

LAFAYETTE

DEPARTMENT OF ECONOMICS



Monitoring Pressure and Billing Practices: Evidence from Medicare Recovery Audits

by Jianjing Lin (UMass Amherst) and Juan Pantano (University of Hong Kong)

Discussant: Adam I. Biener

Department of Economics, Lafayette College

American Health Econometrics Workshop · October 2026

Regulatory Oversight – Bobbing for Bad Apples

Hospitals seek to profit maximize through coding decisions for a given patient volume

Previous literature shows that higher potential reimbursement incentivizes up-coding. Increased monitoring disincentivizes up-coding.

It is not well understood what happens when regulatory oversight is relaxed

Hospitals may increase up-coding when likelihood of being caught decreases

Do the gains from lower monitoring costs justify potential for increased fraud?

Monitoring is costly, but so is fraudulent upcoding that leads to overpayment

This paper uses introduction of auditing limits for Medicare Recovery Audit Program to assess welfare impact of relaxed monitoring

The paper identifies the effect of decreased monitoring through Medicare Recovery Audits on changes in top-coding, which can include intentional inflating of discharge complexity to extract higher payments from Medicare

Medicare Recovery Audit Program (RAP)

When	What happened
2005–10	RAP demonstration in six states: rolled out nationally by end of 2010 with four regional recovery audit contractors (RAC)
Since then...	RACs paid contingency fees (9–12.5% of corrected amounts; 10–14.4% from FY2016), no fee if overturned on appeal. Most improper payments are found in <i>complex</i> reviews that require the medical record from an Additional Documentation Request (ADR)
2011–14	Provider backlash: appealed claims up >500% (2011–13); >800,000 appeals backlogged by 2014; 48% of hospitals report spending >\$25k (11% >\$100k) managing audits; AMA letter on RACs’ “bounty-hunter like tactics”
Oct 2013	RACs barred from inpatient patient-status reviews (Two-Midnight rule era) — Sample starts in 2014

Exogenous Policy Shift - The 2016 ADR limit

When	What happened
Jan 1, 2016	ADR limit: starting on Jan 1, RACs may request records for at most 0.5% of a provider's Medicare claims paid in the prior 12 months. A cap on how many claims can be audited → a sharp drop in audit probability <i>for some hospitals with prior ADR rates > 0.5%</i>, and possibly for all hospitals?
Early 2017	New round of RAC contracts: 3 of 4 original RACs continue — Sample keeps only regions with the same RAC through 2016
2020	Audits paused during the COVID-19 pandemic — Sample ends after 2019

Upcoding in Medicare Part A

Since 1983, hospitals billing Part A paid by Prospective Payment

Each admission is assigned one diagnosis-related group (DRG). Payment $\approx$ DRG weight $\times$ base rate (+ hospital-specific adjustments) — a flat amount regardless of what the stay actually cost.

DRGs sharing a principal diagnosis or procedure form a *base DRG*

Based on the presence of no complication/comorbidity (CC), CC, or major CC (MCC) weight are applied to the base payment. w/ MCC is the **top code** and pays the most.

Example: base DRG “Heart Failure and Shock” (FY2027)

Tier	DRG	Weight	Weight x Base payment (\$6,848)
w/ MCC (Top code)	291	1.2685	\$8 686.69
w/ CC	292	0.9575	\$6 556.96
No CC	293	0.7912	\$5 418.14
Top code – no CC	<i>Potential gains from upcoding</i>		\$3,268.55

Upcoding in Medicare Part A

Since 1983, hospitals billing Part A paid by Prospective Payment

Each admission is assigned one diagnosis-related group (DRG). Payment $\approx$ DRG weight $\times$ base rate (+ hospital-specific adjustments) — a flat amount regardless of what the stay actually cost.

DRGs sharing a principal diagnosis or procedure form a *base DRG*

Based on the presence of no complication/comorbidity (CC), CC, or major CC (MCC) weight are applied to the base payment. w/ MCC is the **top code** and pays the most.

RAP audits claims to try and detect overpayment due to upcoding

Hospitals can select a higher-paying DRG or modifier than the record supports, even if documentation does not support the code.

This paper uses the top-coding rate as the primary outcome

Share of a hospital's discharges in multi-tier base DRGs that are coded to the top tier, which is a mix of *legitimate* and *inappropriate* top codes. Serves as a proxy for upcoding as in Dafny (2005), Silverman & Skinner (2004), Ganju et al. (2021).

Upcoding in Medicare Part A – Literature

Hospitals code to the price

- When Medicare raises the relative payment for a severity tier, the share of discharges coded into that tier rises
- Dominated by For-profits
- No change in admissions or treatment intensity

(Dafny 2005; Dafny & Dranove 2009; Silverman & Skinner 2004; Cook & Averett 2020; Sacarny 2018)

Persistent upward drift in billed severity

- Re-abstraction of Medicare records found about one in five DRGs mis-coded in 1985, typically over-payment
- By 1988, roughly a third of early case-mix growth was coding rather than acuity.
- Highest-severity stays increased through 2014–19

(Hsia et al. 1988, 1992; Carter, Newhouse & Relles 1990; OIG 2021; Crespin et al. 2024)

Coding technology and discretion widen the margin

- EHR makes it easier to document patient complexity and legitimately bill or up-code Medicare
- Hospitals have discretion some over base DRG an ambiguous admission falls into
- Sometimes clinicians don't play ball

(Li 2014; Ganju, Atasoy & Pavlou 2022; Shekhar, Leder-Luis & Akoglu 2026; Sacarny 2018)

Monitoring and enforcement matter

- Audits, fines, whistleblower suits and False Claims Act enforcement deter improper billing
- RAP audits deters wasteful use or coding but raises compliance costs for hospitals

(Shi 2024; Ganju et al. 2022; Howard & McCarthy 2021; Leder-Luis 2025; Groß, Jürges & Wiesen 2021; Becker, Kessler & McClellan 2005; Angerer et al. 2021; League 2024; Olken 2007; Kleven et al. 2011; Duflo et al. 2013; Telle 2013; Avis et al. 2018)

Data: four sources, 2014–2019, hospital-year panel

Source	What it provides	Used for
Recovery Audit Program microdata (universe of audited Medicare inpatient claims)	Claim characteristics plus audit details: date selected, reason, determination, amount recovered	Number of audits, ADRs; recovered overpayments
Healthcare Cost Report Information System (HCRIS)	Medicare inpatient discharges; administrative & general expenses	Total discharges per hospital-year; admin-cost outcome
5% Inpatient Standard Analytic File (LDS)	5% random sample of all Medicare inpatient claims with DRG	Top-coding rate at hospital-year (and hospital-DRG-year) level
AHA Annual Survey	Beds, births, outpatient visits, ownership, system, teaching, physician integration	Hospital characteristics
Inpatient PUF (appendix)	Discharges by hospital × DRG	Denominator for DRG-level audit rates in additional spillover test

Pre-period 2014–2015
Post-period 2016–2019

Sample size ≈5,500 hospitals/year in the SAF → 3,759/year after dropping hospitals whose RAC changed in the 2017 renewal

Estimation Sample: 21,601 hospital-year observations

Empirical Strategy – ADR limit as binding constraint

Hospital audit rate

$\text{AuditRate}_{jt} = (\# \text{ claims selected for RAC review in hospital } j, \text{ year } t) \div (\text{Medicare inpatient discharges in } j, t)$

Hospital top-coding rate

$\text{TopCodeRate}_{jt} = (\# \text{ discharges coded to the top tier of a multi-tier base DRG}) \div (\text{all discharges in those DRGs})$

