Monitoring Pressure and Billing Practices: Evidence from Medicare Recovery Audits

Jianjing Lin*
University of Massachusetts Amherst

Juan Pantano†
The University of Hong Kong

November 1, 2025

Abstract

Monitoring has been widely used to ensure accountability and proper use of resources. However, excessive oversight may cause substantial social costs, calling for a scale-back in monitoring programs. We exploit a dramatic reduction in audit probability and study how hospitals respond in their billing practices when submitting inpatient claims to Medicare. A common approach would be to examine differential claim coding responses across diagnostic groups given varying reductions in audit rates at this level of analysis. However, obtaining unbiased estimates requires no treatment interference across diagnostic groups, and this requirement may not hold, due to potential inter-dependence in coding decisions across these groups within a hospital. We propose a simple model to formalize the idea and empirically assess the presence of spillovers across diagnostic groups. With the evidence of within-hospital across-diagnostic groups externalities, we evaluate the effect of reduced audit pressure at a higher level of aggregation, the hospital level, and find that hospitals code more aggressively as audit rates drop. Our results suggest that for every dollar Medicare saves on reduced monitoring, it may suffer a loss of $17, among which 51% comes from hospitals' more intense coding. It might require thoughtful policy design to balance the gain from lowering social costs of unnecessary audits, against the potentially increased illicit activity arising from lax oversight.

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

*Email: jianjinglin@umass.edu.
†Corresponding Author; email: jpanta.econ@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. A 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 costs and benefits of these policy changes. 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 consider 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 contract 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.

The RAP imposed substantial audit pressure to providers during the first few years since its implementation, resulting in numerous complaints about the inefficiency and burdens it created. 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 submitted by hospitals for reimbursement and paid by Medicare. We call it *ADR limits*, and effectively imposes a cap of *0.5%* on the audit rate RACs. This effectively reduced the proportion of claims that could be reviewed by RACs for each hospital. Our empirical strategy, described more fully below, is built around this regulatory change.

In this paper, we 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.[1] 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.

To identify the effect of audit pressure on billing or coding practices, one approach is to exploit the differential relative declines in audit rates across disease groups—specifically diagnostic related groups (DRGs), each represented by a *code*—as exogenous variation in the incentives for inappropriate coding, similar to the approach in the hospital upcoding literature (Silverman and Skinner, 2004; Dafny, 2005; Ganju et al., 2021). These studies examine how much more likely a claim is to be coded in a top-tier DRG (among all patients from the same *base DRG*) when the spread between the top-tier reimbursement weight and the lower tier weights increases.[2]

One implicit assumption of this approach is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980), which requires no interference across units. Specifically in our case,

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

[2] Base DRGs can be viewed as a way of grouping similar patients on the basis of a shared primary disease or procedure. Within a base DRG, there could be multiple DRGs, each representing a particular level of severity/complexity.

SUTVA would translate into an assumption that the coding decision for a claim in a given DRG is not affected by changing incentives (e.g. the reimbursement weight or, in our context, the audit pressure) in another DRG. This assumption could be violated as there may be spillovers in coding decisions across DRGs within a hospital. For instance, due to limited coding resources or incentive mis-alignment between different parties in a hospital, it may be infeasible to achieve the optimal "upcoding" in every base DRG. When a hospital faces some organization-wide constraints in upcoding, it may maximize the upcoding opportunity in *some* base DRGs while reducing the extent to which it upcodes patients in other base DRGs. In this case, the incidence of reporting a particular DRG due to the change in audit rates could be affected by what would happen to the others.

To formalize this idea, we develop a simple model where hospitals aim to maximize the expected profit from top coding across all base DRGs, subject to a hospital-wide upcoding limit. We apply this limit, seeking to capture the fact that hospitals may not be able to maximize upcoding opportunity in every base DRG. In fact, hospitals may have incentive to keep track of the incidence of overall upcoding and limit it to a certain level, particularly given regulators' increasing monitoring and law enforcement efforts to reduce improper payments in health care. With the implementation of ADR limits, hospitals may re-adjust the level of upcoding for each base DRG. We find that the optimal level of upcoding would be inter-correlated across base DRGs when the hospital-wide constraint is binding. The rationale is that the binding organization-wide constraint creates a linkage in coding decisions across DRGs. When the aggregate level of upcoding is high enough, hospitals have to substitute the less profitable upcoding with the more profitable one in re-balancing the top-coding portfolio, resulting in within-hospital across-DRG "externalities" in coding practices.

Violation of SUTVA could lead to biased estimates of the treatment effect (Miguel and Kremer, 2004; Butts, 2021). Moreover, the direction of bias is ambiguous in our context. To test the presence of such spillovers empirically, we exploit the "multilevel treatments" embedded in the data and define both, treatment at the *hospital* and *DRG* levels. For the hospital-level treatment definition, we leverage the ADR limit and the differential exposure to this cap given how much RACs audited each hospital before the cap is imposed. In particular, we define a treated hospital as one whose *overall* average audit rate prior to the policy change was above the ADR limit, 0.5%. We then define the timing of treatment as the point in time when the ADR limit was implemented. We classify the remaining hospitals into the control group.

The spillover analysis focuses on a set of DRGs that were not directly impacted by the

implementation of ADR limits but may be indirectly affected by other DRGs that were. Thus, for the treatment definition at the DRG level, we define a *control* DRG as one whose pre-treatment audit rate was always zero. We expect that the policy has minimal effects on these DRGs. We then compare the top-coding rate among control DRGs in treated hospitals with that among the same set of (control) DRGs in control hospitals before and after the policy change. The idea is that, absent spillovers across DRGs, one would not expect significant differences in evolution of the top-coding rate across these two sets of control DRGs before and after the ADR limit. Using both the two-way fixed effects model (TWFEs) and a modern DiD estimator developed by Callaway and Sant'Anna (2021) (CSDiD), we find evidence of spillovers: control base DRGs in treated hospitals report a smaller proportion of patients with top-tier DRGs codes relative to the same control base DRGs in control hospitals after the policy change.

To reduce the potential bias due to spillovers, we estimate the effect of audit probability on coding practices at the *hospital* level, sharing a similar idea in Miguel and Kremer (2004). To achieve this, we first apply a DiD research design, comparing the overall top-coding rate in treated hospitals relative to control hospitals from before to after the policy change. In addition to examining the overall policy effect, we also estimate the causal effect of audit pressure on top coding rates using the continuous variation in audit rates across hospitals over time. Concerning that the within-hospital variation in audit pressure might be correlated with time-varying unobserved hospital heterogeneity, we instrument hospital audit rates with the exogenous policy variation from our DiD analysis. The main findings are similar from both approaches, providing suggestive evidence of more aggressive coding by hospitals as audit pressure drops. Moreover, the increase in reported top codes concentrates in for-profit hospitals, similar to a finding in prior literature.

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 $17, among which over half 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 contributes to three 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 pro-

curement (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).[3] 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). We study how hospitals' coding intensity changes following a large scale-back in the monitoring program, where the scale back was spurred precisely due to these burden/social cost concerns. We find that hospitals that previously underwent greater audit pressure tend to code relatively more aggressively post policy change. Our findings suggest that policy-makers might face a tradeoff between reducing administrative burden providers face from the monitoring program and the potential upticks in wasteful spending that may re-emerge when scaling back those programs.

Second, our paper also contributes to the broad literature of spillover effects and treatment interference. Many of the recent studies explore methods to test the presence of spillovers and to generate unbiased estimates of treatment effects under interference (Clarke, 2017; Butts, 2021; Li et al., 2025). We examine spillovers in coding practices across disease groups within a hospital, exploiting the differential exposures to a policy shock across DRGs and across hospitals. Our findings offer new empirical insights on the propagation of treatments in the context of health care, which has previously been under-studied.

Finally, our paper also complements the literature that investigates how healthcare providers respond to regulators' non-financial policy changes, such as stricter requirements for documenting care provision and increasing monitoring and detection of fraud and abuse. Sacarny (2018) finds 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 but resulted in compliance costs; in response, providers adopt new information technology to identify potentially medically unnecessary care. League (2024) examines the impacts of administrative burden on provider

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

behavior and market outcomes and finds that increased claim denial rates result in more healthcare consolidation and higher healthcare spending. Leder-Luis (2025) and Howard and McCarthy (2021) find 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 discusses the situations where SUTVA could be violated and proposes a theoretical model to formalize the idea. Section 4 describes the datasets and reports summary statistics. Section 5 presents the empirical strategy for examining spillovers and discusses the findings. Section 6 describes the empirical strategy for examining changes in coding practices and report the results. 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, a level or a severity subclass 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).[4] 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.[5]

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 Program (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.[6] 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

[4]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.

[5]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.

[6]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.

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.[7] 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.[8] 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.[9] 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.

The imposition of ADR limits mainly arose from hospitals' complaints about the increasing 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.[10] 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

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

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

[9] 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, and so on (Recovery Audit Program, 2015).

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

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.[11] 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."[12] 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 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.[13]

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

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

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

3 Possible Spillovers in Coding Practices

3.1 Violation of SUTVA in coding across disease groups

To study how hospitals vary their coding intensity in response to the change in audit pressure, one approach is to look at variation in the top-tier coding rate across base DRGs that experienced different declines in top-tier audit probability after the imposition of ADR limits. This is similar to the approach in the hospital upcoding literature (Dafny, 2005; Li, 2014; Ganju et al., 2021) that exploits exogenous changes in reimbursement spread (top- vs bottom-tier) across base DRGs. These studies typically compare the change in top-coding rates across base DRGs and hospitals before and after some policy shocks to identify how hospitals respond in coding decisions.

However, to obtain an unbiased estimate of treatment effects, some form of the Stable Unit Treatment Value Assumption (SUTVA) must be invoked (Rubin, 1980), which requires the potential outcome of a top-tier DRG—the incidence of reporting a top code—not to depend on the treatment status—the change in audit rates—of other (top-tier) DRGs. However, this assumption may not necessarily hold because the optimal response in tier coding decisions could be inter-dependent across base DRGs within a hospital. Consider, for simplicity, a hospital that provides care to patients in two base DRGs: A and B, each with multiple severity subclasses. The imposition of ADR limits may force hospitals to re-adjust the top coding in each base DRG. Assume that the audit rate for the top code of B dropped whereas the audit rate of A stayed the same after the policy change. Under SUTVA, the change in reported top codes of A is not affected by the change in audit rates for the top code of B. However, this assumption could be violated if, for instance, the hospital faces some organization-wide constraints, due to limited coding resources or agency conflicts in the hospital, limiting the hospital's capability of achieving the optimal "upcoding" in all base DRGs. If it maximizes the upcoding opportunity in B, the hospital may have to reduce the extent to which it upcodes the patients in A when the hospital-level upcoding has reached the limit.[14] Thus, the adjustment to the upcoding level of A post policy change would be affected by what has happened to B.

Violation of SUTVA could lead to inaccurate estimates of treatment effects. As pointed out by Butts (2021), if the potential outcome of untreated units is (indirectly) affected by the treatment, the evolution of their outcomes post treatment do not serve as a valid coun-

[14] With lower audit pressure in the top code of B post policy change, the hospital may prioritize the upcoding opportunity in B to maximize the overall profits.

terfactual trend for treated units, even with pre-treatment parallel trends. In the presence of spillovers across units, comparing the outcomes between treatment and control groups may not reflect the true treatment effect.

Moreover, the direction of bias is ambiguous in our context, where the inter-dependence in coding decisions may involve all DRGs within a hospital, resulting in spillovers not only from treated onto control DRGs but also between treated DRGs. To illustrate the idea, consider a modification to the example above: the hospital also treats patients from another two multi-tier base DRGs: C and D. Further assume that (i) the post-treatment audit rate for the top code in D decreased and the post-treatment audit for the top code in C remained unchanged; and (ii) the decline in audit rates for the top code in B is greater than that for the top code in D. A simple way to estimate the effect of audit pressure on upcoding is to compare the differential change in top-coding rates between "treated" DRGs (the top codes in B and D) and "control" DRGs (the top codes in A and C) over time. If the hospital takes away some of the upcoding in A and C *and* increases that in B and D, then the magnitude of the negative effect would be over-estimated. However, if upcoding in B is much more profitable than all other base DRGs (due to the more substantial decrease in audit rates) and the hospital chooses to forgo the upcoding opportunities in all others (A, C, and D) to maximize the upcoding benefits from B (when the overall upcoding rate is high enough), then the "flow" of upcoding not only goes from control to treated DRGs (i.e., from A and C to B) but also from one treated DRG to another (i.e., from D to B, due to the larger decline in audit rates in the top code of B than D). In this case, the ultimate direction of bias may not be as clear, depending on the interplay between the substitutions. Considering that a hospital, on average, treats patients from a large number of base DRGs, it becomes even harder to predict the exact bias of the estimated effect.

3.2 Model

We develop a simple model characterizing a hospital's decision to top-code patients from a given base DRG. For simplicity, we assume that there are only two base DRGs, $\mathcal{A}$ and $\mathcal{B}$, each with two tiers. The key implications from the model can be easily extended to a larger number of base DRGs or base DRGs with more than two levels.

For each admission within a given base DRG, we use s to denote the severity of the patient's condition, which we assume follows the uniform distribution between 0 and 1 for simplicity. 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, R, 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, lowest tier in a base DRG, the profit from treating this patient equals $R\omega^{low} - c^{low}$, where ω^{low} denotes the reimbursement weight for the bottom DRG within the given base DRG and c^{low} denotes the cost of treating this lowest-severity patient. If the patient is instead coded to the top tier, it would imply upcoding, leading to higher profit that equals $R\omega^{top} - c^{low}$, relative to the one obtained under proper payment. 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 excess amount 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 still assigns a top code to, but does *not* meet the criteria for this tier according to the medical coding requirements, i.e., $\tilde{s}^u \leq \underline{s}^u$. The expected profit the hospital receives from upcoding a patient in base DRG $\mathcal{A}$ has the following form:

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

where the subscript $\mathcal{A}$ indicates the parameters associated with base DRG $\mathcal{A}$. $\phi^{\text{AUDIT}}_{\mathcal{A}}$ refers to the probability of a *top-coded* admission from base DRG $\mathcal{A}$ being audited by RACs.[15]

[15] $\phi^{\text{AUDIT}}_{\mathcal{A}}$ 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 $\phi^{\text{AUDIT}}_{\mathcal{A}}$ denote the audit probability by RACs.

