	3,219	3,199	3,070	3,067	3,043	2,986
Control group						
Fraction top coding (%)	75.5	75.6	75.6	75.6	76.4	76.2
Average DRG weights	3.04	3.03	3.04	3.06	3.04	3.08
# DRGs	100	100	100	100	100	100
# DRG-RAC pairs	122	122	122	122	122	122
# hospitals	1,889	1,887	2,083	2,030	2,007	1,945

Notes: Table 1 reports the summary statistics for the key variables in the analysis. Fraction top coding is defined as the percentage of patients reported with the top DRG in a given base DRG among all the patients from this base DRG. We obtain the average DRG weights by averaging the DRG weights across all the DRGs considered.

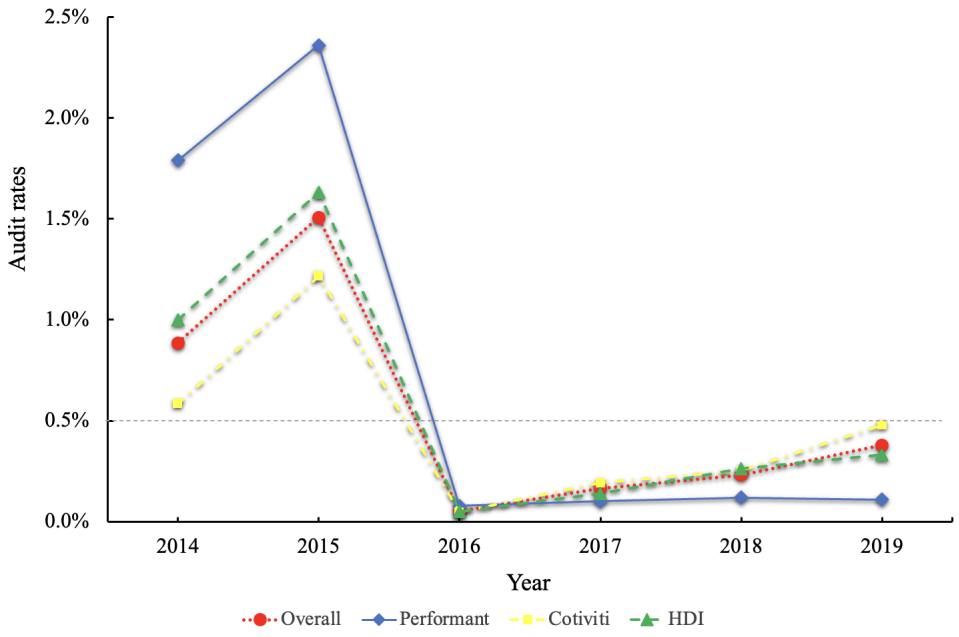
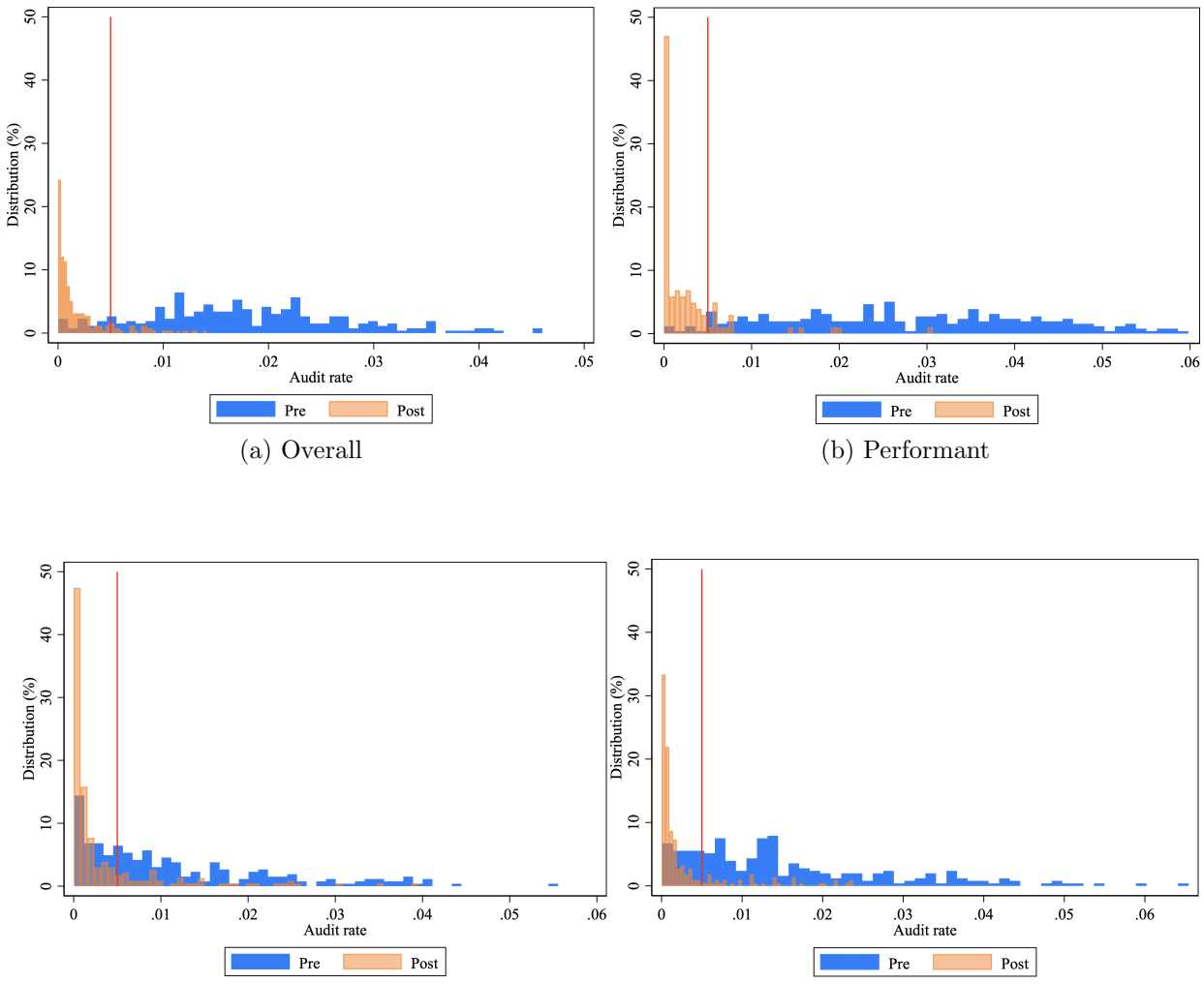


Figure 1: Average Audit Rates among Top-tier DRGs, Overall and by RAC

Notes: We plot the average audit rate overall and separately for each RAC. The overall average is a weighted average across the three RACs, using the share of claims to the overall claims as the weight. We include a dashed line indicating the ADR limit, 0.5%.