A hospital is considered **Treated** if hospital j 's average audit rate from 2014 – 2015 exceeds 0.5% (ADR limit)

721 of 4,108 hospitals (17.6%) are treated

A hospital is considered a **Control** if hospital j 's average audit rate from 2014 – 2015 below 0.5%

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

Estimation approaches include TWFE, CSDID, event study to test pre-trends, IV for share of top codes audited using post-period after ADR limit introduced as IV

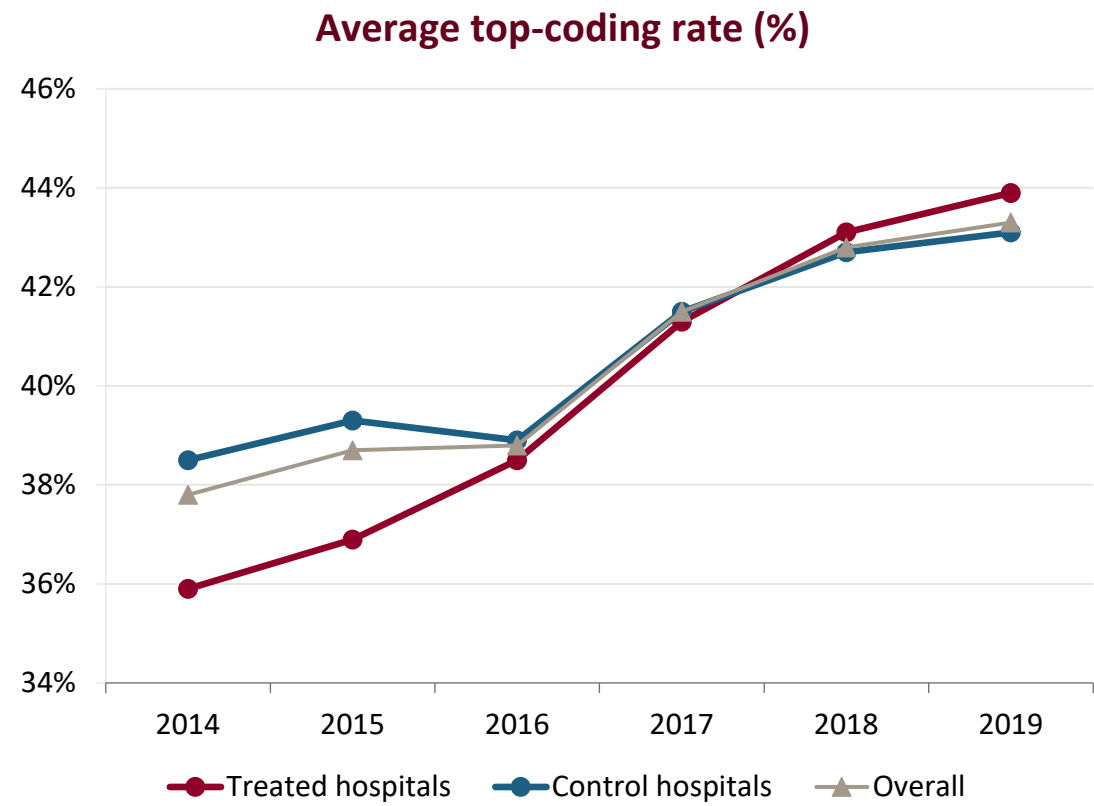
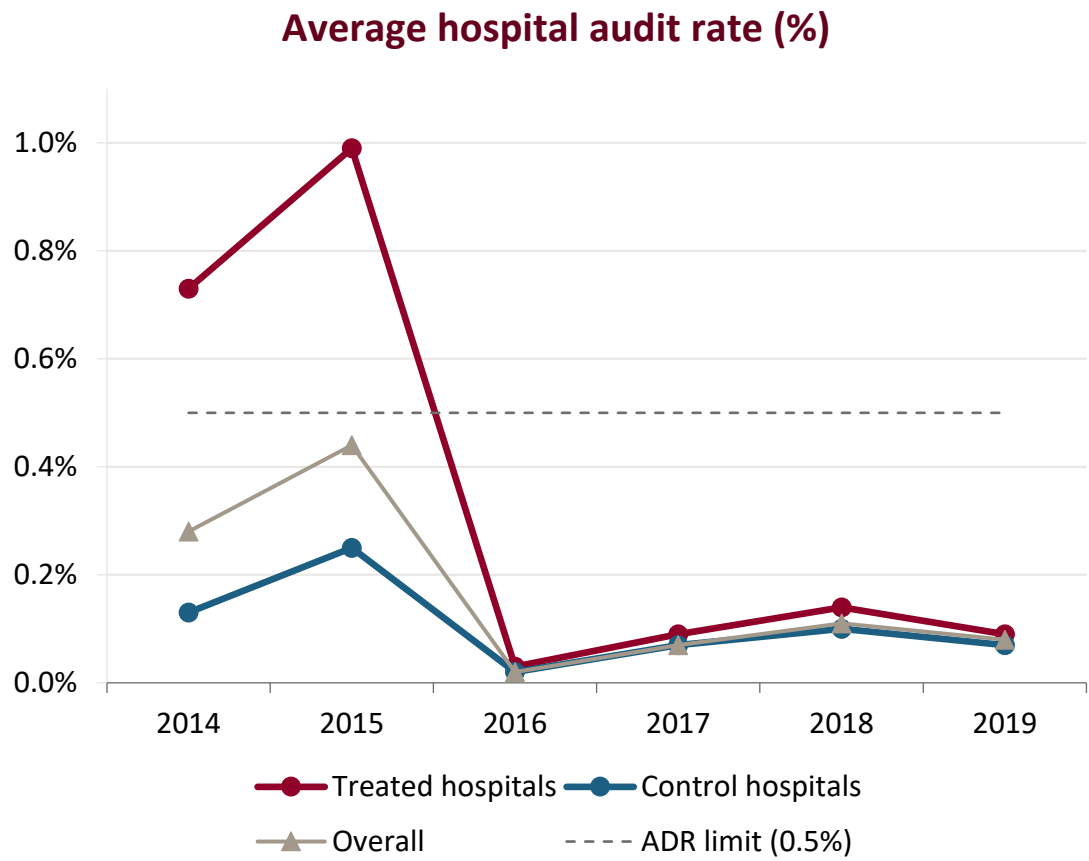
Summary Statistics

