The Inspection Lottery: How Regulatory Stringency Crowds Out Nursing Home Staffing

APEP Autonomous Research* @olafdrw

March 30, 2026

Abstract

Nursing home quality depends on staffing, but does regulatory inspection improve it? I exploit quasi-random variation in state survey agency stringency—the "inspection lottery"—to identify the causal effect of deficiency citations on nurse staffing. Using leave-one-out state mean severity as an instrument for 11,427 facilities in 2025, I find that stricter enforcement *reduces* total nurse staffing by 0.73 hours per resident per day per severity unit (5.2% of the mean for a one-standard-deviation stringency shift). The effect is confirmed within multi-state chains sharing identical management (–0.59 with chain fixed effects), concentrated in for-profit facilities, and accompanied by a 12.7 percentage point increase in nursing staff turnover. These results reveal a compliance crowding mechanism: regulatory pressure diverts resources from clinical care to administrative remediation.

JEL Codes: I11, I18, L51, J23

Keywords: nursing homes, health inspection, regulatory enforcement, staffing, compliance costs

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 25m).

1. Introduction

Every year, a state surveyor walks into each of America's 15,000 nursing homes, clipboard in hand, and decides how many deficiencies to cite and how severely to grade them. The consequences are real: citations trigger mandatory remediation plans, follow-up inspections, and potential financial penalties. But here is the paradox—identical conditions in identical chain-operated facilities receive wildly different regulatory treatment depending on which state's agency conducts the survey. The Government Accountability Office found that 70 percent of state surveys miss deficiencies identified by federal follow-up teams, and 15 percent miss serious ones (Government Accountability Office, 2008). This variation is not merely statistical noise. It is a natural experiment.

The policy stakes are large. Nursing homes serve 1.3 million residents, consume over $100 billion annually in Medicare and Medicaid spending, and have been the subject of persistent quality concerns (Bowblis, 2011; Harrington et al., 2012). The conventional wisdom holds that stricter regulation improves quality: more citations discipline facilities into investing in better care, primarily through higher staffing (Mukamel et al., 2012; Castle et al., 2011). If true, the dramatic cross-state variation in enforcement intensity represents a massive quality gap—and the policy response is straightforward standardization. But this paper tells a different story.

I use state survey agency stringency as an examiner-leniency instrument for deficiency severity, in the spirit of judge-IV designs (Kling, 2006; Doyle, 2007). The instrument is the leave-one-out mean scope-severity score assigned by the same state agency to *other* facilities in the same year. The identifying assumption is that, conditional on facility characteristics, a nursing home's surveyor regime is determined by geography, not by the facility's own quality choices. Multi-state chains provide the sharpest test: a Genesis Healthcare facility in strict New Hampshire faces a different regulatory regime than a Genesis facility in lenient Louisiana, despite identical corporate governance and budget policies.

The first stage is powerful: a one-unit increase in state stringency raises the facility's mean severity score by 0.94 (robust $F = 4{,}764$). But the second stage delivers a surprise. Instrumented deficiency severity *reduces* total nurse staffing by 0.73 hours per resident per day (HPRD), significant at the 5 percent level ($p = 0.024$). This is 15 times larger than the OLS estimate of –0.05, and it goes the wrong way for the enforcement-improves-quality hypothesis. Facilities facing stricter surveyors invest *less* in clinical staffing, not more.

The within-chain specification confirms this finding. Among 8,192 chain-affiliated facilities spanning 604 chains, the IV estimate with chain fixed effects is –0.59 HPRD ($p = 0.003$). Since chain fixed effects absorb all corporate-level policies, this estimate reflects pure within-chain variation in regulatory exposure—the closest feasible approximation to random surveyor

assignment.

The mechanism I identify is *compliance crowding*. Each deficiency citation triggers a cascade of administrative responses: plans of correction must be drafted, submitted, and implemented; follow-up surveys must be prepared for; staff time is redirected from patient care to documentation. I decompose the staffing reduction into its components: registered nurses lose 0.20 HPRD, licensed practical nurses 0.25, and certified nursing assistants 0.28. The reduction is spread across all staff types, with the lowest-paid caregivers bearing the largest share. Simultaneously, nursing staff turnover rises by 12.7 percentage points per unit of instrumented severity ($p = 0.002$), consistent with the hypothesis that regulatory burden drives clinical workers away from cited facilities.

This paper makes three contributions. First, it provides the first examiner-leniency IV estimates of the causal effect of regulatory inspection stringency on nursing home staffing. While a large literature studies nursing home quality determinants (Grabowski, 2004; Bowblis, 2011; Konetzka et al., 2008), no prior work exploits the quasi-random assignment of surveyor regimes to identify the causal chain from citations to staffing investment. Second, it identifies the compliance crowding mechanism, adding to the literature on unintended consequences of regulation (Bardach and Kagan, 2002; Short and Toffel, 2010; Parker and Nielsen, 2009; Johnson, 2020). This mechanism is distinct from the well-studied compliance costs literature, which focuses on firms exiting the market; here, firms *stay* but redirect resources internally. Third, the multi-state chain design provides a novel identification strategy that separates management quality from regulatory pressure, extending the examiner-leniency framework to a setting where the "examiners" are state agencies rather than individual judges.