(a) Overall (b) Performant

(c) Cotviti (d) HDI

Figure 2: Frequency Distribution of Average Audit Rates per DRG, Before and After Policy Change

Notes: We show the difference in frequency distribution of average audit rates per DRG between pre- and post-policy change, for the overall sample (in (a)) and separately for each RAC (in (b)-(d)). For each DRG, we obtain the average audit rate across time for the pre-treatment and post-treatment periods, respectively. The vertical line denotes the ADR limit, 0.5%.

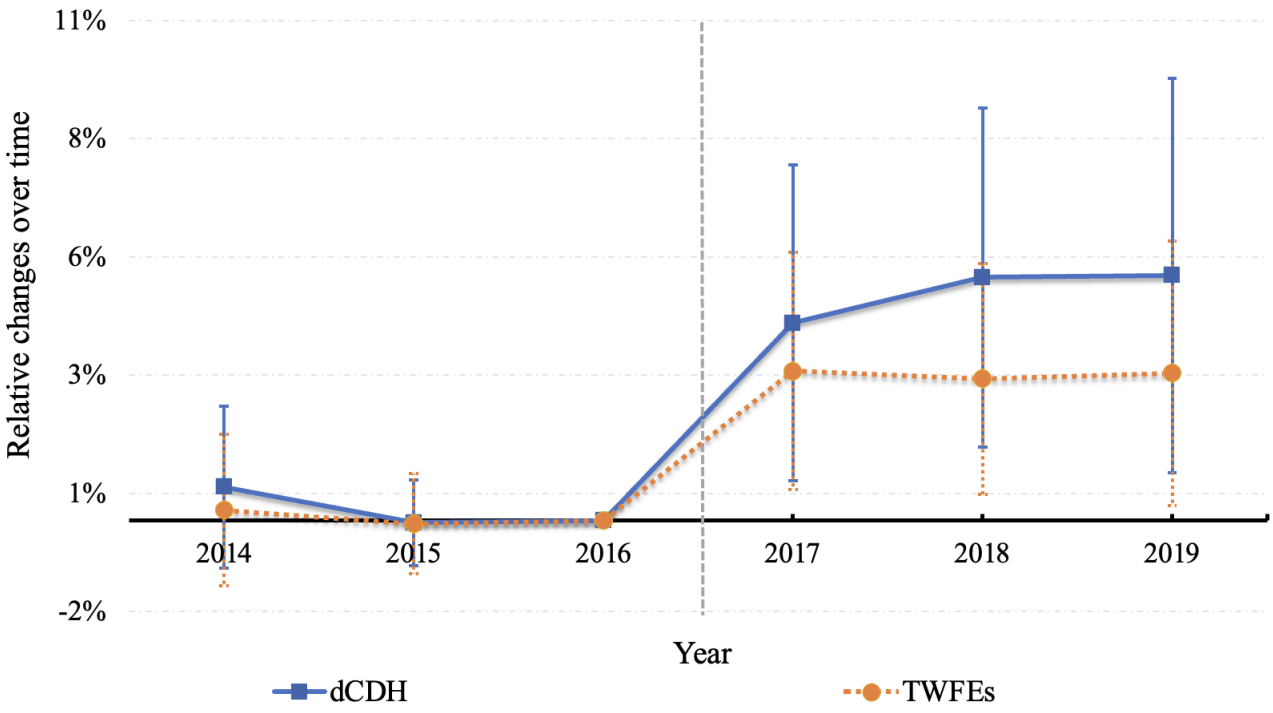


Figure 3: Effect of Reduced Audit Pressure on Top-Coding Rates
(dCDH and TWFEs Estimates)

Notes: The figure shows the event study coefficients (estimated by dCDH and TWFEs). The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.