Pre-period (2014–2015) means

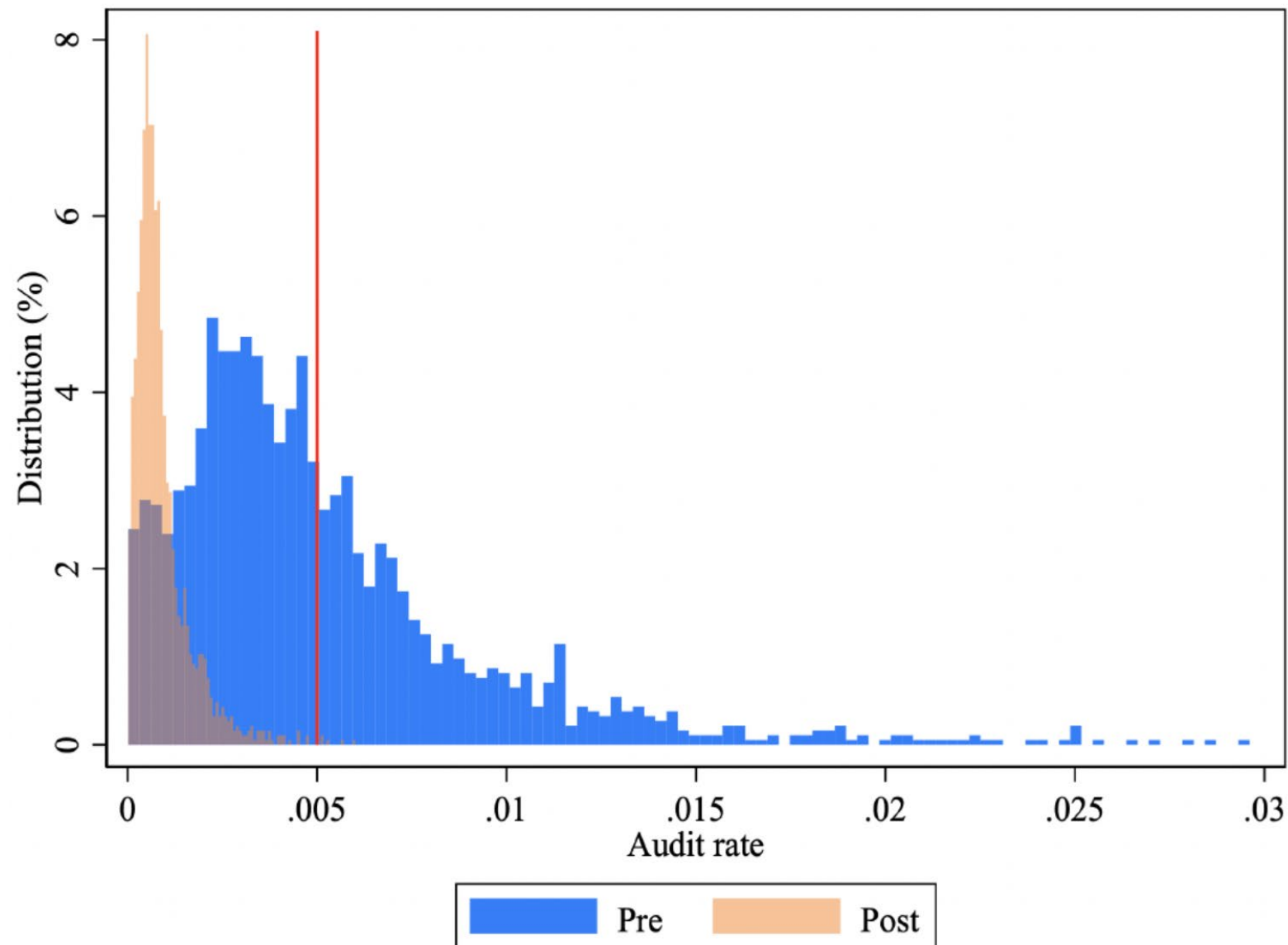
	All	Treated	Controls
Top-coding rate (%)	38.8	32.9	40.0
Audit rate (%)	0.24	0.96	0.09
# top-tier DRGs per hospital	69.8	68.9	70.1
Beds	168	195	160
For-profit (%) / Not-for-profit (%)	29.7 / 41.0	22.7 / 54.8	31.8 / 36.8
Teaching (%) / System-affiliated (%)	5.3 / 53.5	7.0 / 57.5	4.7 / 52.3