$\phi^u_{\mathcal{A}}$ denotes the probability that, in the event of an audit, auditors detect upcoding for the given admission. We assume that $\phi^u_{\mathcal{A}}(\cdot)$ 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 $\phi^u_{\mathcal{A}}$ 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. That is, $\frac{\partial^2 \phi^u_{\mathcal{A}}(\cdot)}{\partial |\underline{s}^u - s|^2} > 0$. 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 involves improper payments. In particular, the last term in total profit from upcoding $\int_{\tilde{s}^u_{\mathcal{A}}}^{\underline{s}^u_{\mathcal{A}}} \left[\phi^{\text{AUDIT}}_{\mathcal{A}} \times \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - s|) \times \xi^u\right] ds$, accounts for the possible costs from a potential audit. The expected profit from upcoding a patient from base DRG $\mathcal{B}$ shares a similar form.

Assume that the hospital seeks to maximize the total expected profit across the two base DRGs, subject to a hospital-wide upcoding limit $\bar{\kappa}$, i.e.,

$$\max_{\{\tilde{s}^u_{\mathcal{A}}, \tilde{s}^u_{\mathcal{B}}\}} \{E[\pi_{\mathcal{A}}(\tilde{s}^u_{\mathcal{A}}) + \pi_{\mathcal{B}}(\tilde{s}^u_{\mathcal{B}})]\}$$
$$s.t.\ \underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}} + \underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}} \leq \bar{\kappa}$$

Implicitly, we have $\bar{\kappa} \leq \underline{s}^u_{\mathcal{A}} + \underline{s}^u_{\mathcal{B}}$, as $\tilde{s}^u_{\mathcal{A}} > 0$ and $\tilde{s}^u_{\mathcal{B}} > 0$.

Hospitals may fail to maximize upcoding opportunity in every base DRG for a variety of reasons. For instance, the health information technology in the hospital may be outdated or the billing staff are poorly trained, leading to missing opportunities of capturing high-revenue codes. Also, there could be misaligned incentives between administrators and physicians due to principal-agent problems, preventing hospitals from billing a higher level (Sacarny, 2018). Moreover, hospitals may set a limit on the overall, hospital-wide upcoding rate. The finding of substantially overbilling Medicare may not only result in financial penalty and reputation damage but also cause more intense monitoring and investigation later on, the process of which could be costly and stressful for providers.[16] As a result, hospitals may want to keep

[16]CMS share the audit findings across different contractors and use these findings to identify payment vulnerabilities. Thus, if a hospital is caught for improper billing by one audit contractor, this finding will be known to all other Medicare contractors, leading to potentially heightened monitoring on this hospital by other contractors. Moreover, CMS issue the Program for Evaluating Payment Patterns Electronic Reports (https://pepper.cbrpepper.org/index.html) from time to time, which deliver provider-specific Medicare data statistics for areas that are identified by CMS as high risk for improper payments due to billing/coding issues. The report compares the provider's billing practices with peers in the state/nation, alerting them to

track of the incidence of improper coding hospital-wide, aggregating across all base DRGs and ensure that it does not go beyond a certain limit, even when there are, ignoring this aggregate constraint, incentives in each base DRG to extract additional profit through more upcoding.

The implementation of ADR limits implies changes in $\phi_{\mathcal{A}}^{\text{AUDIT}}$ and/or $\phi_{\mathcal{B}}^{\text{AUDIT}}$. For simplicity, we only let $\phi_{\mathcal{B}}^{\text{AUDIT}}$ change, assuming that claims reported with the top-tier DRG in $\mathcal{B}$ were much more heavily audited than those in $\mathcal{A}$ prior to the policy change. Thus, imposing the ADR limit (an institute-based calculation) would affect the change in the probability of auditing the top-coded cases in $\mathcal{B}$ more substantially than that in $\mathcal{A}$. We consider the effect of a decline in $\phi_{\mathcal{B}}^{\text{AUDIT}}$ on hospitals' coding intensity in each base DRG and obtain the following proposition:

Proposition 1 As the audit rate of the top code in one base DRG goes down, we expect two possible cases:

(i) When the hospital-wide constraint is binding, i.e., the overall upcoding rate of the hospital reaches $\bar{\kappa}$, the effect of varying audit rates for the top code in base DRG $\mathcal{B}$ would spillover onto the coding intensity decision in base DRG $\mathcal{A}$. Specifically, the proportion of improperly top-coded cases *rises* in base DRG $\mathcal{B}$ and *drops* in base DRG $\mathcal{A}$.

(ii) When the hospital-wide constraint is not binding, the upcoding rate in base DRG $\mathcal{A}$ does not vary with $\phi_{\mathcal{B}}^{\text{AUDIT}}$ whereas a greater share of patients in the bottom tier of base DRG $\mathcal{B}$ will be assigned the top code.

Proof. See Appendix I.

3.3 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 occur 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 possible that hospitals' strategic response in coding due to the introduction of ADR limits

potential over-utilization and potential payment errors. Finally, hospitals may be cautious about excessive overbilling, due to concerns about litigation issues from whistleblowers (Leder-Luis, 2025).

may have changed the composition of claims submitted. This could affect the claim denial risk and hence the probability of facing post-payment audits. Thus, our results may also pick up the effect of the change in pre-payment claim denial risk as a result of implementing ADR limits.

Second, an alternative mechanism that is not captured in our model but could lead to similar empirical implications is the potential *undercoding* of legitimate top-tier claims prior to the policy change. Hospitals might previously choose to undercode or downcode these claims to avoid costly audits, even when they know that they would ultimately prevail in those audits. Then, hospitals might reduce this type of defensive coding after the ADR limit was imposed.[17] 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. Our model here abstracts away from the undercoding mechanism.

Third, we assume that the hospital-wide aggregate constraint on upcoding, $\bar{\kappa}$, does not vary with the imposition of the ADR limit, but hospitals may relax this constraint when audit pressure is reduced. If this is the case, we expect less spillover from one top DRG onto another, considering that the increased limit on upcoding would be less likely to be met. However, it may not completely rule out the inter-dependency in coding practices across DRGs, if the expansion of the constraint is limited, especially in the short run. This would not be unlikely, if the constraint arose due to some "exogenous" factors mentioned above, such as limited resources available for billing and other frictions hospitals face in achieving more aggressive coding. 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. Note that there could be various ways in which the top-coding decision is inter-dependent across base DRGs, and our model here captures only one possible mechanism, where this decision is made in a relatively independent way within a base DRG but the entity-wide binding constraint creates a linkage in this decision across base DRGs.

[17] To model this case, $\tilde{s}^u$ (with $\tilde{s}^u > \underline{s}^u$) would denote 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 or decreased undercoding with less monitoring.

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 the audits, such as the date selected for review, the reason for review, final determinations, and the amounts of recovered payments. From this dataset, we calculate the number of audits at two levels. In testing the presence of spillovers across DRGs, we rely on the audit rate at the hospital/DRG level. In examining the overall effect of reduced monitoring on coding practices, we use the audit rate at the hospital level. Since we are interested in changes in coding intensity in response to the policy change, we focus on base DRGs with multiple levels in the analysis at the hospital/DRG level and obtain the hospital audit rate only based on the top-tier DRGs in these base DRGs for the analysis at the hospital level. We define the audit rate at both levels in more detail below.

Our second dataset is the Inpatient Utilization and Payment Public Use File (Inpatient PUF) published by CMS.[18] We extract the hospital inpatient discharges per DRG per hospital from this dataset.[19] Our third dataset comes from the Healthcare Cost Report Information System, which includes essential provider information that CMS require Medicare-certified institutional providers to report annually.[20] We obtain the total number of Medicare inpatient discharges in a hospital from this dataset. Combining the audit microdata and discharge counts, we construct the audit rate at the hospital/DRG level, defined as the total number of claims selected for review by RACs among all the discharges within the hospital-DRG-year cell. Similarly, we obtain the audit rate for each hospital by dividing the total number of audits for the hospital by the total number of discharges in this hospital.

We divide *hospitals* into treatment and control groups based on the audit rate before the policy change. Specifically, we define a hospital as treated, if its average hospital audit rate is strictly greater than 0.5%, the ADR limit, prior to its implementation in 2016. Presumably these are the hospitals that are "treated" in the sense that they must adapt to the policy

[18]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.

[19]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).

[20]The data is available at https://data.cms.gov/provider-compliance/cost-report/hospital-provider-cost-report.

change given the aggregate audit rate on them is forced to decline. We calculate the pre-treatment average audit rate by averaging the hospital audit rate across the years 2014 and 2015. We group the remaining hospitals as control hospitals. Such a division results in 17.6% of hospitals in the treatment group and 82.4% in the control group.[21]

Our spillover analysis will only involve a comparison of DRGs that were not directly affected by the policy, but may be indirectly affected by other DRGs that were. Therefore, we also define a DRG as a *control DRG* using the following steps. First, we identify the DRGs within the treated hospitals whose audit rates were always zero before the policy change. These will be the control DRGs in treated hospitals. We expect that the ADR limit has little impact on the audit pressure of these DRGs. Second, we focus on this set of DRGs and explore what happened to their audits rates in control hospitals. We keep in our set of control DRGs, the DRGs that exhibit zero audit rates before the policy, in both treated and control hospitals. Keep in mind then, the distinct definitions of treatment at the hospital and DRG levels. In particular, for our subsequent analysis it is important to note that a DRG that is defined as a control DRG, could be so in the context of a treated hospital or a control hospital.

Our fourth dataset comes from the 5% Inpatient Standard Analytic File from the Limited 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 measures, the top coding rate, at both hospital/DRG and hospital levels from this dataset. Specifically, the top coding rate at the hospital/DRG level is defined as the fraction of patients assigned the top-tier DRG within a base DRG per hospital per year. This is similar to the outcome measure used by most of the literature that studies how financial incentives affect hospital upcoding behavior (Silverman and Skinner, 2004; Dafny, 2005; Li, 2014; Ganju et al., 2021). We obtain the top coding rate at the hospital level, by dividing the number of discharges reported with top-tier DRGs in multi-tier base DRGs by the total discharges from these base DRGs.[22] Finally, we complement these datasets with the American Hospital Association (AHA) Annual Survey, where we obtain a rich set of hospital characteristics, including bed size, total births, total outpatient visits, profit status, system affiliation status, an indicator for whether a hospital is a teaching hospital, and whether a hospital has a strong integration with physicians.

Our analysis focuses on the years between 2014 and 2019 for the following reasons. First,

[21] We obtain this ratio by dividing the number of treated hospitals by the total, both reported in the last row of Panel 2 in Table 1.

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

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).[23] Second, we drop the data after the year 2019 because of the audit pause due to the pandemic of coronavirus disease 2019.

Sample construction. There are 4,661 hospitals 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 3,759 hospitals each year.

4.2 Summary statistics

Figure 1 displays the average hospital audit rates for the entire sample and separately for treated and control hospitals over time. Specifically, the average audit rate is a weighted average across all the hospitals, with the share of claims from a given hospital relative to the total claims as the hospital weight. Before 2016, it was 0.28% in 2014 and went up to 0.44% in 2015.[24] The average audit rate plummeted immediately after 2015, below the cap, 0.5%, in every year after the imposition of the limit.[25] This figure also shows the trend in the average audit rates by hospital treatment status. Visually, the relative decline in average audit rates is much more significant in treated hospitals than control hospitals.[26]

As the ADR limit was calculated at the hospital level, we show, in Appendix Figure A2, the frequency distribution of the hospital average audit rates among hospitals with strictly positive audit rates before and after the policy change, respectively. Specifically, we obtain the average audit rate for each hospital separately during the 2014-2015 and 2017-2019 years and plot the frequency distribution. Prior to the policy change, there were approximately only 60.5% of hospitals with average audit rates below 0.5% among those

[23]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.

[24]Appendix Table A1 provides more information.

[25]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.

[26]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 and semi-automated reviews, both of which require ADR. We expect that the results of excluding automated reviews would be very similar because this type of reviews accounts for less than 2.6% of the total investigated claims.

with non-zero average audit rates, whereas this rate went up to 99.7% after the ADR limit was introduced.[27]

Table 1 shows the summary statistics for the key variables during the pre-treatment periods. The top panel reports the summary statistics based on the sample for testing spillovers, which is implemented at the hospital/DRG level. The average top-coding rate is 35.3% based on the entire sample, with a slightly lower rate in treated than control hospitals. In each hospital, there are, on average, 63 top-tier DRGs whose audit rate was always zero before the implementation of ADR limits. This number is smaller in treated hospitals, as in principle there would be a greater number of DRGs whose post-treatment audit rates are "forced" to go down.[28] Panel 2 reports the summary statistics for the sample studying coding practices at the hospital level. The average proportion of patients assigned top-tier DRGs per hospital is 38.8% based on the entire sample, with a slightly higher rate in control hospitals. As expected given our definition of a "treated" hospital, the hospital audit rate before the imposition of ADR limits was higher among treated hospitals, at 0.96%, over ten times that among control hospitals. There are almost 70 top-tier DRGs per hospital, with little difference between treated and control hospitals. Appendix Table A2 presents the summary statistics for select hospital characteristics.[29] Treated hospitals are more likely to be larger, not-for-profit or teaching hospitals. They are also more likely to be part of a healthcare system or have a strong integration with physicians.[30]

5 Empirical Analysis: Testing Spillovers

5.1 Empirical strategy

An implication from our model suggests that the reduction on the audit probability for a top-tier code in a given base DRG could spillover onto the top coding decisions in another base DRG. When a hospital considers coding intensity for the hospital as a whole, the decision for

[27]There are still hospitals with audit rates above 0.5% after the policy change, as CMS granted an exception that RACs could request additional documents above the cap in certain cases, such as previously unusually high claim denial rates from the hospitals.