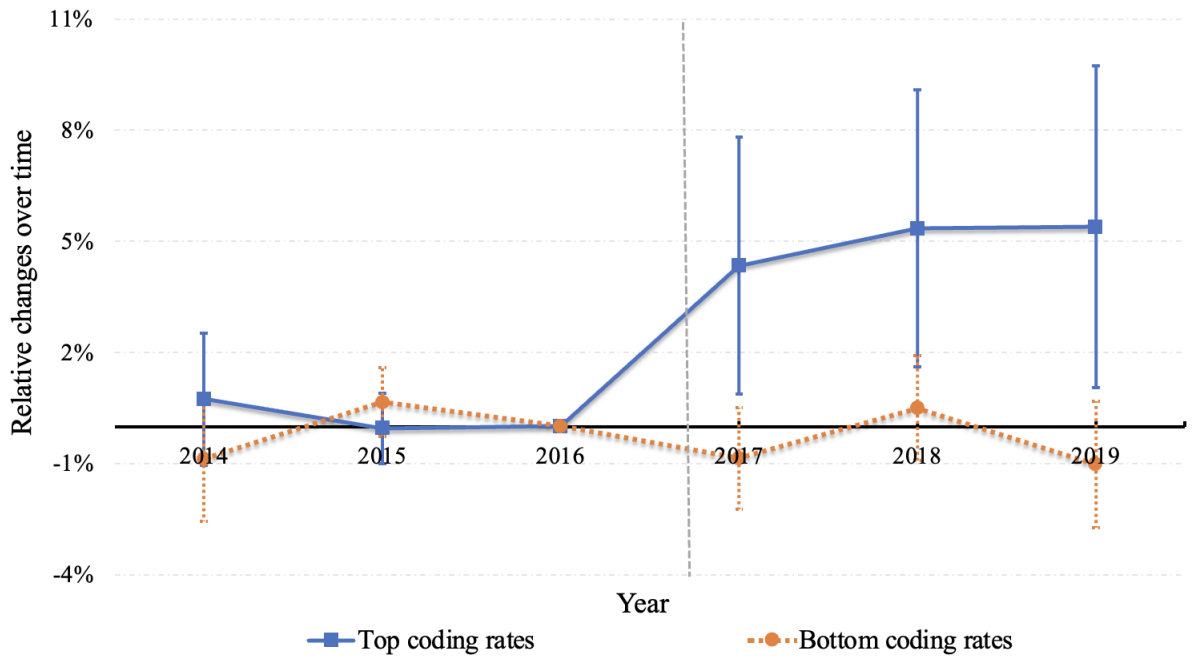


Figure 4: Effect of Reduced Audit Pressure on Top- and Bottom-Coding Rates (dCDH Estimates)

Notes: The figure shows the event study coefficients (estimated by dCDH). Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top (bottom) level DRGs within base DRGs with multiple levels for the analysis on the top (bottom) coding rate. Other regressors in the specification for top coding rates are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects. Other regressors in the specification for bottom coding rates are the weight of the lowest DRG, hospital-DRG fixed effects, and year fixed effects.

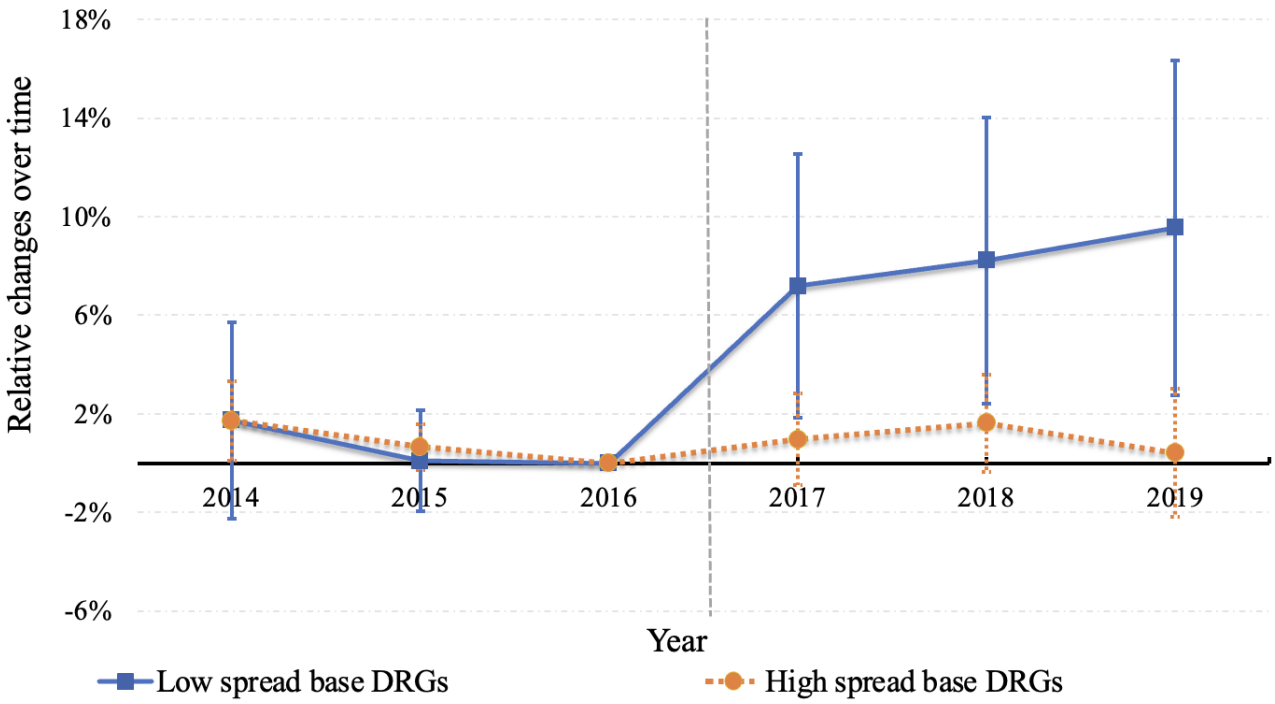
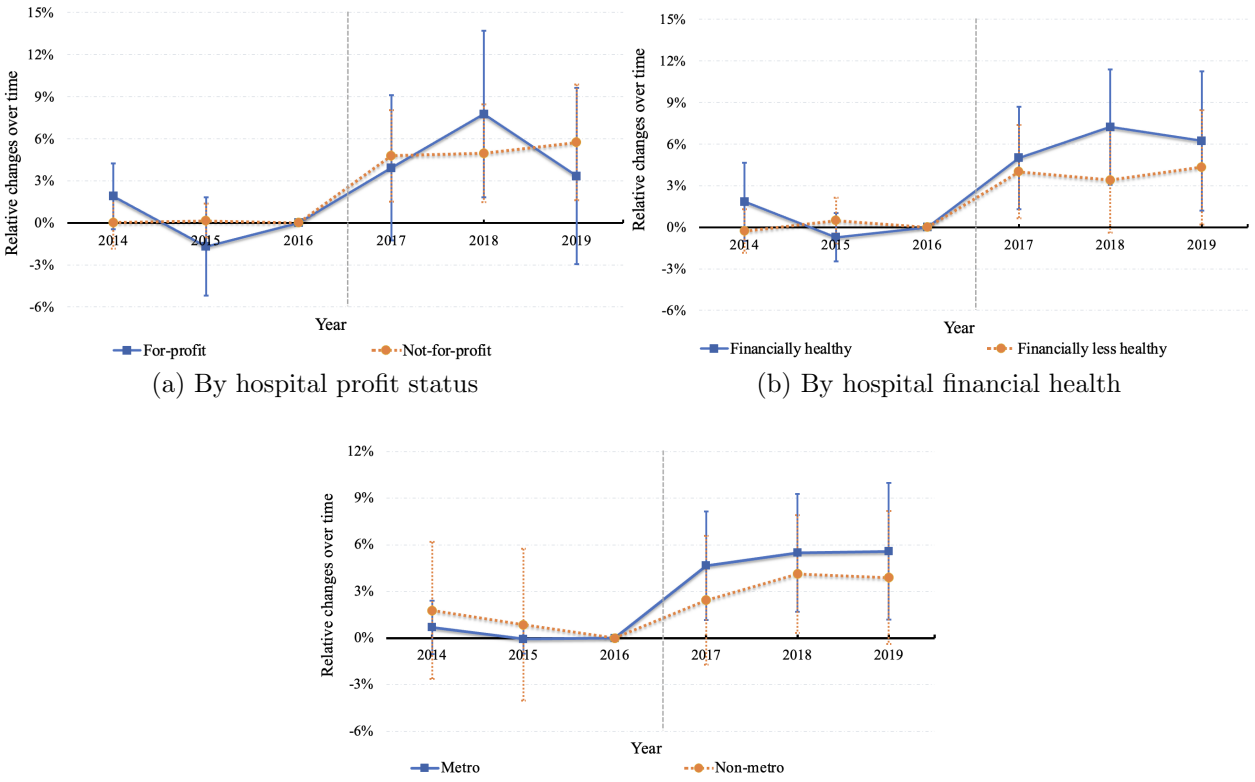