Almost 97% of control DRGs always have zero audit rates post treatment

What the raw series look like

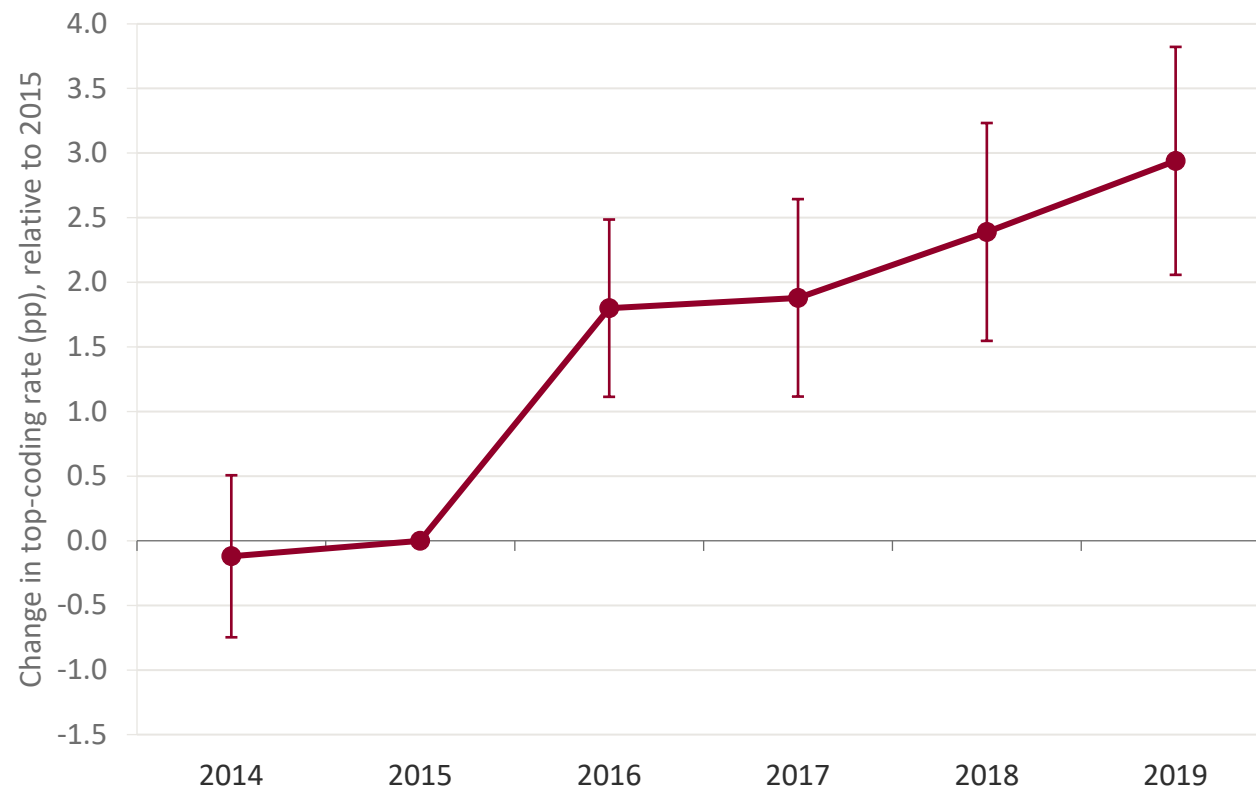


Frequency distribution of average hospital audit rates, pre and post



Main result: top coding rises at treated hospitals following ADR limit

Event-study coefficients, 95% CI

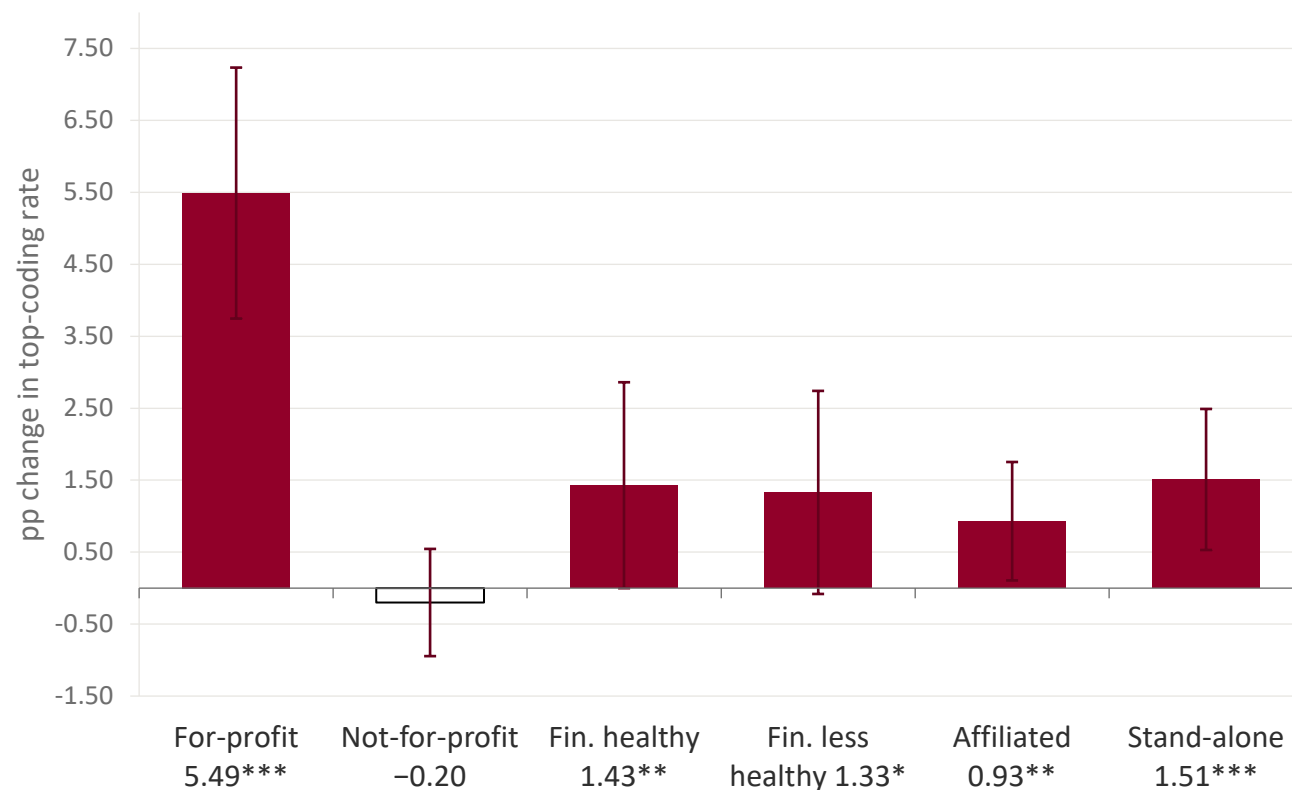


Main Specification	TWFE	CSDiD
Treatment effect (pp)	2.31***	2.23***
(s.e.)	(0.33)	(0.34)
Pre-trend joint test, p	0.71	0.68
N (hospital-years)	21,601	

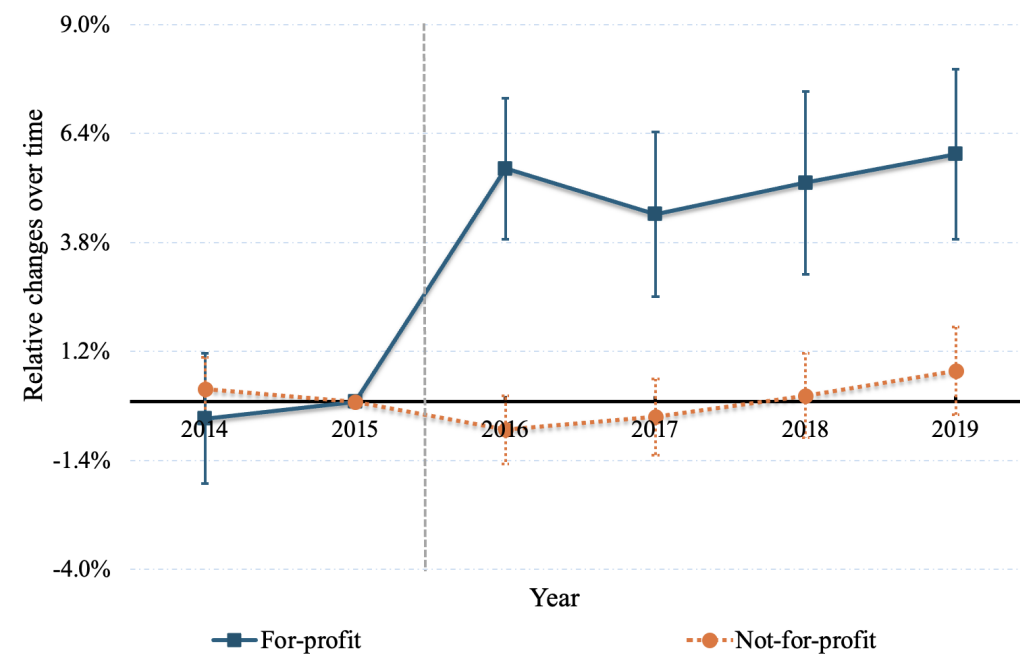
Treated hospitals with pre-period audit rates above the ADR limit (0.5%) increased top-coding by 7% more relative to hospitals with pre-period audit rates below 5%

Heterogeneity – For-profit hospitals drive upcoding response

Treatment effect by hospital type 95% CI



Effect of reduced auditing on top-coding rate



Robustness - results similar across specifications

Is it patient severity or treatment intensity? Alternative outcomes

Outcome (TWFE)	Effect	(s.e.)	Pre-trend p
Average # diagnoses	-0.114	(0.089)	0.06
Average length of stay	-0.040	(0.033)	<0.001
Avg. Elixhauser Comorbidity Index	-0.035**	(0.016)	0.99

No sign that patients at treated hospitals get sicker relative to controls — if anything, slightly fewer comorbidities. (The # diagnoses and LOS series show positive 2014 coefficients, some evidence of non-parallel pre-trends.)

Other alternative specifications

Specification	Effect (TWFE)	(s.e.)	N
Baseline (Table 3)	2.31***	(0.33)	21,601
+ hospital controls	1.14***	(0.32)	16,401
+ controls × pre/post	1.28***	(0.33)	16,401
dCDH estimator	2.26***	(0.34)	16,401
Balanced panel	2.25***	(0.33)	18,606
Without regression weights	11.28***	(0.54)	22,173

Scaling by exposure: the IV estimate in “per percentage point of audit pressure”

Structural equation

$\text{TopCodeRate}_{jt} = \beta \cdot \text{AuditRate}_{jt}^{\text{top}} + \gamma_j + \lambda_t + \varepsilon_{jt}$ $\text{AuditRate}_{jt}^{\text{top}} = \text{top-coded claims reviewed by RACs} \div \text{all top-coded claims in } j, t$

Instrument Relevance and Exclusion

- $\text{AuditRate}_{jt}^{\text{top}}$ likely endogenous: RACs choose whom to review, possibly based on hospital characteristics that also influence top-coding
- Instrument: $\mathbb{1}\{t > 2015\} \times \text{Treat}_j$ (Authors discuss as a “degenerate shift-share”: one shock, binary exposure (Borusyak, Hull & Jaravel 2022))
- Needs the same parallel-trends assumption plus exclusion: Presence of ADR cap moves top coding only through audit pressure

SubSample	β_{IV}	(s.e.)
All hospitals	−3.30***	(0.51)
For-profit	−14.97***	(2.72)
Not-for-profit	0.25	(0.48)
Financially healthy ; less healthy	−2.82** ; −2.10*	(1.44) ; (1.18)
Affiliated ; Stand-alone	−1.56** ; −1.74***	(0.72) ; (0.59)