[28]The number of hospitals here is greater than that mentioned in Section 4.1 because our sample of hospitals is not balanced. As a robustness check, we rerun our main specifications using the balanced sample. The results are similar. We provide more information in Section 5.3.

[29]The number of hospitals here is smaller than that in Table 1, due to loss of observations in merging with the AHA annual survey.

[30]We define a strong hospital-physician integration if a hospital forms one of the following types of alliances with physicians: integrated salary model, equity model, or foundations. Under these classes of alliances, hospitals either hire physicians as employees or have some ownership stake in physician practices (Cuellar and Gertler, 2006).

each base DRG may no longer be independent. To assess the presence of these spillovers, we apply a DiD framework to a sample of control DRGs from both treated and control hospitals. In particular, we compare the evolution of the top-coding rate among control DRGs in treated hospitals with that among the same set of control DRGs in control hospitals before and after the policy change, similar to the idea in Sinclair et al. (2012). In the jargon of the literature on spillovers, unlike the control DRGs in treated hospitals, the control DRGs in control hospitals would be typically referred to as "pure control" DRGs. Specifically, we estimate the following regression equation using all control DRGs: the "control-in-treated" DRGs and the "pure control" DRGs:

$$\texttt{TopCodeRate}_{jdt} = \beta \times \mathbb{1}\{t > 2015\} \times \text{Treat}_j + \gamma_{jd} + \lambda_t + \varepsilon_{jdt}, \tag{2}$$

where $\texttt{TopCodeRate}_{jdt}$ denotes percent of claims that were assigned to the top tier reimbursement code amongst claims in base DRG d in hospital j and year t. $\text{Treat}_j = 1$ if hospital j is a treated hospital. Since the ADR limit became effective in 2016, we expect the treatment to start to take effect in this year. We also include hospital-base DRG fixed effects, γ_{jd}, capturing the time-invariant unobserved heterogeneity at this level, such as the pre-existing differences in patient characteristics, coding patterns, healthcare practices, etc. λ_t denotes year fixed effects, capturing, for instance, the changes in patient population or Medicare audit strategies at the national level over time. ε_{jdt} is the error term.

The parameter β captures the difference between the average change in top-coding rate post-2015 relative to pre-2015 among control DRGs in treated hospitals and the average change in this rate among "pure" control DRGs in control hospitals over the same period.[31] The estimated β then informs us about the presence/absence of spillovers. If the change in top tier coding for a given DRG in response to the policy shock is only affected by the variation in audit probability for that DRG, ceteris paribus, we expect that $\beta = 0$, i.e., there are no systematic differences in the evolution of coding intensity between the control DRGs in treated hospital and the same set of DRGs in control hospitals over time. In contrast, a non-zero β, in particular $\beta < 0$, may imply that the variation in coding intensity owing to the policy shock may be simultaneously determined across base DRGs within a hospital. Hospitals that increase top-coding in response to reduced monitoring may prioritize top-

[31] An implicit assumption is that the audit rate for non-top DRGs is zero. We believe that this assumption is reasonable, as one of the most common reasons for DRG audits is DRG validation, which mainly targets at high-reimbursement codes (https://hiacode.com/blog/comprehensive-coding-audit-vs-drg-validation-audit. Also, during our sample period, the average audit rate of bottom-tier DRGs is very close to zero.

coding among DRGs that experience a lower audit probability after the policy change. If hospitals are confronted with a binding cap on the incidence of overall upcoding, they may choose to "forgo" upcoding opportunities in certain DRGs, particularly those in which the audit rate did not decrease after the policy.

Note that implicitly we use `TopCodeRate` 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. An identifying assumption is that levels of sickness in Medicare inpatient population or thresholds for the top codes do not change over this period or that if any does, the changes at the DRG level are not systematically different for patients in "control-in-treated" DRGs relative to patients in "pure control" DRGs.

To investigate whether there were differential pre-treatment trends in the top-coding rate for control DRGs in treated hospitals relative to the same set of DRGs in control hospitals, we extend the specification in Equation (2) using an event study research design:

$$\texttt{TopCodeRate}_{jdt} = \sum_{\tau=2014,\, \tau\neq 2015}^{2019} \beta^{\tau} \times \mathbb{1}\{t = \tau\} \times \text{Treat}_j + \gamma_{jd} + \lambda_t + \varepsilon_{jdt}. \tag{3}$$

The specification above is almost the same as that in Equation (2) except that we replace the key variable of interest in Equation (2)—$\mathbb{1}\{t > 2015\} \times \text{Treat}_j$—with the interaction terms between Treat_j and a series indicators capturing the number of years since the policy takes place. The estimated β^{τ}'s measure the effect relative to the baseline year 2015. Interacting the treatment group indicator with the year 2014 dummy also enables us to test for any pre-existing trend in the year before the change in ADR limits. In the presence of spillovers, our theoretical prediction is that β^{τ} would start to be negative when $\tau \geq 2016$. This would provide evidence that control DRGs in treated hospitals have lower top coding rates after the policy change relative to "pure control" DRGs in control hospitals.

The identification comes from the relative variation in the top coding rate at the hospital/base DRG level across time. The key identifying assumption is that, absent ADR limits, any pre-existing differences in coding intensity among the control DRGs between treated and control hospitals would have evolved similarly before and after the policy change. While the "treatment"—defined by a hospital's overall audit rate pre-treatment—is not randomly assigned, our identification comes from the differential evolution of top-coding rates across these two types of control DRGs around the time of the implementation of ADR limits.

We estimate our main specification in Equations (2) and (3) using a standard Two-Way Fixed Effects (TWFEs) approach. We cluster standard errors at the hospital 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 all hospital-base 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 Callaway and Sant'Anna (2021), referred as CSDiD from now on—to estimate the treatment effects.

5.2 Main results

We now discuss the results for the specifications where we test the presence of spillovers in coding practices across DRGs within a hospital. Figure 2 shows the estimated coefficients and their 95% confidence intervals for the key parameters of interest in Equation (3)—the interaction terms between the indicator for the control DRGs in treated hospitals and the series of leads/lags. We present the estimates from both the TWFEs model and the CSDiD estimator. We include a dashed line before the year 2016 to indicate when the treatment starts to have an impact. Columns (1) and (2) of Appendix Table A3 provide more details on the results.

Starting in 2016, we find that treated hospitals reported a smaller proportion of patients with top-tier DRGs in (control) base DRGs relative to the same set of DRG's in control hospitals (i.e. the "pure control" DRGs). The effect is significantly negative in most years after the treatment, whereas the differences in top-coding rates among these DRGs in treated and control hospitals were statistically insignificant before the policy. The estimated treatment effect from TWFEs is equivalent to an average reduction of 1.44 percentage points in top coding, or an average decrease by 4.43% (=0.0144/0.325 $\times$ 100%) among treated hospitals following the policy change, relative to control hospitals.[32]

The results show that with regard to the same set of top-level DRGs for which the audit probability did not vary after the policy, treated hospitals assigned fewer of the claims to these top tier codes, relative to control hospitals. This evidence challenges the assumption of no spillovers across base DRGs, common in the literature, but it is instead consistent with the hypothesis that hospitals substitute top coding in some base DRGs with that in others when re-optimizing coding intensity in response to declining monitoring pressure.

[32]We obtain the 4.43% by dividing the estimated treatment effect from the TWFE estimator (shown in Column (1) in Panel A of Appendix Table A3) by the average percent top codes among control DRGs in treated hospital during pre-treatment periods (shown in Panel 1 of Table 1).

5.3 Additional analyses and robustness checks

In the following, we discuss the results on additional analyses that we use to assess the robustness of our findings above. First, we examine whether the differential coding pattern in control DRGs between treated and control hospitals is driven by the different patient population between these hospitals. For instance, we want to test whether the finding of relatively fewer reported top codes among the control DRGs in treated hospitals is due to the less severe patients in these hospitals. To achieve this, we replace percent top codes with the following outcome measures: average length of stay and average number of diagnoses at the hospital-DRG level. If patients among these DRGs in control hospitals become sicker across time relatively to those in treated hospitals, we might expect relatively longer hospital stays and relatively more underlying medical conditions in control hospitals over time. Appendix Figure 3 shows the event study results separately for each of these measures. In neither of them do we see a similar pattern to our findings above.

Second, we try different specifications for our main analyses. First, we further add a set of common hospital characteristics (mentioned in Section 4.1) to Equation (3). Figure 4(a) shows the estimates. The results are materially the same. Second, we apply a different DiD estimator, developed by De Chaisemartin and d'Haultfoeuille (2024), and present the results in Figure 4(b). The estimates are similar. Third, we re-run the main specifications using a balanced sample of hospital-DRG pairs. To achieve this, we drop hospital-DRG pairs which only show up in parts of the sample periods. Figure 4(c) shows that the estimates are similar to our main results. Finally, we redo the estimation without using sampling weights. The main findings still hold, as shown in Figure 4(d). Overall, these robustness checks confirm that our main results are consistent across alternative specifications.

6 Empirical Analysis at the Hospital Level: Top-Coding and Audit Pressure

Given the evidence on spillovers at the hospital-base DRG level, in this section we move on to a hospital-level analysis that is not hindered by spillovers across DRGs within hospitals. Imposing the ADR limit reduces the audit pressure facing hospitals and could potentially lead to more improper coding. We now examine the overall effect of reduced monitoring at the hospital level. We first apply a DiD research design and discuss the findings. Going beyond the overall policy effect, we then estimate the causal effect of audit pressure on

the top coding rate using a fixed effects model with a continuous treatment variable, the audit rate. Further, to address potential endogeneity concerns, we use the exogenous policy variation from our DiD analysis as an instrumental variable.

6.1 Empirical strategy in a DiD framework

To examine empirically the extent to which hospitals vary the reporting of top-coded patients in response to the change in audit pressure, we use the following specification:

$$\texttt{TopCodeRate}_{jt} = \beta \times \mathbb{1}\{t > 2015\} \times \text{Treat}_j + \gamma_j + \lambda_t + \varepsilon_{jt}, \tag{4}$$

where $\texttt{TopCodeRate}_{jt}$ denotes the fraction of discharges reported with top-tier DRGs among all discharges from multi-level base DRGs in hospital j and year t. $\text{Treat}_j = 1$ if hospital j is a treated hospital, for which the average *pre*-treatment audit rate was strictly greater than 0.5%. We also include year fixed effects (λ_t) and hospital fixed effects (γ_j), with the latter capturing the time-invariant unobservables at this level.[33]

Our key coefficient of interest, β, measures the differential response in the share of claims that are coded into top-tier DRGs among treated hospitals relative to control hospitals before and after the introduction of ADR limits. Noting that treatment here really means a reduction in audit pressure, the finding of $\beta > 0$ implies that hospitals would assign top codes to a greater proportion of patients as the audit probability drops, providing suggestive evidence of more aggressive coding when monitoring declines. We also apply an event study framework to Equation (4):

$$\texttt{TopCodeRate}_{jt} = \sum_{\tau=2014,\ \tau\neq 2015}^{2019} \beta^{\tau} \times \mathbb{1}\{t = \tau\} \times \text{Treat}_j + \gamma_j + \lambda_t + \varepsilon_{jt}, \tag{5}$$

where the β^{τ}'s help examine whether there were differential pre trends in the outcome measure between treated and control hospitals and capture the dynamic effects post-treatment.

The identification is based on variation within a hospitals over time. The key identifying assumption shares a similar idea to that in Section 5.1, requiring that, absent the implementation of ADR limits, the difference in coding intensity between treated and control hospitals would have had the same trends. We also use both TWFEs and CSDiD to estimate Equa-

[33] As a robustness check, we also control for some important hospital characteristics, and the results are similar. More details can be found in Section 6.3.

tions (4) and (5). We weight the regressions using total discharges from multi-level base DRGs in a hospital and cluster standard errors at the hospital level.

6.2 Main results

Figure 5 shows the event study estimates for the interaction terms with year dummies in a similar format to Figure 2. Immediately after 2015, we observe a larger fraction of discharges reported with top codes in treated hospitals (where the overall audit rate declined to be under the cap) relative to the control group. The effect is significantly positive in every year post-treatment and lasts for several years. In contrast, there is no significant difference in fraction top-coding between the two types of hospitals during the pre-treatment period. We provide more details on the results in Columns (1) and (2) of Appendix Table A5.

The treatment effect estimated from the TWFEs estimator corresponds to an average increase of 2.3 percentage points in top coding, or an average increase by 7.0% (=0.0230/0.329× 100%) among the treatment group following the policy change, relative to the control group.[34] The results suggest that hospitals with higher pre-treatment audit rates top-code relatively more patients following the implementation of ADR limits.[35] It is consistent with hospitals engaging in more aggressive coding under reduced audit pressure.

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). Therefore we now turn to the estimated treatment effects by hospital type. Figure 6(a) shows the main estimates separately for for-profit versus not-for-profit hospitals based on

[34]We obtain the 7.0% by dividing the estimated treatment effect from the TWFEs estimator (shown in Column (1) in Panel A of Appendix Table A3) by the average percent top codes among the treatment group during the years 2014-2016 (shown in Table 1).