Figure 5: Effect of Reduced Audit Pressure on Top-Coding Rates, Below vs Above Median Pre-treatment Spread (dCDH Estimates)

Notes: The figure shows the event study coefficients (estimated by dCDH) by median of pre-treatment DRG spread. High-spread (Low-spread) base DRGs refer to those with spread greater (smaller) than the median of the pre-treatment spread. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.



(a) By hospital profit status

(b) By hospital financial health

(c) By metropolitan status

Figure 6: Effect of Reduced Audit Pressure on Top-Coding Rates by Hospital Type (dCDH Estimates)

Notes: The figures show the event study coefficients (estimated by dCDH) by hospital type. Financially healthy hospitals are those whose debe-asset ratio is below the pre-treatment median, and financially less healthy hospitals include the remaining hospitals. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.

Appendix I Extra Tables and Figures

Table A1: Average Audit Rates among Top-tier DRGs, Overall and by RAC (%)

Year	Overall	Performant	Cotiviti	HDI
2014	0.88	1.79	0.58	1.00
2015	1.51	2.36	1.22	1.63
2016	0.05	0.08	0.05	0.05
2017	0.16	0.10	0.19	0.14
2018	0.23	0.12	0.25	0.26
2019	0.38	0.11	0.48	0.33
# DRGs	749	241	256	252

Notes: The overall average is a weighted average across the three RACs, using the share of claims to the overall claims as the weight.

Table A2: Treatment Effects and Dynamic Treatment Effects on Top- and Bottom-coding rates

	DV: top-coding rates				DV: bottom-coding rates	
	Overall		Low-spread base DRGs	High-spread base DRGs	Overall	
	dCDH (1)	TWFEs (2)	dCDH (3)	dCDH (4)	dCDH (5)	TWFEs (6)
Panel A						
Treatment Effect	0.0498***	0.0317**	0.0824***	0.0100	-0.0049	-0.0007
	(0.0187)	(0.0147)	(0.0294)	(0.0092)	(0.0072)	(0.0046)
Panel B						
Treated×Year 2014	0.0074	0.0022	0.0174	0.0172	-0.0088	-0.0014
	(0.0091)	(0.0085)	(0.0203)	(0.0082)	(0.0086)	(0.0040)
Treated×Year 2015	-0.0005	-0.0007	0.0009	0.0066	0.0065	0.0051
	(0.0048)	(0.0056)	(0.0104)	(0.0047)	(0.0047)	(0.0039)
Treated×Year 2016	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted
Treated×Year 2017	0.0434***	0.0329***	0.0718***	0.0097	-0.0086	-0.0017
	(0.0177)	(0.0133)	(0.0273)	(0.0095)	(0.0070)	(0.0039)
Treated×Year 2018	0.0535***	0.0311***	0.0823***	0.0162	0.0049	0.0074*
	(0.0190)	(0.0129)	(0.0296)	(0.0100)	(0.0072)	(0.0045)
Treated×Year 2019	0.0539***	0.0324**	0.0955***	0.0041	-0.0102	-0.0043
	(0.0221)	(0.0148)	(0.0346)	(0.0133)	(0.0087)	(0.0052)
N	261,304	261,304	139,141	122,163	239,920	239,920
$P-$value of testing joint significance of pre trends	0.5935	0.7666	0.6649	0.1003	0.2506	0.1242

Notes: In Columns (1)-(4), sample is the top level DRGs within base DRGs with multiple levels. In the last two columns, sample is the bottom level DRGs within base DRGs with multiple levels. Unit of observation is hospital/base DRG/year. Coefficients are reported in percentage point form. Other regressors for the analysis on top-coding rates include the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects. Other regressors for the analysis on bottom-coding rates include the weight of the lowest DRG, hospital-DRG fixed effects, and year fixed effects. Standard errors are clustered at the RAC/DRG level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Treatment Effects and Dynamic Treatment Effects on Top-coding rates by Hospital Type