Economic Impact of ADR limits

Fewer recoveries

- Re-run TWFE with *recovered overpayments* as the outcome: DiD estimate **-\$107,927** per treated hospital per year
- Of which RAC contingency fees (10.75% = avg. of 9% and 12.5%): **\$11,602**
- Net recoveries Medicare no longer gets: **\$96,325**

More top codes

- Median Medicare discharges per treated hospital (2014–15): **811**
- $\times 2.3$ pp more top codes $\times 0.9$ average weight spread (top – bottom tier) $\times$ \$5,961 per weight unit
- = **\$100,071** extra payments per treated hospital per year

Net Loss for Medicare

- Loss per \$1 saved on fees: $(\$96,325 + \$100,071) \div \$11,602 \approx$ **\$17**
- **51%** of the loss is the coding response;
- **49%** is overpayments that would have occurred anyway but now go undetected
- No detectable reduction in hospitals' administrative costs

Findings and Conclusions

- 1 Hospitals code more aggressively when audit pressure falls**
+2.3 pp ($\approx 7\%$) more top codes at treated hospitals; immediate and growing through 2019
- 2 The response comes from for-profit hospitals**
5.5 pp among for-profits vs. ≈ 0 among not-for-profits; little difference by financial health; smaller in systems.
- 3 No robust cross-DRG spillovers, but they cannot be ruled out**
Hospital-level analysis robust to across DRG-spillovers of changes in upcoding behavior
- 4 Scaling back cost Medicare $\approx$ \$17 per \$1 saved on RAC fees**
Compliance costs sunk, deterrence is recurring. *“Simply scaling back the program might invite improper billing practices without lowering providers’ administrative burden.”*

Alternative interpretations of the findings – Hospital Response

Less defensive down-coding, not more upcoding

- Hospitals under heavy audit may have coded *legitimate* MCCs down to avoid the hassle, even when they would win on appeal
- Some increases in top-coding may be legitimate
- **Can you look at trends in appeals?**

A mechanical channel in the outcome data?

- Are the DRG codes in the data what was submitted in the claim initially, or the final DRG after adjudication from auditing.

Catch-up to a secular trend

- Treated hospitals start from *lower* top-coding rates
- Were treated hospitals behind on clinical-documentation programs and catching up?
- OIG (2021) documents a nationwide shift toward highest-severity stays over 2014–2019

Coincident shocks

- ICD-10-CM/PCS went live on October 1, 2015, after years of delays.
- National change is absorbed by year FE only if its effect on CC/MCC coding was the same across hospitals, was there some adoption patterns?

Why the hospital level? Upcoding spillovers across DRGs

Consider a DRG-level DiD analysis that exploits DRG-level variation (Dafny 2005; Li 2014; Ganju et al. 2021) on a sample of codes with low audit rates in the pre-period

$$\text{TopCodeRate}_{djt} = \beta \times \mathbb{1}\{t > 2015\} \times \text{Treat}_j + \gamma_{jd} + \eta_{dt} + \lambda_t + \varepsilon_{djt}$$

Where γ_{jd} is a DRG-in-hospital FE – will absorb flat changes in top-coding for each DRG within a hospital consistent across those DRGs at different hospitals. η_{dt} will absorb secular trends individual DRG volume

Under SUTVA, Decreased audit monitoring in frequently up-coded DRGs cannot influence coding in DRGs where there is no change in monitoring (e.g., audit rate below 0.5%)

If fail to reject $\beta < 0$, suggestive of declining top-coding for untreated DRGs

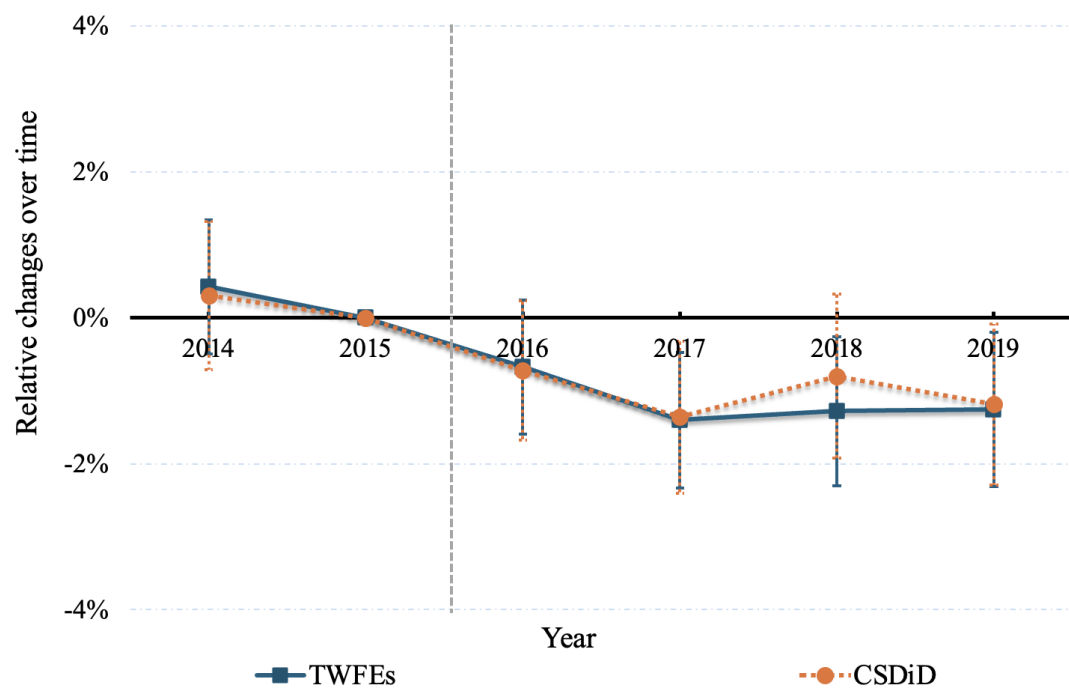
- Lack of internal capacity/will to upcode more (Sacarny, 2018)
- Up-coding resources shifted to DRGs with lessened oversight

If fail to reject $\beta > 0$, suggestive of *increased* top-coding for untreated DRGs