[35]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 hospital decision-making regarding coding admissions into higher reimbursement levels, the focus of our paper, 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 in which unnecessary admissions are adjudicated. Thus, the extent to which hospitals respond to the respective policy changes 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 Dafny (2005), where she found no significant impacts on admissions but more patients reported with higher codes in response to a change in reimbursement rates.

TWFE models.[36] Interestingly, the positive response to reduced monitoring mainly occurs among for-profit hospitals, whereas among the not-for-profit hospitals, the treatment effect is is not statistically significant. It is consistent with the findings in the hospital upcoding literature (Silverman and Skinner, 2004; Dafny, 2005; Li, 2014): for-profit hospitals tend to code more aggressively, compared to not-for-profit hospitals.

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.[37] Following the hospital upcoding literature (Dafny, 2005; Li, 2014), we define a hospital to be financially healthy if its average debt-asset ratio during the pre-treatment period is below the bottom 25^{th} percentile and a hospital to be financially less healthy if this ratio is above the top 25^{th} percentile. Following the policy change, both types of hospitals saw an uptick in the fraction of top-coded patients in the treatment group, relative to the control group.[38]

Finally, we estimate the differential effects of reduced monitoring between stand-alone hospitals and hospitals affiliated with a healthcare system. Given the trend of consolidation in health care, it may be important to understand to what extent the coding pattern among affiliated hospitals differ from that among stand-alone hospitals, in response to varying audit pressure. Figure 6(c) presents the estimates separately for each type of hospitals. It seems that both types of hospitals seize the opportunity to enhance revenues when monitoring becomes less intense, with a slightly higher increase in stand-alone hospitals.[39]

6.3 Additional analyses and robustness checks

We conduct the following analyses for robustness checks. First, we examine whether our finding of the relatively greater proportion of top-coded patients in treated hospitals is driven by a (potentially) increasingly sicker population in those hospitals over time. We use the same set of outcome measures as a proxy for patient severity, as those in Section 5.3. Appendix Figure 7 shows the event study estimates for all these measures. It does not appear that the increase in percent top codes among treated hospitals post-treatment to pre-treatment is due to changes in patient population at those hospitals relative to control hospitals.

[36] We report the estimates in Columns (1) and (2) of Appendix Table A6 and present the results from the CSDiD in Appendix Figure A4(a).

[37] We present the results in Columns (3) and (4) of Appendix Table A6 and plot the estimates from CSDiD in Appendix Figure A4(b).

[38] Appendix Table A6 provides more details.

[39] We include more information on the estimates in Appendix Table A6 and present the results using CSDiD in Appendix Figure A4(c).

Second, we try different specifications to examine the robustness of the findings above. For instance, we additionally control for common hospital characteristics (the same set of hospital controls in Section 5.3), use an alternative DiD estimator, or drop regression weights. Our findings are rather robust in all the alternative specifications, as shown in Figure 8.[40]

6.4 The Causal Effects of Audits: IV estimation

While the previous analysis provides convincing evidence of the policy effects brought by the ADR limits, it is worthwhile to also obtain an estimate of the causal effect of the audit pressure. To this end we exploit the continuous variation in audit pressure across hospitals over time and estimate a TWFE model relating the top coding rate to the audit rate, our (continuous) measure of audit pressure faced by the hospitals. Specifically, we start from the following specification:

$$\texttt{TopCodeRate}_{jt} = \beta \times AuditRate_{jt}^{top} + \gamma_j + \lambda_t + \varepsilon_{jt}, \tag{6}$$

where $AuditRate_{jt}^{top}$ denotes the hospital audit rate based on top-tier DRGs, defined as the number of top-coded claims reviewed by RACs among all the claims reported with top codes in the given hospital and year. Our key coefficient of interest, β, measures the change in overall top-coding rate in hospital j given a unit change in audit rate at the same level. A negative estimate of β would imply an increase in reported top codes in a hospital as a result of declining audit pressure facing the hospital. This provides evidence that hospitals are more aggressive in their coding practices when the monitoring they face is less intense.

Concerns may arise regarding the endogeneity in $AuditRate_{jt}^{top}$. In particular, the within-hospital variation in the audit pressure that hospitals face over time could be correlated with time-varying unobserved hospital heterogeneity. For instance, RACs may adjust the claims selection strategy for a hospital according to their knowledge about the hospital's certain characteristics, such as managerial skills of administrators, information technology capacity, experience of billing staff, and so on, which may also affect the hospital's coding practices. To address this concern, we instrument $AuditRate_{jt}^{top}$ with $\mathbb{1}\{t > 2015\} \times \text{Treat}_j$, the interaction between the treated hospital indicator and post treatment indicator. It is similar to the idea of shift-share instruments, a product of the exposure share to a policy shock—measured by the treatment status of a hospital—and the quasi-experimental shock—measured by the

[40] While the pre-treatment effect is significantly negative in the specification without regression weights, we believe that the main findings still hold given the sharp increase in top-coding immediately after 2015 and the continuously increasing trend post-treatment.

introduction of ADR limits. We expect that treated hospitals, whose pre-treatment audit rate was above the limit, would be impacted more substantially by the implementation of this policy. Even though the exposure share is not as good as randomly assigned, the instrument could still be valid, as the exogeneity mainly arises from the policy change, which is plausibly exogenous (Borusyak et al., 2022).

We present the IV estimates in Panel C of Appendix Tables A5 – A7. Column (1) of Appendix Tables A5 shows the estimated effect of audit rates on percent top codes based on entire sample. It is significantly negative, suggesting an average increase of 3.26 percentage points as the audit rate drops by 1 percentage point. The magnitude is not substantially different from that implied by our DiD estimate (2.3 percentage points), considering that the overall audit rate dropped by approximately 1 percentage points post implementation of ADR limits (shown in Appendix Table A1). In all, the IV estimates are consistent with the key take-away of our findings from the DiD estimation: hospitals assign top codes to a larger proportion of patients when they face lower audit pressure; these findings do not appear to be driven by differential patient severity trends across hospitals over time.

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 level, and present the results in Appnedix Figure A3.[41] We see an immediate decline in corrected overpayments among treated hospitals relative to control hospitals following the imposition of ADR limits. The magnitude corresponds to an average reduction of $107,927 per hospital. Thus, the direct savings from reclaimed payments dropped by $96,325, and Medicare paid RACs a smaller amount of contingency fee, by $11,602, at the same level.[42]

We next quantify the magnitude of additional reimbursement claims submitted to Medicare due to the more intense coding by hospitals after the implementation of ADR limits, and the resulting reduction in audit pressure. The median numer of discharges per treated hospital is 811. From Column (1) of Appendix Table A5, treated hospitals assign more patients to top-tier DRGs by 2.30 percentage points post-treatment to pre-treatment. The

[41] We focus on overpayments assuming that the policy change has little impact on underpayments and considering that most of the reclaimed improper payments arose from overpayments during our sample period—over 85%, on average.

[42] We obtain the contingency fee by multiplying $107,927 by the contingency fee ratio, 10.75% (the average of 9% and 12.5%), similar to the approach in Shi (2024). The reduction in direct saving is derived from subtracting $11,602 from $107,927.

pre-treatment average spread of the base DRGs among treated hospitals is approximately 0.9. 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 treated hospital due to more aggressive coding is $100,071.

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 $17, among which 51% 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.

7 Conclusion

We seek to understand how hospitals re-optimize their coding decisions after a cap (ADR limit) was imposed on the number of document requests that auditors could solicit from hospitals during audits. This cap effectively reduced audit probability a hospital faced. A common approach would be exploiting the resulting variation in coding intensity across diagnostic groups (DRGs), given the differential changes in audit probability at this level before and after the policy shock. To obtain unbiased estimates though, some form of the Stable Unit Treatment Value Assumption (SUTVA) must be invoked, requiring no spillovers across DRGs in coding response due to the change in audit pressure. However, this assumption may be violated if hospitals are confronted with some entity-wide binding constraint that created inter-dependence in coding decisions across different DRGs. In particular, the optimal level of upcoding for a DRG could change even if the audit pressure faced by that DRG did not change. We formalize the idea with a simple model.

To test the presence of spillovers across DRGs within hospitals, we compare the top coding rate among control DRGs in treated hospitals with that among the same set of DRGs in control hospitals (that is, the "pure control" DRGs) over time, before and after the ADR limit is imposed. We find relatively fewer patients reported with the top codes for control DRGs in treated hospitals, post-treatment versus pre-treatment relative to the pure control DRGs, which is inconsistent with SUTVA.

To minimize the bias from potential spillovers, we examine the effect of reduced audit pressure on coding decisions at the more aggregate hospital level. We first apply a difference-in-differences framework and find that hospitals, whose overall pre-treatment audit rate

exceeded 0.5%, top-code relatively more patients after the introduction of ADR limits. Our results are quantitatively similar when we use this strategy to instrument for the continuous audit rate, thus obtaining an estimate of the causal effect of the audit pressure. A back of the envelope calculation shows that for every dollar Medicare saves on the program after the scale-back, it may have to bear a loss of $17, with more than half coming from the increased payments to hospitals due to 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 for-profit hospitals.

A recent report from the Office of Inspector General (OIG, 2021) has found that hospitals assigned a greater proportion of inpatient stays to the highest severity level, after an analysis based on Medicare Part A claims during 2014 – 2019. While we do not seek to attribute the increasing trend of hospital billing and Medicare payments to the more limited oversight, our results suggest that the Medicare Recovery Audit Program has potential in combating fraud, waste, and abuse in healthcare spending. Scaling back the program helps decrease social costs, such as lowering administrative burden for healthcare providers, but 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 thus 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. 7

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, 6.2

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy 76*(2), 169–217. 7

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-experimental shift-share research designs. *The Review of economic studies 89*(1), 181–213. 6.4

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies 91*(6), 3253–3285. 5.1

Butts, K. (2021). Difference-in-differences with spatial spillovers. *arXiv preprint arXiv:2105.03737*. 1, 3.1

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230. 1, 5.1

Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers. 1

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

Cuellar, A. E. and P. J. Gertler (2006). Strategic integration of hospitals and physicians. *Journal of health economics 25*(1), 1–28. 30

Dafny, L. (2005). How do hospitals respond to price changes? *American Economic Review 95*(5), 1525–1547. 1, 3.1, 4.1, 35, 6.2

De Chaisemartin, C. and X. d'Haultfoeuille (2024). Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*, 1–45. 5.1, 5.3

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*. 1, 3.1, 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, 3.3

Leder-Luis, J. (2025). Can whistleblowers root out public expenditure fraud? evidence from medicare. *Review of Economics and Statistics*, 1–18. 1, 16

Li, B. (2014). Cracking the codes: Do electronic medical records facilitate hospital revenue enhancement. *Working paper*. 3.1, 4.1, 6.2

Li, X., Y. Shen, and Q. Zhou (2025). Causal inference with social interactions: A structural break viewpoint. *Available at SSRN 5183952*. 1

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217. 1

OIG (2021). Trend Toward More Expensive Inpatient Hospital Stays in Medicare Emerged Before COVID-19 and Warrants Further Scrutiny. *Office of Inspector General*. 7

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*. 9, 4.1

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association 75*(371), 591–593. 1, 3.1

Sacarny, A. (2018). Adoption and learning across hospitals: The case of a revenue-generating practice. *Journal of Health Economics 60*, 142–164. 1, 3.2

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*. 1, 2.2, 35, 42

Silverman, E. and J. Skinner (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics 23*(2), 369–389. 1, 4.1, 6.2

Sinclair, B., M. McConnell, and D. P. Green (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science 56*(4), 1055–1069. 5.1

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of economic literature 57*(4), 904–954. 3

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199. 5.1

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

	Panel 1: Samples for examining spillovers		
	All	Treated	Controls
Top-coding rates (%)	35.3	32.5	36.1
	(40.3)	(41.4)	(39.9)
# control DRGs per hospital	62.9	42.2	68.2
	(37.3)	(19.8)	(38.8)
# hospitals	4,098	718	3,380
	Panel 2: Samples for studying top coding		
	All	Treated	Controls
Top-coding rates (%)	38.8	32.9	40.0
	(27.8)	(11.6)	(30.0)
Audit rates (%)	0.24	0.96	0.09
	(0.51)	(0.88)	(0.15)
# top-level DRGs per hospital	69.8	68.9	70.1
	(38.9)	(35.6)	(40.1)
# hospitals	4,108	721	3,387

Notes: Table 1 reports the summary statistics for the key variables in the pre-treatment period. The top-coding rate at the hospital/base DRG level is defined as the percentage of patients reported with the top-tier DRG in a given base DRG among all the patients from this base DRG. The top-coding rate at the hospital level is defined as the number of discharges reported with top-tier DRGs among all discharges from multi-level base DRGs. The number of hospitals is slightly smaller in the sample of examining spillovers is due to loss of observations in constructing the sample.