	For-profit	Not-for-profit	Financially-healthy	Financially less-healthy	Metro	Non-metro
	dCDH (1)	dCDH (2)	dCDH (3)	dCDH (4)	dCDH (5)	dCDH (6)
Panel A						
Treatment Effect	0.0497*	0.0513***	0.0609***	0.0391**	0.0521***	0.0340*
	(0.0257)	(0.0175)	(0.0209)	(0.0177)	(0.0189)	(0.0184)
Panel B						
Treated×Year 2014	0.0189	0.0002	0.0184	-0.0027	0.0069	0.0178
	(0.0119)	(0.0095)	(0.0144)	(0.0080)	(0.0087)	(0.0226)
Treated×Year 2015	-0.0169	0.0013	-0.0073	0.0048	-0.0006	0.0085
	(0.0179)	(0.0063)	(0.0089)	(0.0084)	(0.0048)	(0.0249)
Treated×Year 2016	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted
Treated×Year 2017	0.0392	0.0477***	0.0500***	0.0401**	0.0466***	0.0244
	(0.0265)	(0.0166)	(0.0188)	(0.0172)	(0.0179)	(0.0212)
Treated×Year 2018	0.0775***	0.0495***	0.0723***	0.0339*	0.0549***	0.0413**
	(0.0303)	(0.0178)	(0.0213)	(0.0192)	(0.0193)	(0.0193)
Treated×Year 2019	0.0333	0.0574***	0.0622***	0.0433**	0.0558***	0.0389*
	(0.0321)	(0.0210)	(0.0257)	(0.0211)	(0.0225)	(0.0218)
N	40,306	148,799	131,508	116,516	210,509	50,795
$P-$value of testing joint significance of pre trends	0.1173	0.9778	0.1916	0.6670	0.6594	0.7318

Notes: Unit of observation is hospital/base DRG/year. Coefficients are reported in percentage point form. For the analysis by profit status, the omitted category is public hospitals. For the analysis by financial health status and metropolitan location, there is no omitted category. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors include the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects. Standard errors are clustered at the RAC/DRG level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

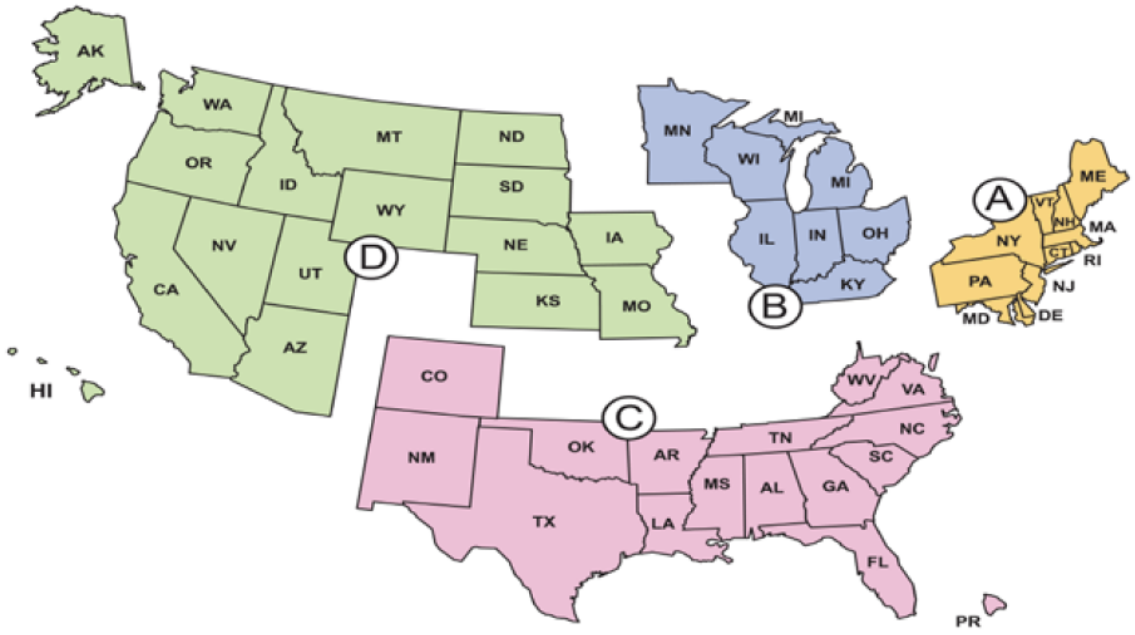


Figure A1: Map of Recovery Auditor Regions

Notes: The following lists the recovery audit contractors for each region (assigned in 2010): Performant Recovery for Region A, CGI for Region B, Connolly for Region C, and HealthData Insights (HDI) for Region D.