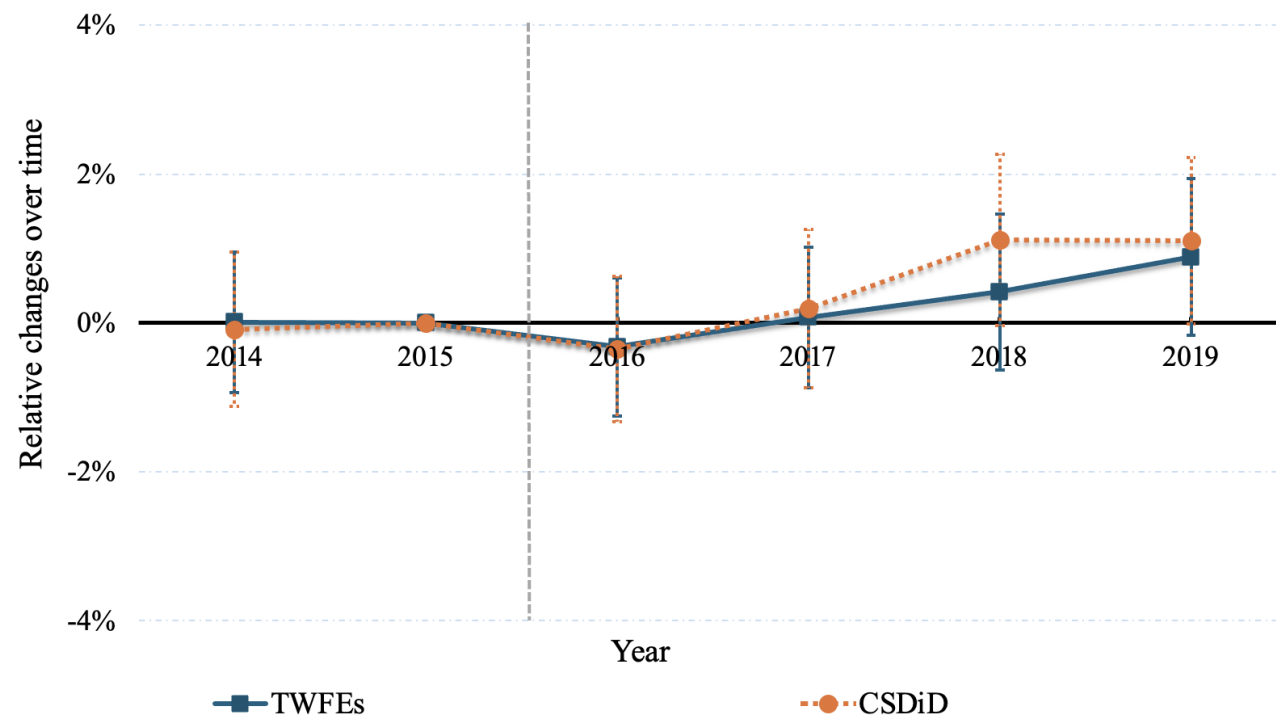
- Hospital may be seeking to recoup lost profits from scaling back up-coding on more scrutinized DRGs

Why the hospital level? Upcoding spillovers across DRGs

Without DRG-year FE



With DRG-year FE (preferred)



SUTVA assertion depends on knowing who receives what exposure to treatment

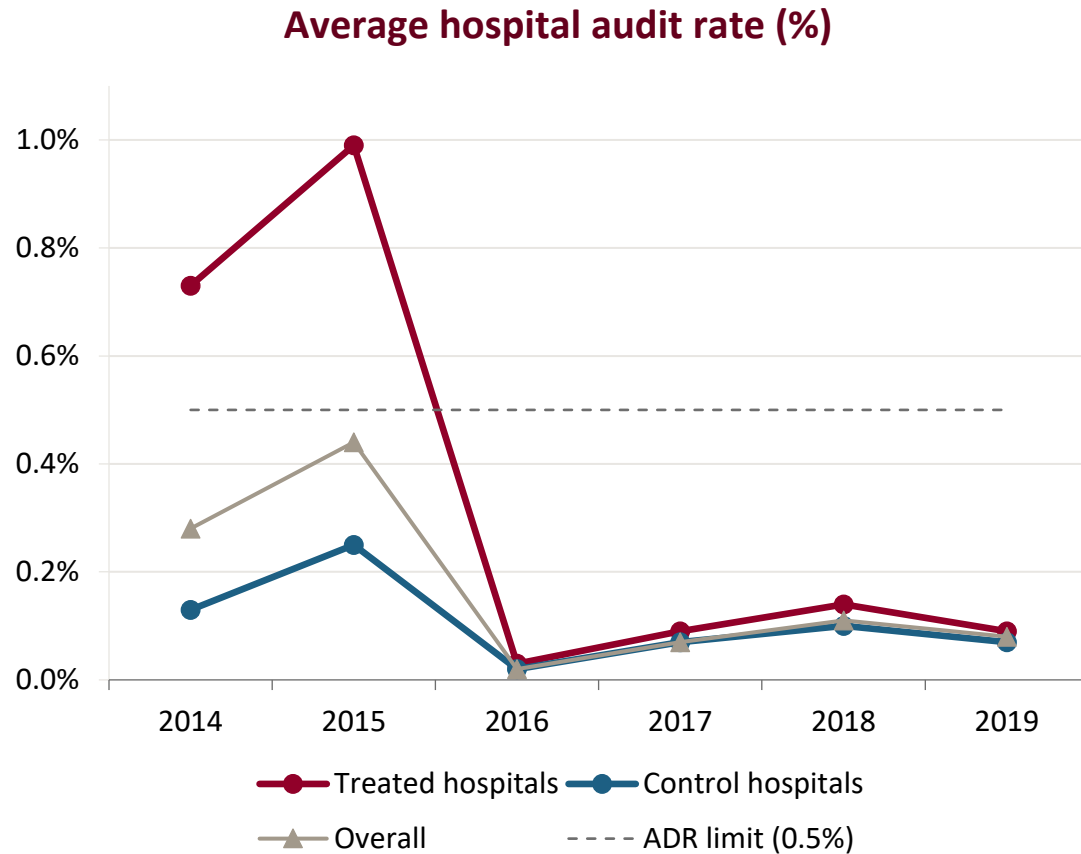
$\text{TopCodeRate}_{djt} = \beta \times \mathbb{1}\{t > 2015\} \times \text{Treat}_j + \gamma_{jd} + \eta_{dt} + \lambda_t + \varepsilon_{djt}$ for a subsample of zero audit rate DRGs

- *Hospitals have some leeway to assign base DRG*
- *DRGs with zero audits in pre-period may be fundamentally different in hospital's ability to up-code, regardless of monitoring (which DRGs are these? How common are they?)*
 - *For instance, surgical base DRGs*
- *Is the spread in DRG weights similar to DRGs with more frequent upcoding*

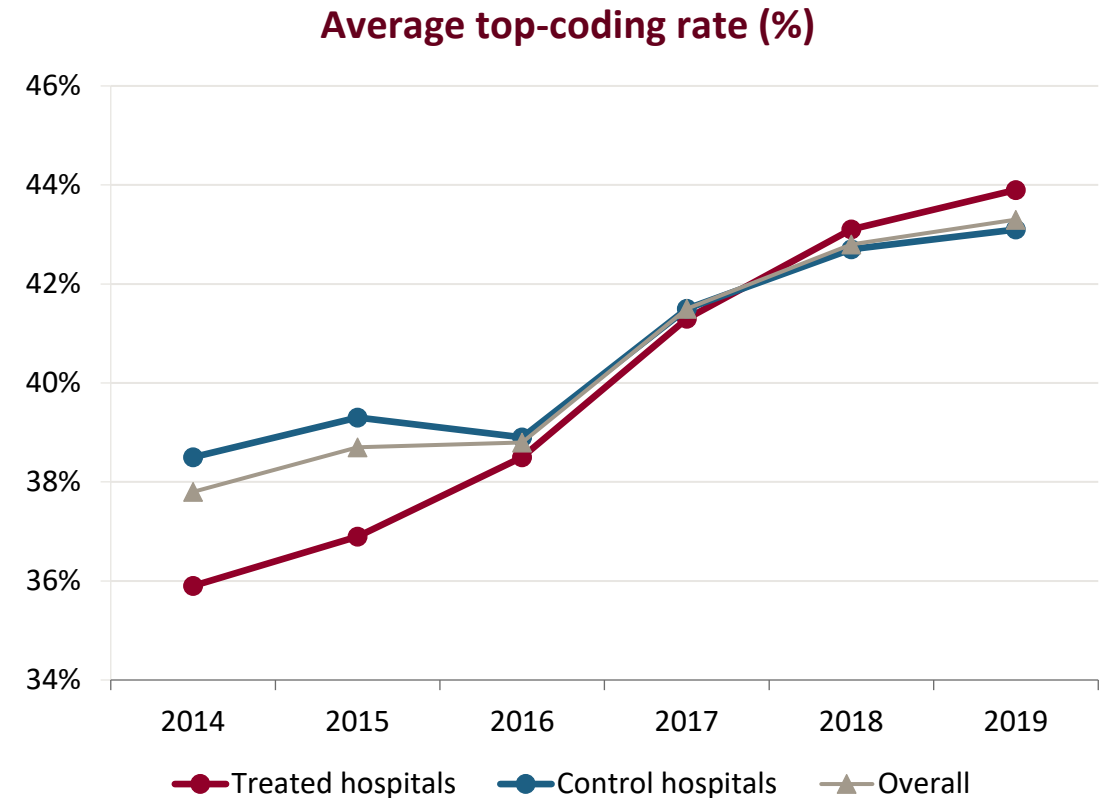
$\text{TopCodeRate}_{jt} = \beta \times \mathbb{1}\{t > 2015\} \times \text{Treat}_j + \gamma_j + \lambda_t + \varepsilon_{jt}$ for a sample of hospitals