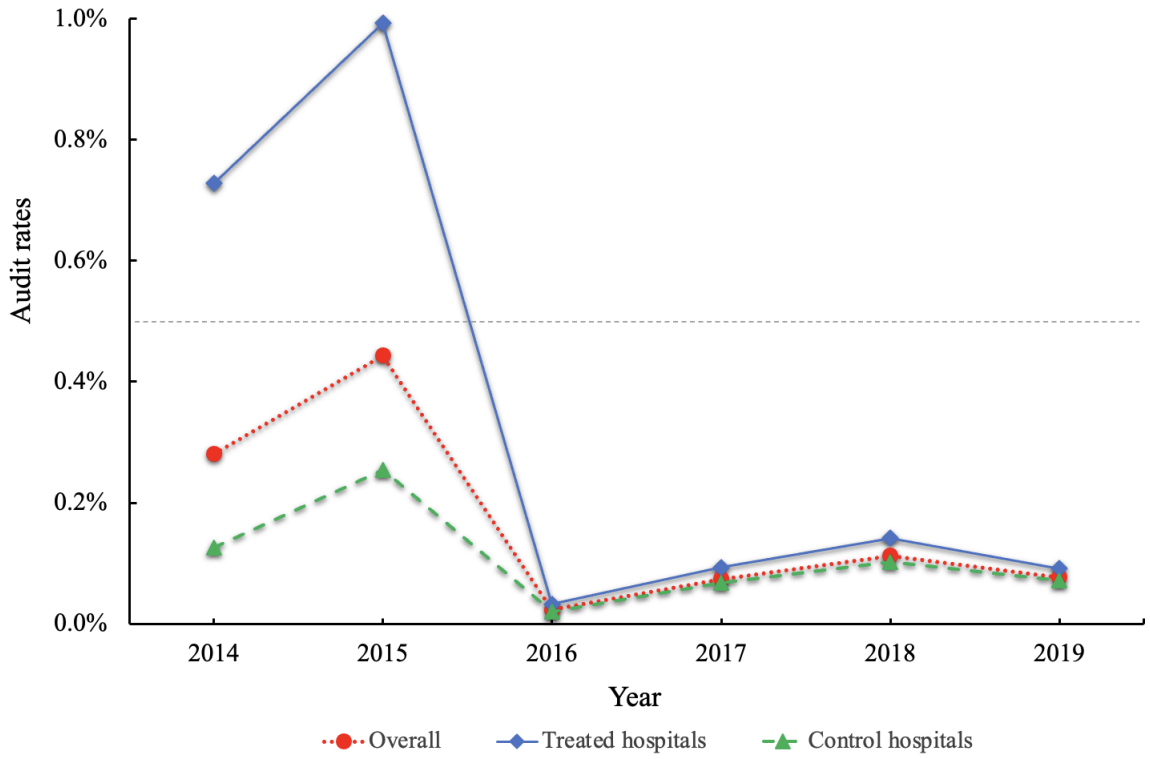


Figure 1: Average Hospital Audit Rates, Overall and by Hospital Treatment Status

Notes: We plot the average audit rate overall and separately for treated and control hospitals. The overall average is a weighted average across hospitals, using the share of claims from a given hospital relative to the total claims as the hospital weight. We calculate the average audit rates among treated and control hospitals in a similar way. We include a dashed line indicating the ADR limit, 0.5%. Appendix Table A1 provides more information.

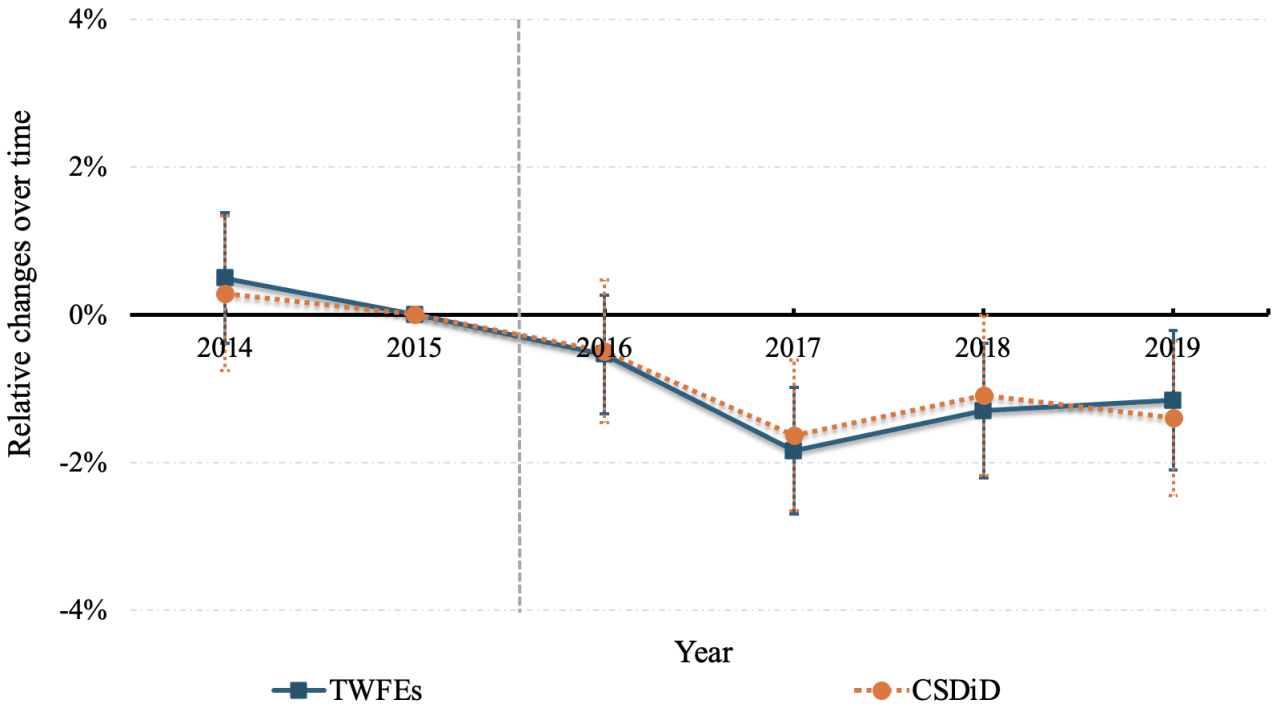
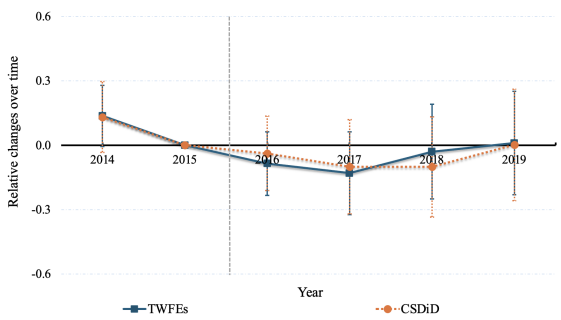
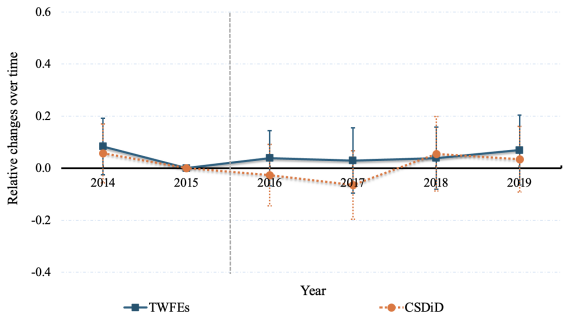


Figure 2: Examining Spillover across DRGs

Notes: The figure shows the event study coefficients (estimated by TWFEs and CSDiD) of examining the differential effect of imposing the ADR limit on top-coding rates between control DRGs in treated hospitals and the same set of DRGs in control hospitals. The lines report the coefficients for the indicator of control DRGs in treated hospitals 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 hospital level. Unit of observation is hospital/DRG/year. Sample includes the same set of control DRGs in treated hospitals and control hospitals. Other regressors are hospital-DRG fixed effects and year fixed effects.



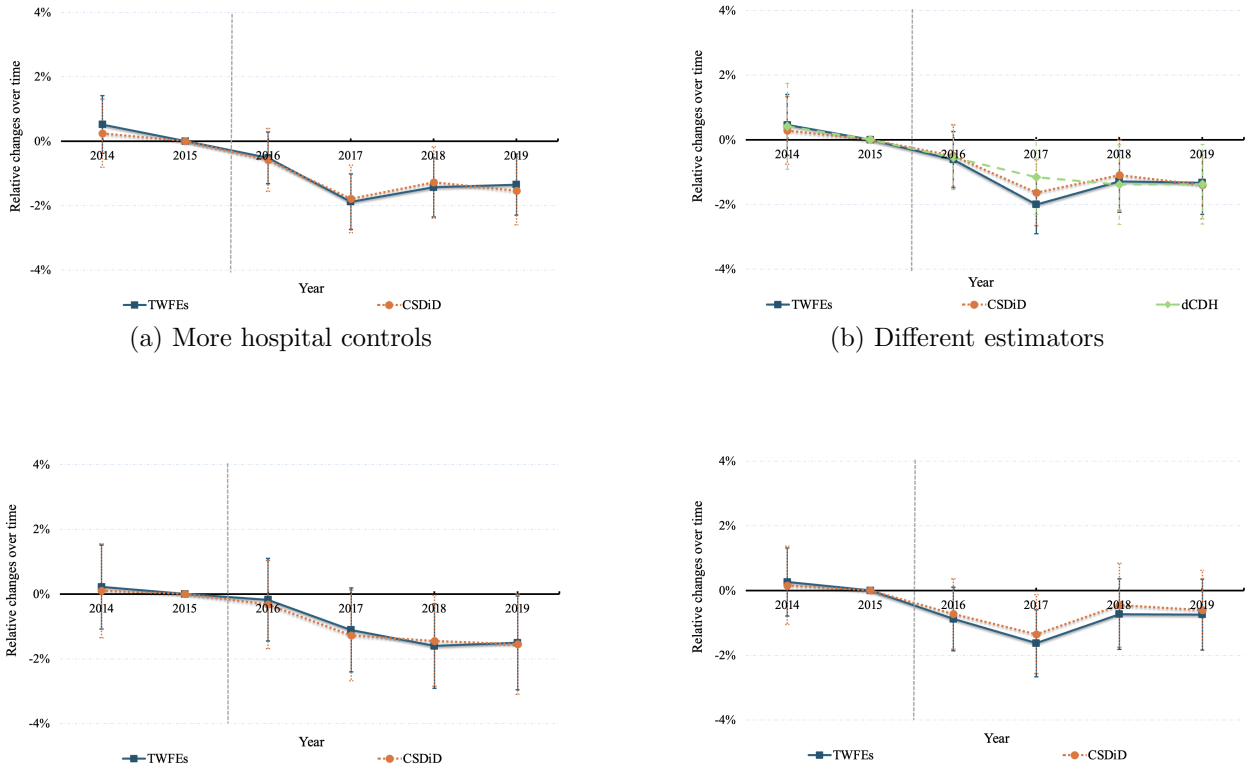
(a) Number of diagnoses



(b) Average length of stay

Figure 3: Spillover Analyses, Using Additional Outcomes

Notes: The figures show the event study coefficients (estimated by TWFEs and CSDiD) for other outcome measures. The lines report the coefficients for the indicator of control DRGs in treated hospitals interacting with year dummies. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the hospital level. Unit of observation is hospital/DRG/year. Sample includes the same set of control DRGs in treated hospitals and control hospitals. Other regressors are hospital-DRG fixed effects and year fixed effects.



(a) More hospital controls

(b) Different estimators

(c) Using a balanced panel

(d) Without sampling weights

Figure 4: Examining Spillover across DRGs, in Different Specifications

Notes: The figure shows the event study coefficients from the robustness checks for the spillover analysis. The lines report the coefficients for the indicator of control DRGs in treated hospitals 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 hospital level. Unit of observation is hospital/DRG/year. Sample includes the same set of control DRGs in treated hospitals and control hospitals. Other regressors are hospital-DRG fixed effects and year fixed effects. Additional hospital controls are specified in Section 4.1.

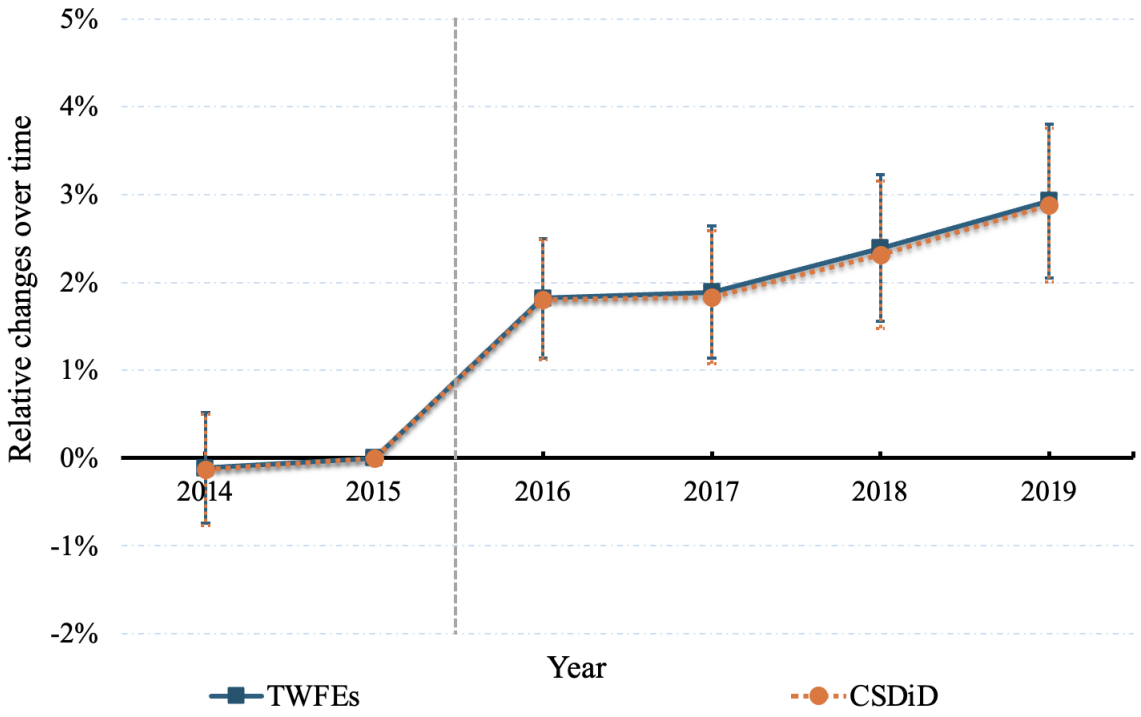
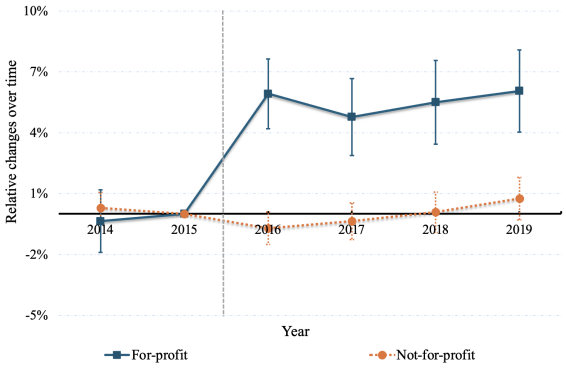
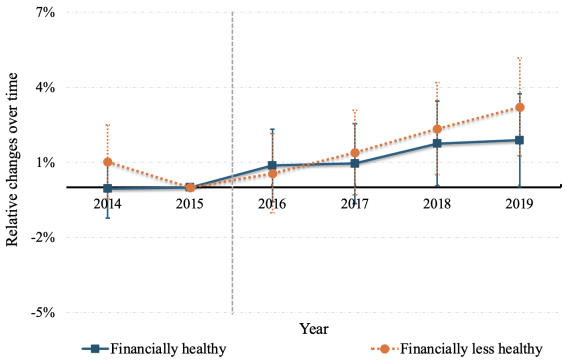