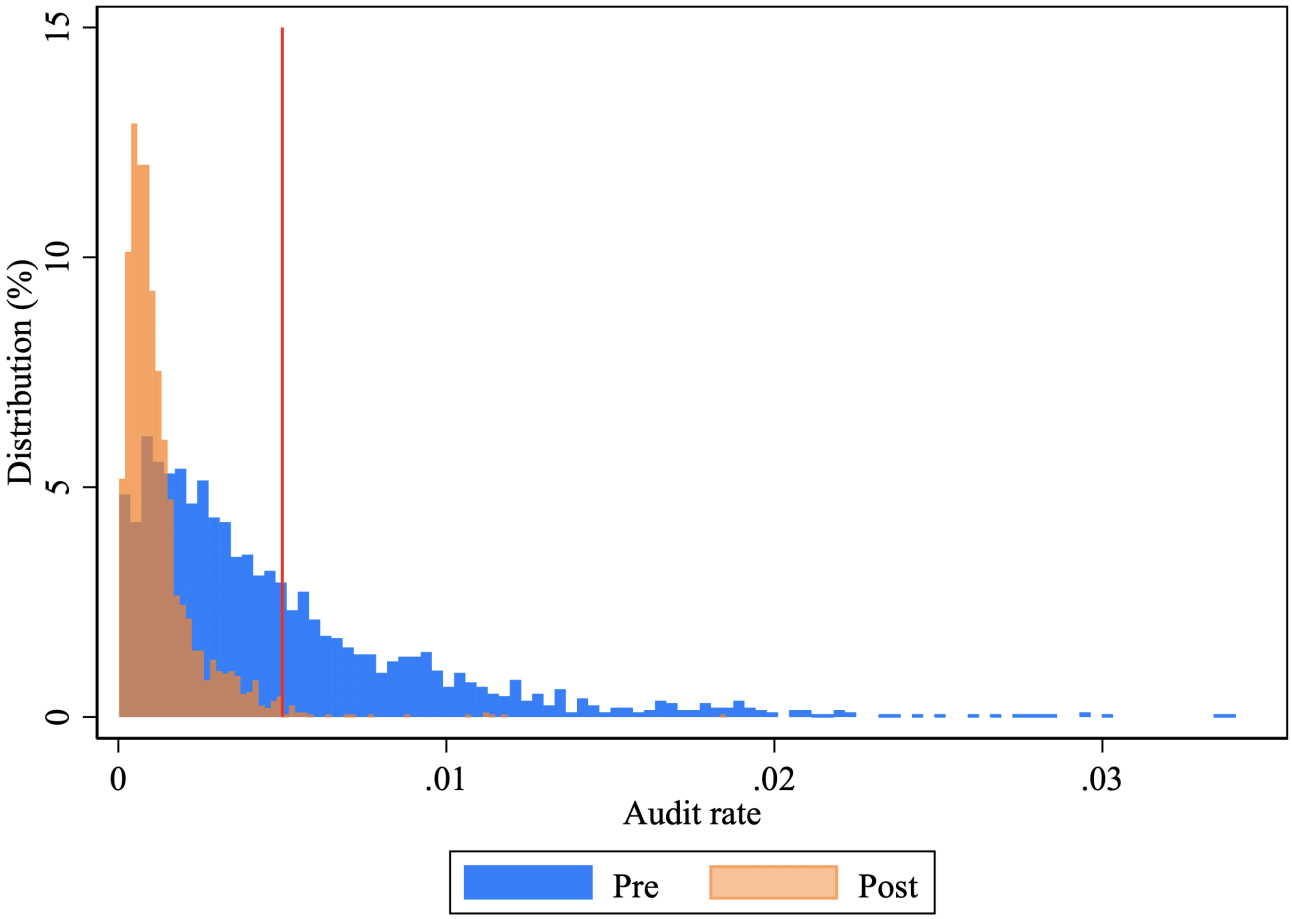
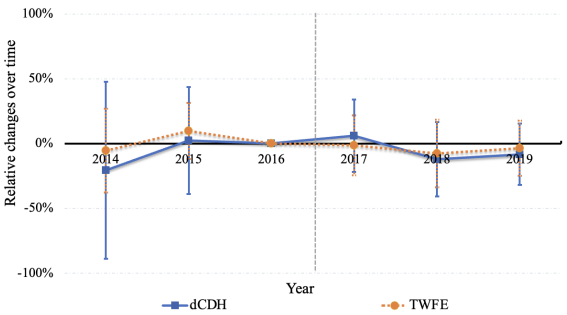
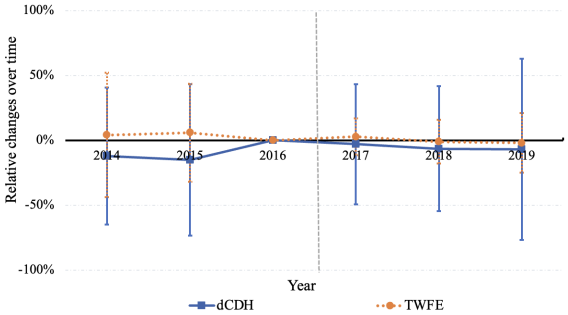


Figure A2: Frequency Distribution of Average Audit Rates per Hospital, Before and After Policy Change

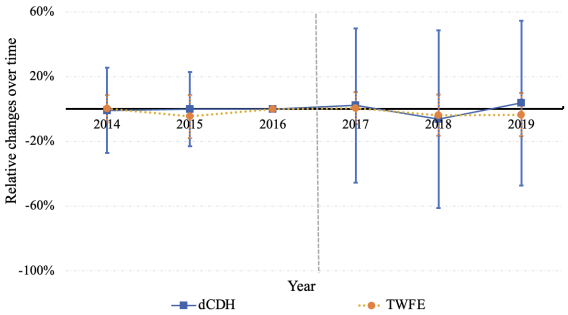
Notes: We show the difference in frequency distribution of average audit rates per hospital between pre- and post-policy change. We obtain the average audit rate per hospital across time during the pre-treatment and post-treatment period, respectively. The vertical line denotes the ADR limit, 0.5%. We trim outliers above 3.5%.