- *Reduced form model comingles legitimate top-coding and up-coding to the top code*
- *Effects identified from changes in top-coding rate, assuming underlying legitimate top-coding evolves similarly for treated and control hospitals.*
- *Then why does the trend in top-coding among **control** hospitals seem to change after ADR limit in 2016, not the **treated** hospitals*

A second look at the raw trends



Auditing drops for both treated and control groups, then is nearly identical for all hospitals



Positive DiD effect comes from differencing out a flat control pre-trend from rising top-coding trend among treated hospitals in pre-period

What is treatment exactly?

Treatment is the dose-response to changing likelihood of ADR request

*In this **dose–response** story, treated hospitals with larger reductions in audit rates would **overshoot** the controls in their increase in top-coding, not land on top of them*

Treatment is the binding constraint of the 0.5% limit in ADR requests

We would expect auditing of ADR hospitals to fall to just below ADR limit, and no movement among control hospitals. Instead both audit rates converge to new lows

Treatment is the mere introduction of the ADR limits

Best identified using the main specification, then outcome is merely heterogeneity in response among the equally treated sample (i.e., no treatment/control groups)

Do RACs adjust audit rates based on prior knowledge of hospital top-coding

Seems not to be the case, or audit rates would not converge. Likewise, IV estimates similar to TWFE estimates. Then why does the ADR rate drop so low below the ADR limit for both treatment and control hospitals?

RAC auditing constraints – Dynamic ADR limits (

The 0.5% limit is a baseline in the first of eight 45 day cycles

RACs are constrained to make similar numbers of ADR requests in each 45 day cycle

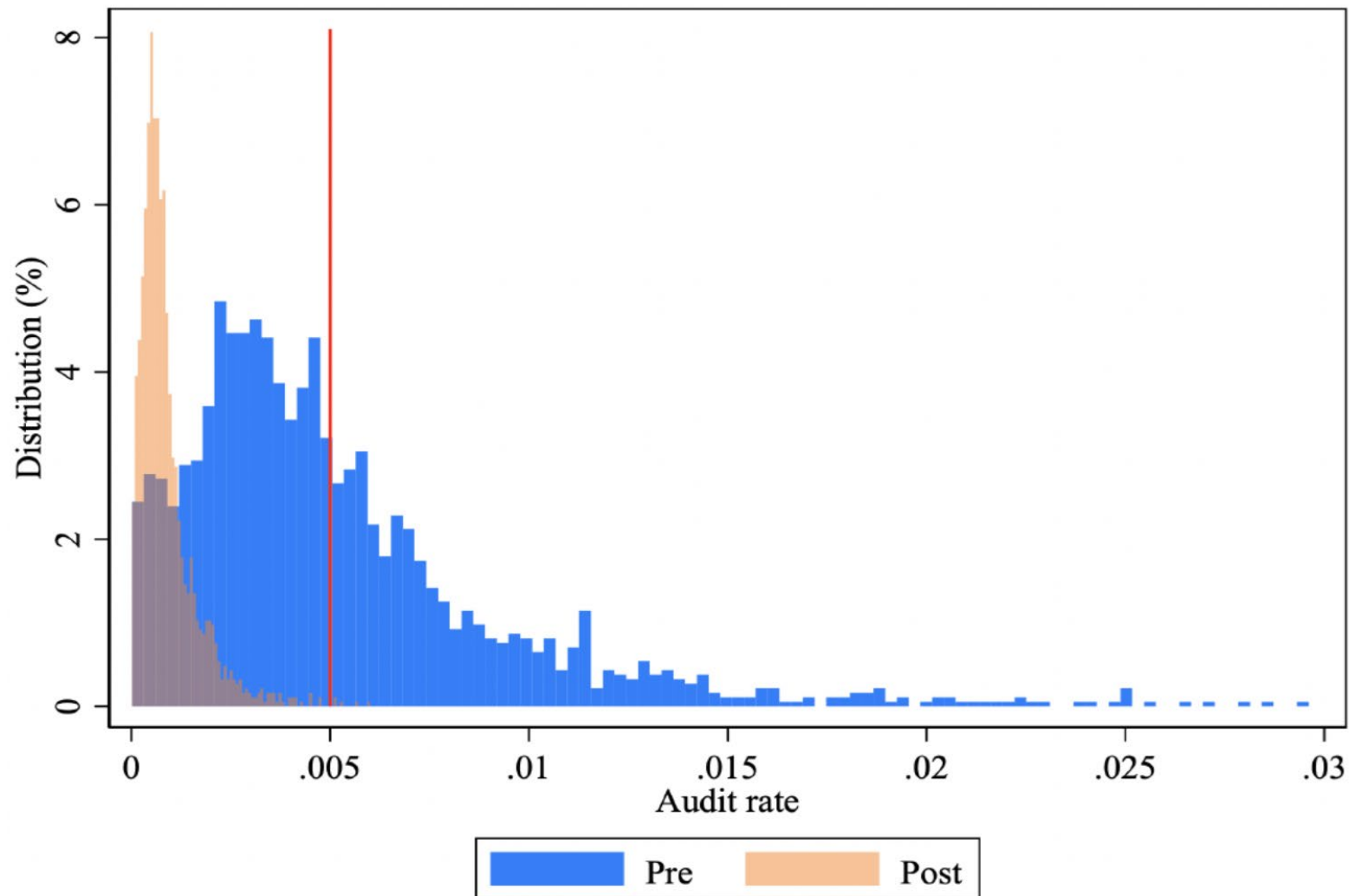
After three 45 day cycles, CMS calculates the provider's denial rate (# overpaid claims less appealed claims /# claims reviewed), using the most recent 12 months of completed reviews

CMS sets the next three cycles' limit from a tier table, recalculated every 3 cycles:

Denial rate	Adjusted annual ADR limit		Denial rate	Adjusted annual ADR limit
0–3%	No reviews for the next three cycles (then back to 0.5%)		36–50%	1.5%
4–9%	0.25%		51–70%	3.0%
10–20%	0.5% (baseline)		71–90%	4.0%
21–35%	1.0%		91–100%	5.0%