Figure 5: Effect of Reduced Audit Pressure on Hospital Top-Coding
(TWFEs and CSDiD Estimates)

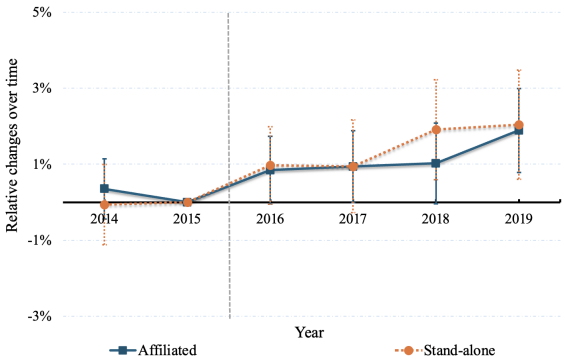
Notes: The figure shows the event study coefficients (estimated by different estimators) in the analysis of examining top-coding and audit pressure. The lines report the coefficients for the treated hospital indicator interacting with year dummies. Each dot is a regression coefficient expressed as a percentage point. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.



(a) By profit status



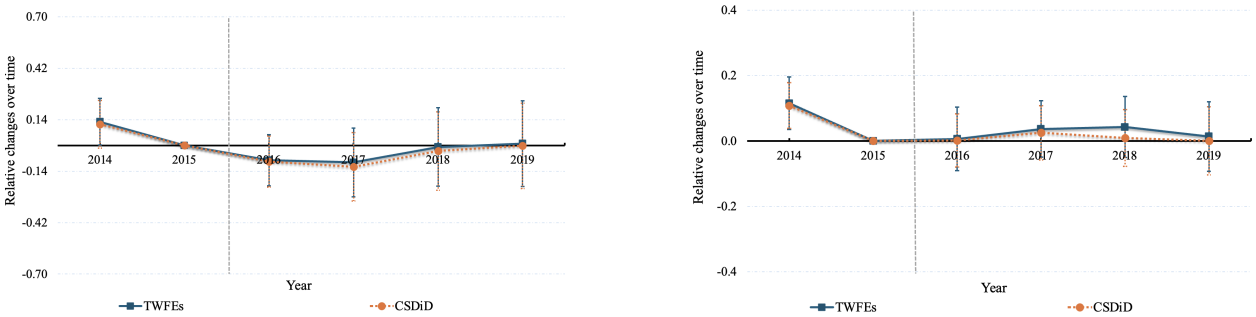
(b) By financial health status



(c) By system affiliation

Figure 6: Effect of Reduced Audit Pressure on Hospital Top-Coding, by Hospital Type

Notes: The figures show the event study coefficients (estimated by TWFEs) in the analysis by hospital type. The lines report the coefficients for the treated hospital indicator 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 hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.

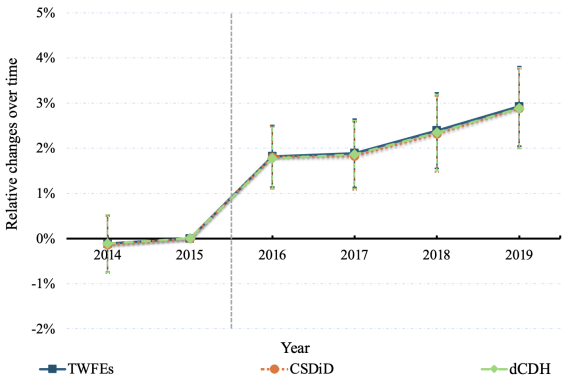


(a) Number of diagnoses

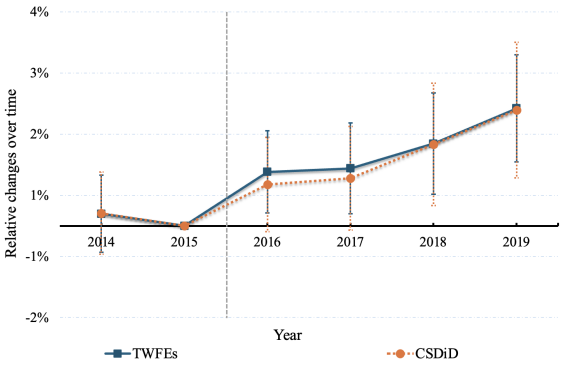
(b) Average length of stay

Figure 7: Effect of Reduced Audit Pressure on Alternative Outcomes

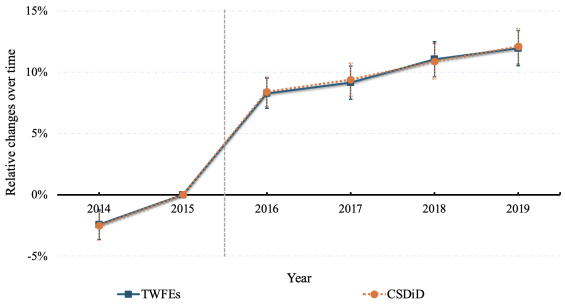
Notes: The figures show the event study coefficients (estimated by TWFEs and CSDiD) for other outcome measures. The lines report the coefficients for the treated hospital indicator interacting with year dummies. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.



(a) Different estimators



(b) More hospital controls



(c) Without sampling weights

Figure 8: Effect of Reduced Audit Pressure on Hospital Top-Coding, in Different Specifications

Notes: The figure shows the event study coefficients (estimated by different estimators) from the robustness checks for examining top-coding and audit pressure. The lines report the coefficients for the treated hospital indicator 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 hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.

Appendix I Proof of Proposition 1

In this section, we seek to prove Proposition 1 in Section 3.2. The implementation of ADR limits implies changes in $\phi_{\mathcal{A}}^{\text{AUDIT}}$ and/or $\phi_{\mathcal{B}}^{\text{AUDIT}}$. For simplicity, we only let $\phi_{\mathcal{B}}^{\text{AUDIT}}$ change, assuming that claims reported with the top-tier DRG in $\mathcal{B}$ were much more heavily audited than those in $\mathcal{A}$ prior to the policy change. Thus, imposing the ADR limit (an institute-based calculation) would affect the change in the probability of auditing the top-coded cases in $\mathcal{B}$ more substantially than that in $\mathcal{A}$. We consider the effect of a decline in $\phi_{\mathcal{B}}^{\text{AUDIT}}$ on hospitals' coding intensity in each base DRG and obtain the following proposition:

Proposition 1 As the audit rate of the top code in one base DRG goes down, we expect two possible cases:

(i) When the hospital-wide constraint is binding, i.e., the overall upcoding rate of the hospital reaches $\bar{\kappa}$, the effect of varying audit rates for the top code in base DRG $\mathcal{B}$ would spillover onto the decision of coding intensity in base DRG $\mathcal{A}$. Specifically, the proportion of improperly top-coded cases rises in base DRG $\mathcal{B}$ and drops in base DRG $\mathcal{A}$.

(ii) When the hospital-wide constraint is not binding, the upcoding rate in base DRG $\mathcal{A}$ does not vary with $\phi_{\mathcal{B}}^{\text{AUDIT}}$ whereas a greater proportion of disqualified patients in base DRG $\mathcal{B}$ will be assigned the top code.

Proof. With a lower $\phi_{\mathcal{A}}^{\text{AUDIT}}$ after the policy change, the Lagrangian function has the following form:

$$\mathscr{L}(\tilde{s}_{\mathcal{A}}^{u}, \tilde{s}_{\mathcal{B}}^{u}, \lambda) \quad = E[\pi_{\mathcal{A}}(\tilde{s}_{\mathcal{A}}^{u})] + E[\pi_{\mathcal{B}}(\tilde{s}_{\mathcal{B}}^{u})] + \lambda \left(\underline{s}_{\mathcal{A}}^{u} - \tilde{s}_{\mathcal{A}}^{u} + \underline{s}_{\mathcal{B}}^{u} - \tilde{s}_{\mathcal{B}}^{u} - \bar{\kappa} + \mu^{2}\right),$$

where μ^2 is the slack variable and λ is the Lagrange multiplier. Write down the gradients:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \tilde{s}_{\mathcal{A}}^{u}} &= R\omega_{\mathcal{A}}^{low} - c_{\mathcal{A}}^{low} - [R\omega_{\mathcal{A}}^{top} - c_{\mathcal{A}}^{low} - \phi_{\mathcal{A}}^{\text{AUDIT}} \times \phi_{\mathcal{A}}^{u}(|\underline{s}_{\mathcal{A}}^{u} - \tilde{s}_{\mathcal{A}}^{u}|) \times \xi^{u}] - \lambda \\
\frac{\partial \mathscr{L}}{\partial \tilde{s}_{\mathcal{B}}^{u}} &= R\omega_{\mathcal{B}}^{low} - c_{\mathcal{B}}^{low} - [R\omega_{\mathcal{B}}^{top} - c_{\mathcal{B}}^{low} - \phi_{\mathcal{B}}^{\text{AUDIT}} \times \phi_{\mathcal{B}}^{u}(|\underline{s}_{\mathcal{B}}^{u} - \tilde{s}_{\mathcal{B}}^{u}|) \times \xi^{u}] - \lambda \\
\frac{\partial \mathscr{L}}{\partial \lambda} &= \underline{s}_{\mathcal{A}}^{u} - \tilde{s}_{\mathcal{A}}^{u} + \underline{s}_{\mathcal{B}}^{u} - \tilde{s}_{\mathcal{B}}^{u} - \bar{\kappa} + \mu^{2}.
\end{aligned}$$

According to the Krush-Kuhn-Tucker conditions, there are two possible cases: (i) when the constraint is *not* active, i.e., $\underline{s}_{\mathcal{A}}^{u} - \tilde{s}_{\mathcal{A}}^{u} + \underline{s}_{\mathcal{B}}^{u} - \tilde{s}_{\mathcal{B}}^{u} < \bar{\kappa}$; and (ii) when the constraint is active, i.e., $\underline{s}_{\mathcal{A}}^{u} - \tilde{s}_{\mathcal{A}}^{u} + \underline{s}_{\mathcal{B}}^{u} - \tilde{s}_{\mathcal{B}}^{u} = \bar{\kappa}$.

In Case (i), $\lambda = 0$ and $\mu^2 > 0$. By $\frac{\partial \mathscr{L}}{\partial \tilde{s}^u_{\mathcal{A}}} = 0$ and $\frac{\partial \mathscr{L}}{\partial \tilde{s}^u_{\mathcal{B}}} = 0$, we have

$$R\omega^{low}_{\mathcal{A}} - c^{low}_{\mathcal{A}} - [R\omega^{top}_{\mathcal{A}} - c^{low}_{\mathcal{A}} - \phi^{\text{AUDIT}}_{\mathcal{A}} \times \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|) \times \xi^u] = 0$$
$$R\omega^{low}_{\mathcal{B}} - c^{low}_{\mathcal{B}} - [R\omega^{top}_{\mathcal{B}} - c^{low}_{\mathcal{B}} - \phi^{\text{AUDIT}}_{\mathcal{B}} \times \phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|) \times \xi^u] = 0$$

Take the derivative w.r.t. $\phi^{\text{AUDIT}}_{\mathcal{B}}$ on both sides of the two equations above:

$$-\phi^{\text{AUDIT}}_{\mathcal{A}} \times \frac{\partial \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)}{\partial(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \times \xi^u = 0 \tag{7}$$

$$\phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|) \times \xi^u - \phi^{\text{AUDIT}}_{\mathcal{B}} \times \frac{\partial \phi^u_{\mathcal{B}}(|\underline{s}^u - \tilde{s}^u_{\mathcal{B}}|)}{\partial(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \times \xi^u = 0 \tag{8}$$

Equation (7) $\implies \frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} = 0$. Equation (8) $\implies \frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} > 0$. The prediction from Equation (7) suggests that hospitals tend to report more patients with the top code if the audit probability for this code drops. It is relatively standard and in line with seminal work on the economic of illicit behavior (Becker, 1968; Allingham and Sandmo, 1972). The implication from Equation (8) is that when the hospital-wide constraint is not binding, there are no spillovers in coding practices between $\mathcal{A}$ and $\mathcal{B}$.