(a) Average length of stay



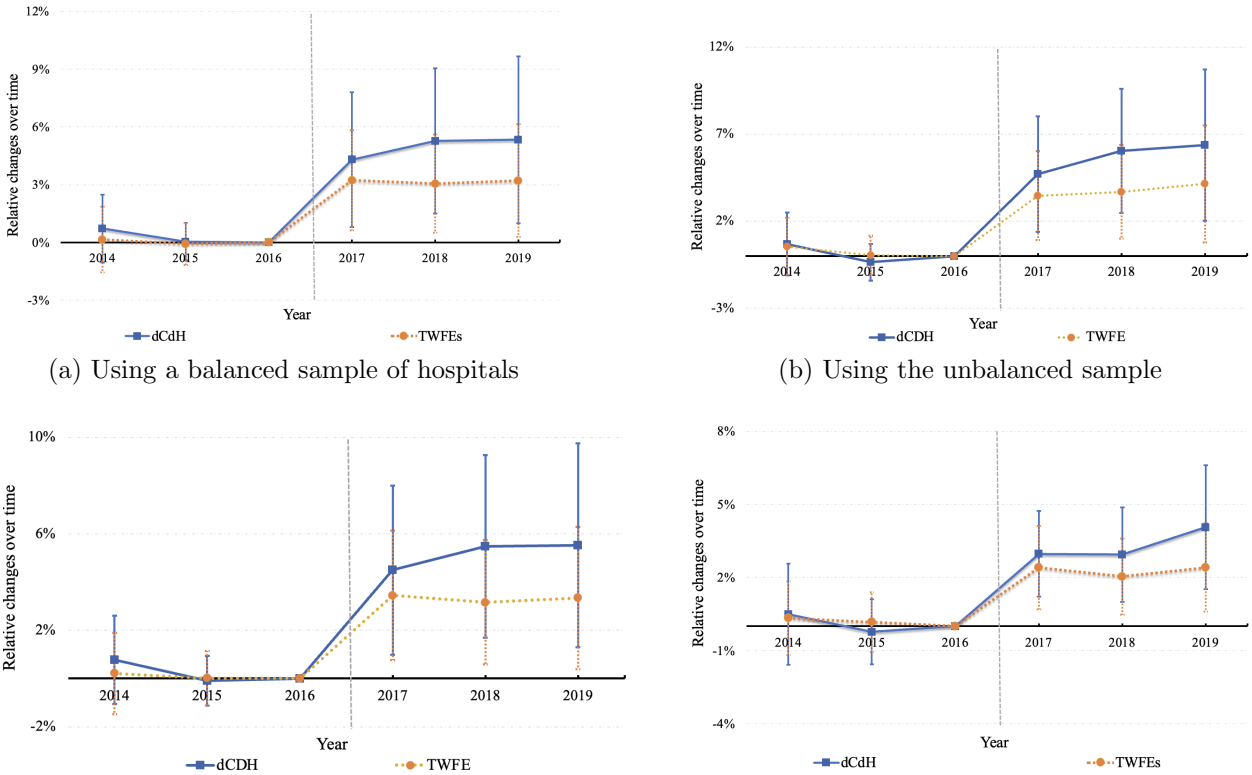
(b) Number of diagnoses



(c) Average Charlson Comorbidity Index

Figure A3: Event Studies for Additional Outcomes

Notes: The figures show the event study coefficients (estimated by dCDH and TWFEs) for other outcome measures. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.



(a) Using a balanced sample of hospitals

(b) Using the unbalanced sample

(c) Including hospital controls

(d) Without sampling weights

Figure A4: Effect of Reduced Audit Pressure on Top-Coding Rates, Robustness Checks

Notes: The figures show the event study coefficients (estimated by dCDH and TWFEs) using different samples or alternative specifications. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Section 6.6 provides more details on the sample for each figure. For all the four figures, other regressors include the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects. For the specification in (c), additional hospital controls are included: logged bed size, logged net patient revenues, logged non-medicare bad debt expense, logged uncompensated care costs, and year fixed effects interacted with for-profit and not-for-profit hospital indicators, respectively. For the specification in (d), we do not use sample weights.

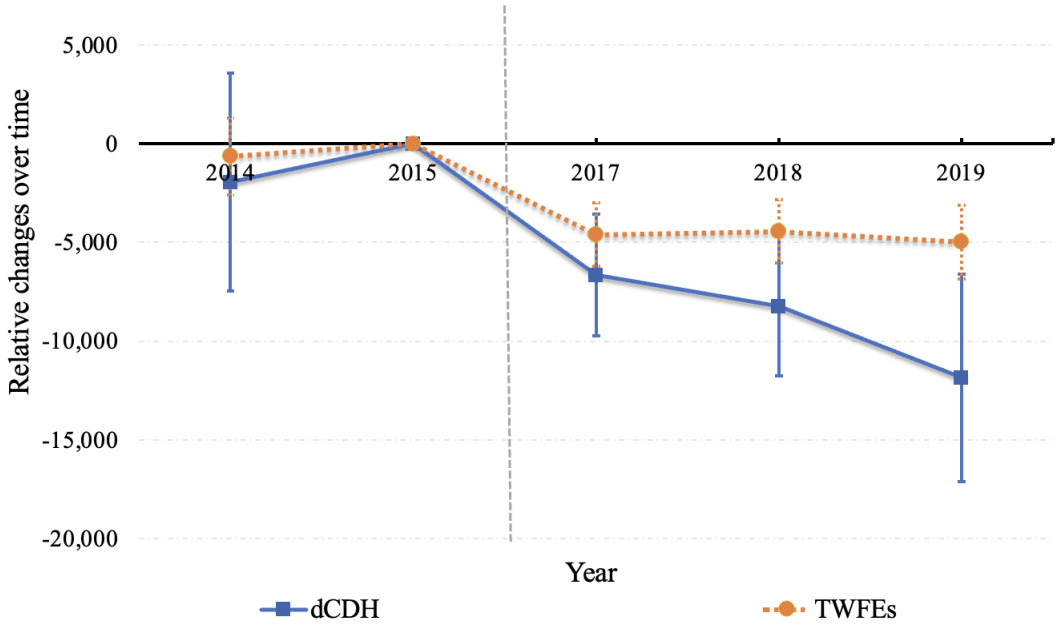


Figure A5: Effect of Scale-back of the RAP on Reclaimed Overpayments

Notes: The figure shows the event study coefficients (estimated by dCDH and TWFEs) for the analysis on reclaimed overpayments. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.