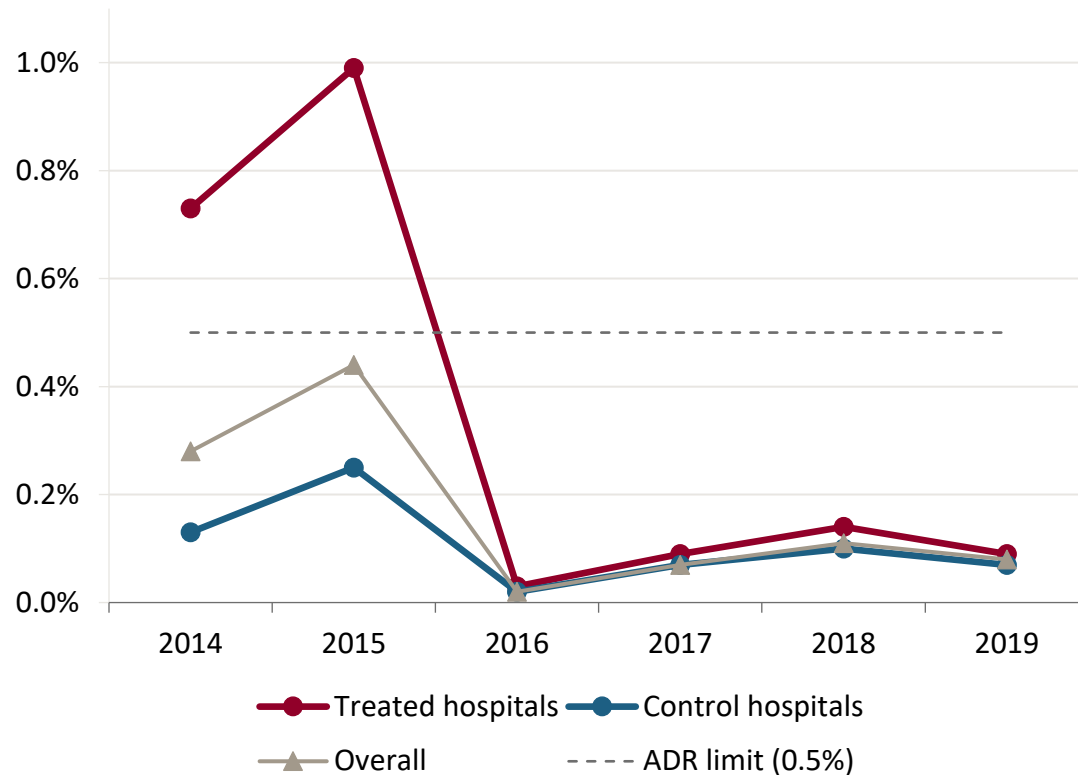
[CMS ADR limits for institutional providers, May 2016 \(rev. Dec 2018\)](#) · [CMS ADR limits for institutional providers, May 1, 2022](#) · [CMS Recovery Audit Program Improvements, Nov 24, 2017](#) · [CMS ADR limits, Jan 29, 2018 version](#)

Post-2016 audit exposure is endogenous to each hospital's own recent audit outcomes



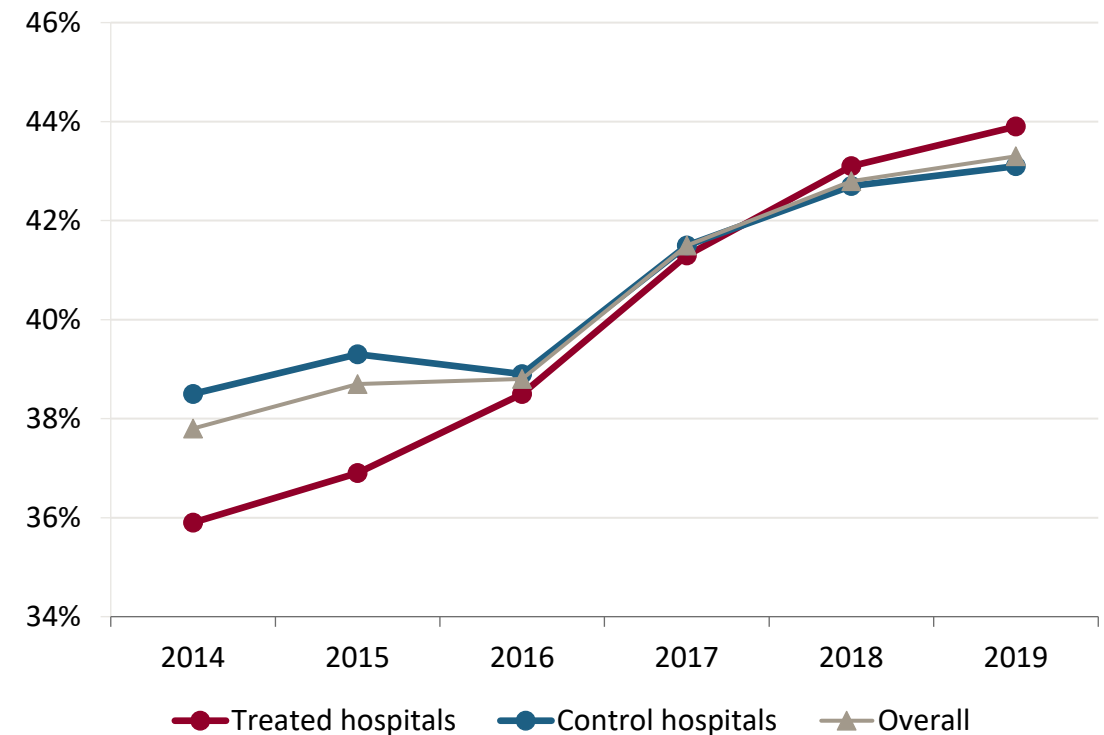
A third look at the raw trends

Average hospital audit rate (%)



Under dynamic limits, possible that most pre-period ADRs did not find up-coding as often as expected (contingent on appeals)

Average top-coding rate (%)



Can also be explained by hospital learning to up-code beyond RAC budget/capacity to catch, explains why previously low-audit hospitals up-code more over time

Some suggestions

Other up-coding outcomes

- Hospitals may “nickel and dime” by up-coding to middle tier DRG (CC) in three-tier base DRGs
- ER visits resulting in admissions w/n 3 days?
- What about Part B?

Sample Management

- Include all RACs and use Region FEs instead?
- Consider omitting small hospitals or low discharge hospitals

Investigate Appeals

- Show how pre-period denial rates vary with audit rates
- Should be doable in the audit microdata (determinations, appeals)

Investigate Appeals

- Show how pre-period denial rates vary with audit rates
- Should be doable in the audit microdata (determinations, appeals)

Avoid broad group means

- Evolution of top-coding within finer groups of pre-period audit rates can show distinct patterns
- Using two group means suggests convergence of top-coding

Exploit dynamic ADR limits in post-period

- Potential to use alternative specifications, such as a dynamic fuzzy RD using ADR cutoffs
- Look for ADR bunching below, appeals above cutoffs

Contributions and future research areas

- 1 Hospitals code more aggressively when audit pressure falls**
Hospitals may be very sophisticated in how they up-code and appeal, possibly new incentives to up-code
- 2 The response comes from for-profit hospitals**
Worst offenders may outpace other hospitals in top-coding (beyond apparent convergence), may be more strategic as well
- 3 No robust cross-DRG spillovers, but they cannot be ruled out**
Sign of coefficients in spillover models can identify aspects of hospitals' response to changes in oversight.
- 4 Scaling back cost Medicare $\approx$ \$17 per \$1 saved on RAC fees**
CMS cares a lot about accounting gains. We also care about welfare gains. Only monitoring costs, compliance costs, and undetected improper payments are DWL.

LAFAYETTE

DEPARTMENT OF ECONOMICS

Thank you!

Adam I. Biener · bienera@lafayette.edu

Department of Economics, Lafayette College

American Health Econometrics Workshop · October 2026