In Case (ii), $\lambda \neq 0$ and $\mu^2 = 0$. By $\frac{\partial \mathscr{L}}{\partial \tilde{s}^u_{\mathcal{A}}} = 0$, $\frac{\partial \mathscr{L}}{\partial \tilde{s}^u_{\mathcal{B}}} = 0$, and $\frac{\partial \mathscr{L}}{\partial \lambda} = 0$, we have

$$R\omega^{low}_{\mathcal{A}} - c^{low}_{\mathcal{A}} - [R\omega^{top}_{\mathcal{A}} - c^{low}_{\mathcal{A}} - \phi^{\text{AUDIT}}_{\mathcal{A}} \times \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|) \times \xi^u] = \lambda$$
$$R\omega^{low}_{\mathcal{B}} - c^{low}_{\mathcal{B}} - [R\omega^{top}_{\mathcal{B}} - c^{low}_{\mathcal{B}} - \phi^{\text{AUDIT}}_{\mathcal{B}} \times \phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|) \times \xi^u] = \lambda$$
$$\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}} + \underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}} - \bar{\kappa} = 0$$

Take the derivative w.r.t. $\phi^{\text{AUDIT}}_{\mathcal{B}}$ on both sides of the three equations above:

$$-\phi^{\text{AUDIT}}_{\mathcal{A}} \times \frac{\partial \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)}{\partial(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \times \xi^u = \frac{\partial \lambda}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \tag{9}$$

$$\phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|) \times \xi^u - \phi^{\text{AUDIT}}_{\mathcal{B}} \times \frac{\partial \phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|)}{\partial(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \times \xi^u = \frac{\partial \lambda}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} \tag{10}$$

$$\frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} + \frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} = 0 \tag{11}$$

Equation (11) $\implies$ $\frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}}$ and $\frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}}$ have the opposite signs.

Let the LHS of Equation (9) = LHS of Equation (10):

$$\phi^{\text{AUDIT}}_{\mathcal{B}} \times \frac{\partial \phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|)}{\partial(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} - \phi^{\text{AUDIT}}_{\mathcal{A}} \times \frac{\partial \phi^u_{\mathcal{A}}(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)}{\partial(|\underline{s}^u_{\mathcal{A}} - \tilde{s}^u_{\mathcal{A}}|)} \times \frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} = \phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|).$$

Because $\phi^{\text{AUDIT}}_i > 0$, $\frac{\partial \phi^u_i(|\underline{s}^u_i - \tilde{s}^u_i|)}{\partial(|\underline{s}^u_i - \tilde{s}^u_i|)} > 0$, $\phi^u_{\mathcal{B}}(|\underline{s}^u_{\mathcal{B}} - \tilde{s}^u_{\mathcal{B}}|) > 0$, for any $i = \mathcal{A}, \mathcal{B}$, and $\frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}}$ and $\frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}}$ have the opposite signs, we can conclude $\frac{\partial \tilde{s}^u_{\mathcal{B}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} > 0$ and $\frac{\partial \tilde{s}^u_{\mathcal{A}}}{\partial \phi^{\text{AUDIT}}_{\mathcal{B}}} < 0$. The results here suggests that the upcoding decision among patients in base DRG $\mathcal{A}$ could vary with the change in audit rates for the top code in base DRG $\mathcal{B}$, implying spillovers in coding practices across disease groups. □

Appendix II Extra Tables and Figures

Table A1: Average Hospital Audit Rates, Overall and by Treatment (%)

Year	Overall	Treated	Control
2014	0.28	0.73	0.13
2015	0.44	0.99	0.25
2016	0.02	0.03	0.02
2017	0.07	0.09	0.07
2018	0.11	0.14	0.10
2019	0.08	0.09	0.07
Avg. # hospitals	3,759	3,064	695

Notes: The overall average is a weighted average across hospitals, using the share of claims from a given hospital relative to the total claims as the hospital weight. We calculate the average audit rate for treated and control hospitals in a similar way.

Table A2: Hospital characteristics for the sample for studying top coding

	All	Treated	Controls
Bed size	168.2	195.2	159.9
	(199.9)	(172.4)	(207.0)
Total births	902.3	1,143.4	827.9
	(1,487.0)	(1,510.0)	(1,472.2)
Total outpatient visits	109,156.3	149,648.7	96,661.9
	(176,416.8)	(197,062.6)	(167,622.1)
For-profit hospitals (%)	29.7	22.7	31.8
	(45.7)	(41.9)	(46.6)
Not-for-profit hospitals (%)	41.0	54.8	36.8
	(49.2)	(49.8)	(48.2)
Teaching hospitals (%)	5.25	7.04	4.69
	(22.3)	(25.6)	(21.2)
Affiliated hospitals (%)	53.5	57.5	52.3
	(49.9)	(49.5)	(50.0)
Hospitals with strong integra-	24.4	27.3	23.5
-tion with physicians (%)	(42.9)	(44.6)	(42.4)
# hospitals	3,011	710	2,301

Notes: Table A2 reports the summary statistics for select hospital characteristics in the pre-treatment period for the top-coding analysis. The number of hospitals is slightly smaller than that in Table 1 due to loss of observations in merging with the AHA annual survey.

Table A3: Treatment Effects and Dynamic Treatment Effects on Testing Spillovers, Main Results and Alternative Specifications

	Overall		W/ hospital controls		Other DiD estimator	Balanced samples		W/o regression weights	
	TWFEs (1)	CSDiD (2)	TWFEs (3)	CSDiD (4)	DCDH (5)	TWFEs (6)	CSDiD (7)	TWFEs (8)	CSDiD (9)
Panel A									
Treatment Effect	-1.44***	-1.16***	-1.54***	-1.30***	-1.07***	-1.20***	-1.15*	-1.13***	-0.78*
	(0.31)	(0.40)	(0.31)	(0.40)	(0.44)	(0.48)	(0.61)	(0.35)	(0.43)
Panel B									
Treated×Year 2014	0.49	0.29	0.52	0.25	0.42	0.22	0.11	0.26	0.16
	(0.45)	(0.54)	(0.45)	(0.54)	(0.68)	(0.66)	(0.74)	(0.54)	(0.62)
Treated×Year 2015	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted
Treated×Year 2016	-0.54	-0.50	-0.52	-0.58	-0.55	-0.18	-0.33	-0.88*	-0.72
	(0.41)	(0.49)	(0.41)	(0.50)	(0.51)	(0.65)	(0.70)	(0.50)	(0.55)
Treated×Year 2017	-1.84***	-1.63***	-1.88***	-1.79***	-1.15*	-1.11*	-1.28*	-1.63***	-1.35**
	(0.44)	(0.52)	(0.44)	(0.54)	(0.59)	(0.66)	(0.71)	(0.53)	(0.62)
Treated×Year 2018	-1.30***	-1.09**	-1.43***	-1.28**	-1.39**	-1.61***	-1.44**	-0.73	-0.46
	(0.46)	(0.56)	(0.47)	(0.57)	(0.63)	(0.67)	(0.72)	(0.56)	(0.66)
Treated×Year 2019	-1.16***	-1.40***	-1.35***	-1.54***	-1.37**	-1.51**	-1.55**	-0.74	-0.61
	(0.48)	(0.53)	(0.48)	(0.54)	(0.63)	(0.74)	(0.79)	(0.56)	(0.62)
N	501,064	501,064	486,626	486,626	501,064	143,346	143,346	501,064	501,064
$P-$value of testing joint significance of pre trends	0.274	0.588	0.253	0.646	0.540	0.741	0.886	0.627	0.798

Notes: Sample is the same set of control DRGs in treated and control hospitals. Unit of observation is hospital/DRG/year. Coefficients are reported in percentage point form. Other regressors are hospital-base DRG fixed effects and year fixed effects. Standard errors are clustered at the hospital level. The hospital controls include bed size, total births, total outpatient visits, profit status, system affiliation status, the indicator for a teaching hospital, and whether the hospital has a strong integration with physicians. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Additional Analyses on Testing Spillovers

	# diagnoses		LOS	
	TWFEs	CSDiD	TWFEs	CSDiD
	(1)	(2)	(3)	(4)
Panel A				
Treatment Effect	-0.127	-0.059	0.003	-0.001
	(0.088)	(0.097)	(0.039)	(0.048)
Panel B				
Treated×Year 2014	0.137*	0.131	0.083	0.056
	(0.073)	(0.084)	(0.055)	(0.058)
Treated×Year 2015	Omitted	Omitted	Omitted	Omitted
Treated×Year 2016	-0.085	-0.038	0.038	-0.027
	(0.075)	(0.089)	(0.054)	(0.061)
Treated×Year 2017	-0.130	-0.101	0.029	-0.065
	(0.098)	(0.112)	(0.064)	(0.067)
Treated×Year 2018	-0.030	-0.101	0.038	0.054
	(0.112)	(0.119)	(0.061)	(0.073)
Treated×Year 2019	0.010	0.002	0.069	0.035
	(0.123)	(0.132)	(0.069)	(0.064)
N	500,939	500,939	500,939	500,939
$P-$value of testing joint significance of pre trends	0.061	0.119	0.132	0.327

Notes: Sample is the same set of control DRGs in treated and control hospitals. Unit of observation is hospital/DRG/year. Other regressors are hospital-base DRG fixed effects and year fixed effects. Standard errors are clustered at the hospital level. The hospital controls include bed size, total births, total outpatient visits, profit status, system affiliation status, the indicator for a teaching hospital, and whether the hospital has a strong integration with physicians. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Treatment Effects and Dynamic Treatment Effects on Testing Coding Practices, Main Results and Alternative Specifications

	Overall		W/ hospital controls		Other DiD estimator	W/o regression weights	
	TWFEs (1)	CSDiD (2)	TWFEs (3)	CSDiD (4)	DCDH (5)	TWFEs (6)	CSDiD (7)
Panel A							
Treatment Effect	2.30***	2.21***	1.16***	1.17***	2.21***	11.28***	10.20***
	(0.33)	(0.34)	(0.32)	(0.41)	(0.34)	(0.54)	(0.57)
Panel B							
Treated×Year 2014	-0.11	-0.13	0.20	0.21	-0.11	-2.43***	-2.50***
	(0.32)	(0.32)	(0.32)	(0.34)	(0.33)	(0.61)	(0.60)
Treated×Year 2015	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted
Treated×Year 2016	1.82***	1.81***	0.88***	0.68*	1.77***	8.27***	8.40***
	(0.35)	(0.35)	(0.34)	(0.39)	(0.35)	(0.62)	(0.62)
Treated×Year 2017	1.89***	1.83***	0.94***	0.78*	1.86***	9.15***	9.39***
	(0.38)	(0.38)	(0.38)	(0.43)	(0.39)	(0.69)	(0.69)
Treated×Year 2018	2.39***	2.32***	1.34***	1.33***	2.34***	11.05***	10.89***
	(0.43)	(0.43)	(0.42)	(0.51)	(0.43)	(0.74)	(0.74)
Treated×Year 2019	2.93***	2.88***	1.92***	1.89***	2.89***	11.95***	12.10***
	(0.45)	(0.45)	(0.45)	(0.57)	(0.46)	(0.73)	(0.74)
N	22,173	22,173	16,611	16,611	22,173	22,173	22,173
$P-$value of testing joint significance of pre trends	0.72	0.68	0.54	0.55	0.74	< 0.001	< 0.001
Panel C							
Treatment Effect (IVs)	-3.26***	–	-1.65***	–	-9.27***	-18.29***	–
	(0.50)	–	(0.46)	–	(0.42)	(1.17)	–

Notes: Sample is the top-tier DRGs within base DRGs with multiple levels. Unit of observation is hospital/year. Coefficients are reported in percentage point form. Other regressors are hospital fixed effects and year fixed effects. Standard errors are clustered at the hospital level. The hospital controls include bed size, total births, total outpatient visits, profit status, system affiliation status, the indicator for a teaching hospital, and whether the hospital has a strong integration with physicians. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Treatment Effects and Dynamic Treatment Effects on Testing Coding Practices, by Hospital Type

	For-profit	Not-for profit	Financially--healthy	Financially-less -healthy	Affiliated	Stand-alone
	TWFEs (1)	TWFEs (2)	TWFEs (3)	TWFEs (4)	TWFEs (5)	TWFEs (6)
Panel A						
Treatment Effect	5.73***	-0.22	1.38*	1.33*	0.99***	1.48***
	(0.86)	(0.38)	(0.72)	(0.72)	(0.42)	(0.49)
Panel B						
Treated×Year 2014	-0.36	0.30	-0.05	1.03	0.35	-0.06
	(0.78)	(0.38)	(0.60)	(0.75)	(0.40)	(0.54)
Treated×Year 2015	Omitted	Omitted	Omitted	Omitted	Omitted	Omitted
Treated×Year 2016	5.92***	-0.71*	0.88	0.56	0.84*	0.97*
	(0.87)	(0.41)	(0.74)	(0.80)	(0.45)	(0.52)
Treated×Year 2017	4.77***	-0.36	0.95	1.38	0.94*	0.94
	(0.96)	(0.46)	(0.82)	(0.86)	(0.48)	(0.62)
Treated×Year 2018	5.50***	0.09	1.75**	2.34***	1.02*	1.91***
	(1.05)	(0.51)	(0.86)	(0.94)	(0.54)	(0.67)
Treated×Year 2019	6.06***	0.75	1.88**	3.21***	1.89***	2.03***
	(1.03)	(0.53)	(0.94)	(0.99)	(0.56)	(0.73)
N	4,915	7,063	3,870	3,466	9,099	7,512
$P-$value of testing joint significance of pre trends	0.65	0.44	0.93	0.17	0.38	0.91
Panel C						
Treatment Effect (IVs)	-15.65***	0.27	-2.69*	-2.11*	-1.64**	-1.69***
	(2.71)	(0.47)	(1.40)	(1.17)	(0.71)	(0.57)

Notes: Sample is the top-tier DRGs within base DRGs with multiple levels. Unit of observation is hospital/year. Coefficients are reported in percentage point form. Other regressors are hospital fixed effects and year fixed effects. Standard errors are clustered at the hospital level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Treatment Effects and Dynamic Treatment Effects of Audit Pressure, Additional Outcomes