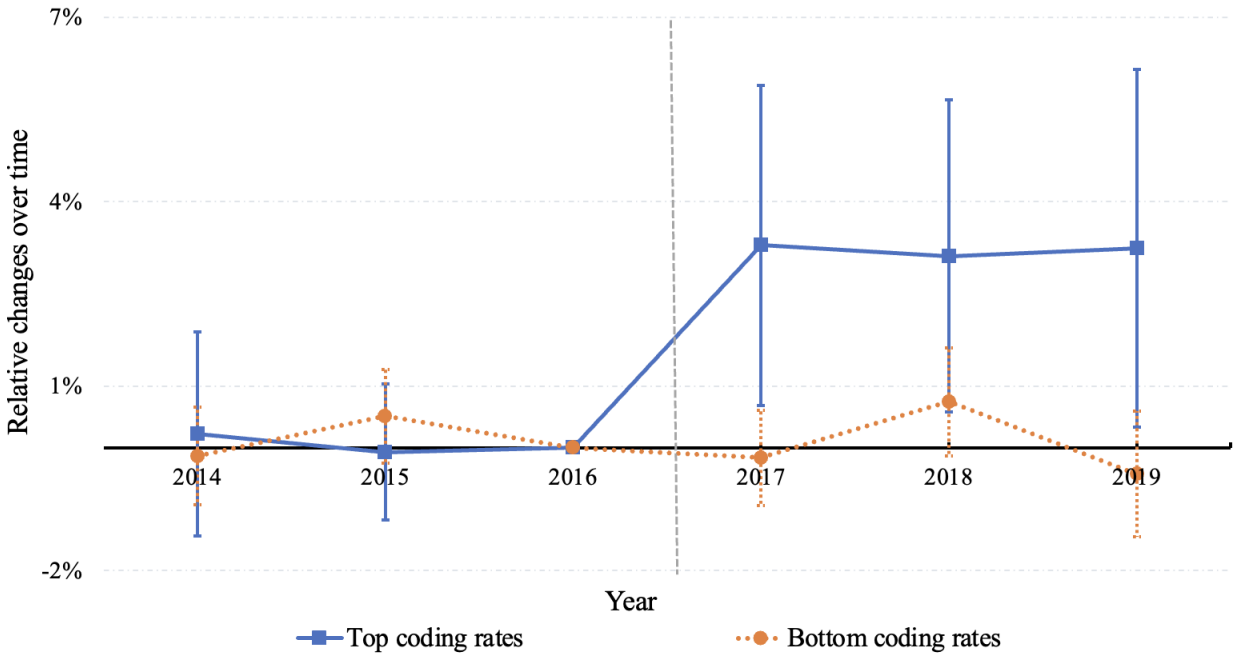
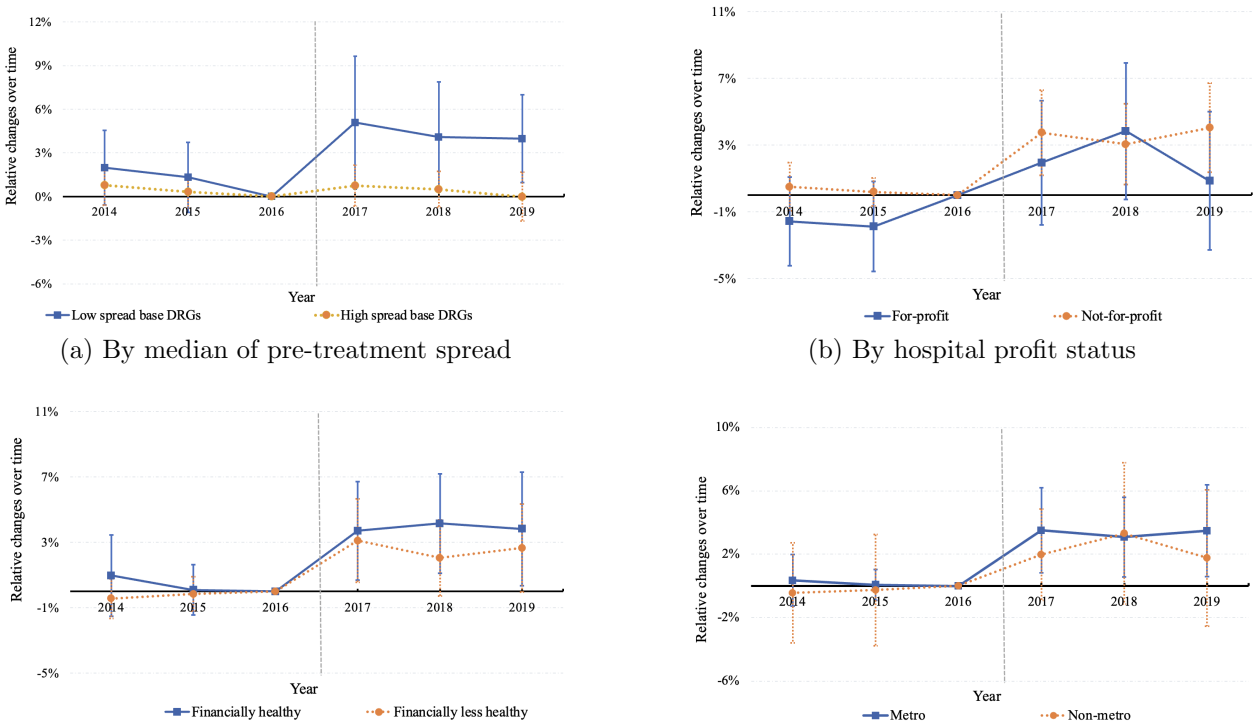


Figure A6: Effect of Reduced Audit Pressure on Top- and Bottom-Coding Rates (TWFEs Estimates)

Notes: The figure shows the event study coefficients (estimated by TWFEs). Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top (bottom) level DRGs within base DRGs with multiple levels for the analysis on the top (bottom) coding rate. Other regressors in the specification for top coding rates are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects. Other regressors in the specification for bottom coding rates are the weight of the lowest DRG, hospital-DRG fixed effects, and year fixed effects.



(a) By median of pre-treatment spread

(b) By hospital profit status

(c) By hospital financial health

(d) By metropolitan status

Figure A7: Effect of Reduced Audit Pressure on Top-Coding Rates (TWFEs Estimates), by DRG and hospital type

Notes: The figures show the event study coefficients (estimated by TWFEs) by DRG and hospital type. High-spread (Low-spread) base DRGs refer to those with spread greater (smaller) than the median of the pre-treatment spread. Financially healthy hospitals are those whose debe-asset ratio is below the pre-treatment median, and financially less healthy hospitals include the remaining hospitals. The lines report the coefficients for DRGs in the treatment group interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the RAC-DRG level. Unit of observation is hospital/DRG/year. Sample is the top level DRGs within base DRGs with multiple levels. Other regressors are the weight of the focal DRG, the DRG weight at the next lower level, hospital-DRG fixed effects, and year fixed effects.