	# diagnoses		LOS	
	TWFEs (1)	CSDiD (2)	TWFEs (3)	CSDiD (4)
Panel A				
Treatment Effect	-0.108	-0.059	-0.032	0.009
	(0.088)	(0.087)	(0.031)	(0.034)
Panel B				
Treated×Year 2014	0.129**	0.114*	0.115***	0.108***
	(0.065)	(0.066)	(0.041)	(0.036)
Treated×Year 2015	Omitted	Omitted	Omitted	Omitted
Treated×Year 2016	-0.081	-0.087	0.006	0.002
	((0.071)	(0.071)	(0.049)	(0.041)
Treated×Year 2017	-0.093	-0.117	0.036	0.025
	(0.096)	(0.095)	(0.044)	(0.042)
Treated×Year 2018	-0.009	-0.031	0.042	0.009
	(0.109)	(0.109)	(0.047)	(0.044)
Treated×Year 2019	0.010	-0.003	0.013	0.000
	(0.119)	(0.119)	(0.054)	(0.053)
N	22,152	22,152	22,152	22,152
$P-$value of testing joint significance of pre trends	0.047	0.083	0.005	0.002
Panel C				
Treatment Effect (IVs)	15.22	–	4.53	–
	(12.49)	–	(4.43)	–

Notes: Sample is the top-tier DRGs within base DRGs with multiple levels. Unit of observation is hospital/year. Other regressors are hospital fixed effects and year fixed effects. Standard errors are clustered at the hospital level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

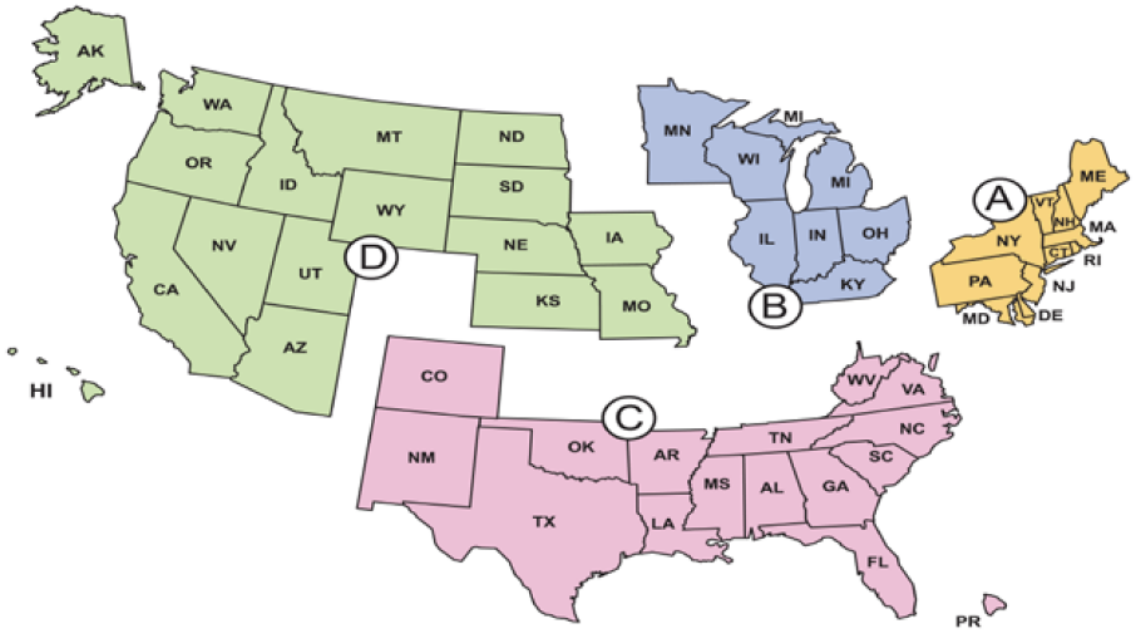


Figure A1: Map of Recovery Auditor Regions

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.

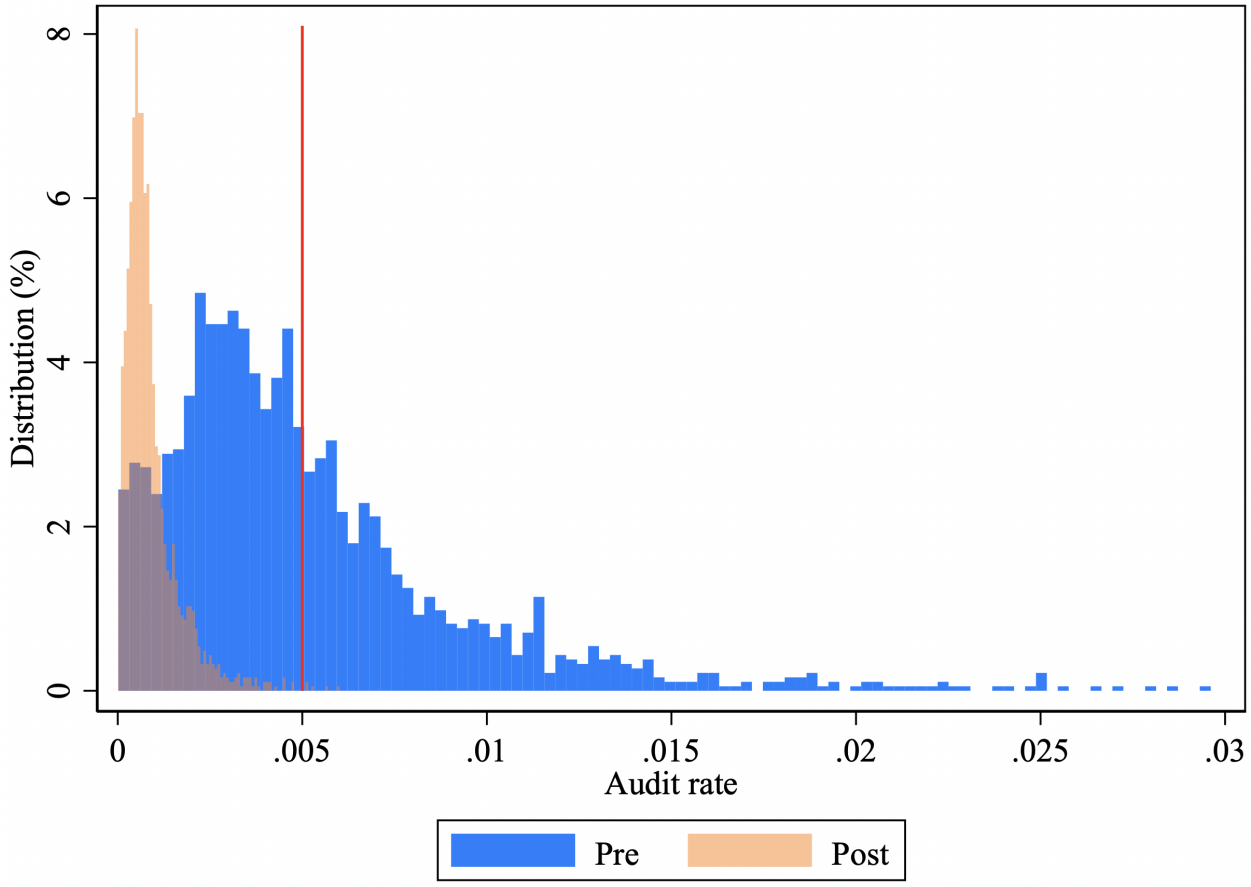


Figure A2: Frequency Distribution of Average Hospital Audit Rates, Before and After Policy Change

Notes: We show the difference in frequency distribution of average audit rates per hospital between pre- and post-policy change. For each hospital, 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%. This figure shows the distribution for audit rates up to 0.03; less than 0.5% hospitals have audit above this value.

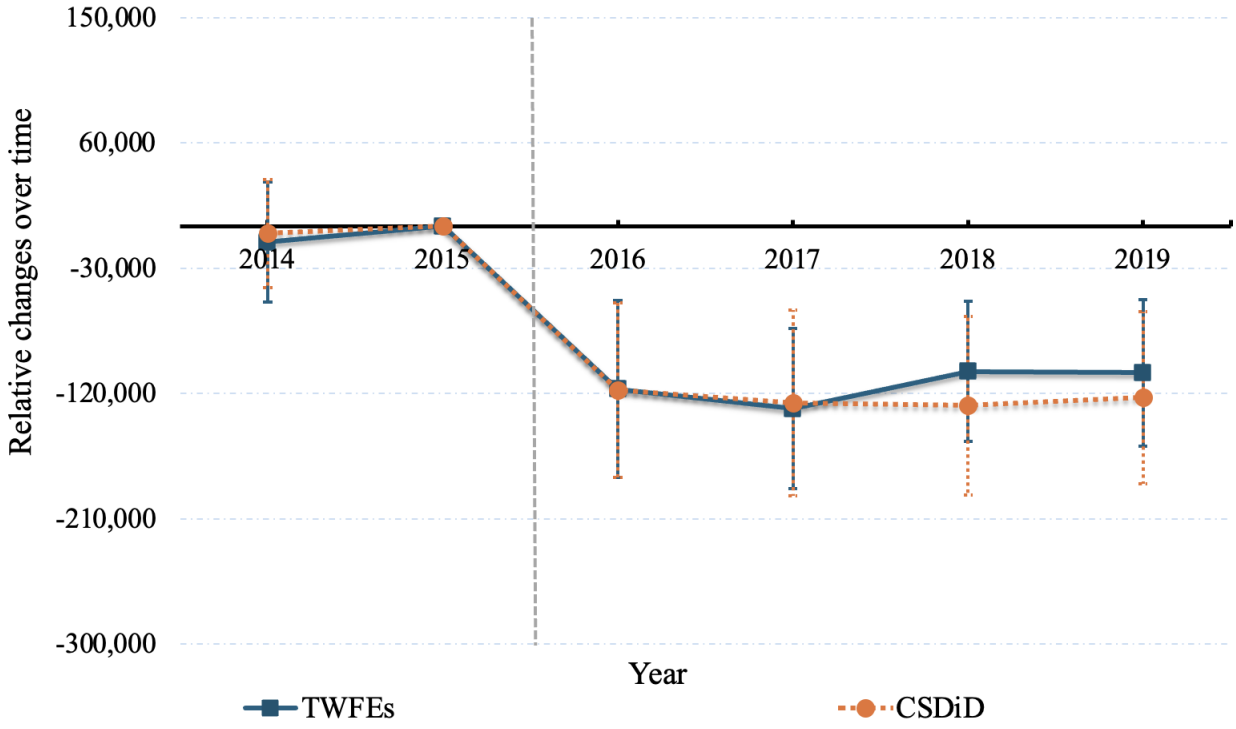
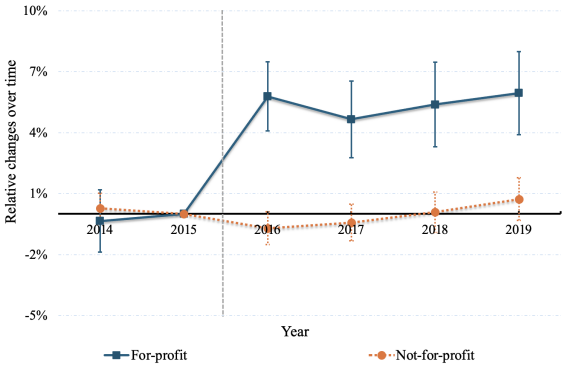
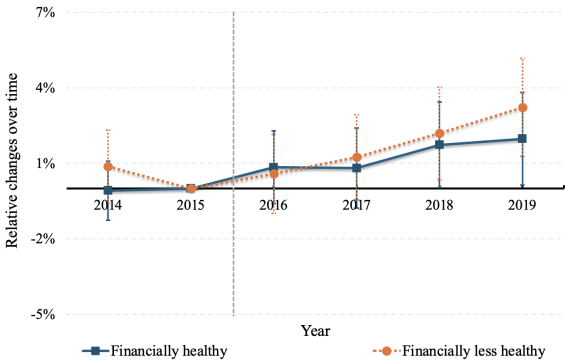


Figure A3: Effect of Scale-back of the RAP on Reclaimed Overpayments

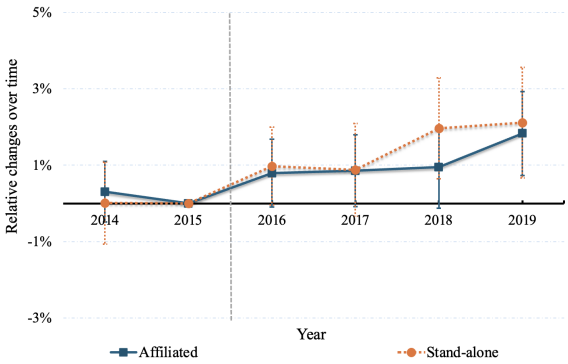
Notes: The figure shows the event study coefficients (estimated by CSDiD and TWFEs) for the analysis on reclaimed overpayments. The lines report the coefficients for the treated hospital indicator interacting with year dummies. Each dot is a regression coefficient measured in dollar value. The whiskers correspond to 95% confidence intervals, with standard errors clustered at the hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.



(a) By profit status



(b) By financial health status



(c) By system affiliation

Figure A4: Effect of Reduced Audit Pressure on Hospital Top-Coding, by Hospital Type (CSDiD)

Notes: The figures show the event study coefficients (estimated by CSDiD) in the analysis by hospital type. The lines report the coefficients for the treated hospital indicator 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 hospital level. Unit of observation is hospital/year. Sample includes all base DRGs with multiple levels. Other regressors are hospital fixed effects and year fixed effects.