The results have direct policy implications for the Centers for Medicare and Medicaid Services (CMS). If stringent enforcement crowds out the very staffing it is supposed to promote, then uniform citation standards—a frequent policy proposal—may be counterproductive without complementary reforms that reduce the administrative burden of compliance. The findings suggest that the optimal regulatory design must balance deterrence against the real resource costs that enforcement imposes on frontline care.

The paper proceeds as follows. Section 2 describes the institutional background of nursing home regulation. Section 3 presents the data. Section 4 develops the empirical strategy. Section 5 reports results. Section 6 discusses implications and limitations.

2. Institutional Background

The Federal-State Survey System. Medicare and Medicaid certification requires nursing homes to meet federal quality standards under 42 CFR Part 488. CMS delegates inspection

authority to state survey agencies, which conduct unannounced standard health inspections approximately annually. Surveyors assess compliance with over 180 federal requirements covering areas from infection control to resident rights (Centers for Medicare and Medicaid Services, 2017). When a violation is identified, it is classified on a scope-severity matrix with 12 cells, from A (isolated deficiency, no actual harm, potential for minimal harm) to L (widespread deficiency, immediate jeopardy to resident health or safety). The citation severity determines the required corrective action and potential penalties.

Cross-State Variation in Enforcement. The decentralized structure creates substantial variation in enforcement intensity. States differ in surveyor staffing ratios (from 2.5 to 12.5 surveyors per 1,000 nursing home residents), training protocols, citation norms, and organizational culture (Government Accountability Office, 2008; Centers for Medicare and Medicaid Services, 2019). Several studies have documented that identical conditions elicit different regulatory responses across states (Harrington et al., 2004; Mukamel et al., 2012). Federal comparative surveys, where CMS resurveys a sample of recently state-surveyed facilities, consistently find that state agencies undercount deficiencies relative to federal teams, with the degree of undercounting varying by state (Government Accountability Office, 2008).

Consequences of Citations. Deficiency citations trigger a graduated enforcement response. At the lower end (A–F, no actual harm), facilities must submit plans of correction and may face follow-up surveys. At the moderate level (G–I, actual harm), facilities face directed plans of correction and potential civil monetary penalties (CMPs) of up to $21,393 per day. At the highest level (J–L, immediate jeopardy), facilities face immediate CMPs, potential termination from Medicare/Medicaid, and state monitoring (Centers for Medicare and Medicaid Services, 2017). The administrative burden of responding to citations is substantial: plans of correction require detailed documentation, implementation, and verification; resurveys consume staff time in preparation and during the visit.

Multi-State Chains. Approximately 69 percent of U.S. nursing homes are affiliated with chains. The largest—Ensign (14 states), Sabra Healthcare (43 states), and Genesis Healthcare (25 states)—operate facilities across diverse regulatory environments. Chain management typically sets corporate-wide policies on staffing ratios, budget allocation, and quality improvement, creating a setting where within-chain variation in regulatory exposure is driven by geography rather than management quality (Grabowski, 2004; Bowblis, 2011).

3. Data

Sources. I use four datasets from the CMS Provider Data Catalog, all publicly available without authentication. The Health Deficiencies dataset contains 418,972 individual deficiency citations with facility identifiers, survey dates, deficiency tags, and scope-severity codes. The Provider Information dataset provides facility characteristics for 14,703 Medicare/Medicaid-certified nursing homes, including certified bed count, ownership type, chain affiliation, and reported nurse staffing hours per resident per day from Payroll-Based Journal (PBJ) submissions. The Quality Measures dataset contains 249,951 MDS-derived quality indicators. I focus on the 2025 survey year (the latest complete year) for the cross-sectional analysis.

Variable Construction. I convert scope-severity codes to a numeric scale (A = 1, ..., L = 12) and aggregate to the facility level, computing mean severity, number of deficiencies, and the proportion classified as actual harm or immediate jeopardy (codes G–L). The leave-one-out (LOO) state stringency instrument for facility i in state s equals the mean severity score of all *other* facilities in state s, excluding facility i. Staffing outcomes use PBJ-reported hours per resident per day (HPRD) for total nursing staff, registered nurses (RN), licensed practical nurses (LPN), and certified nursing assistants (CNA). Turnover is the annual percentage of nursing staff who separated from the facility.

Sample. The analysis sample includes 11,427 facilities with non-missing staffing data and valid instrument values. The chain subsample comprises 8,192 facilities across 604 chains.

3.1 Summary Statistics

Table 1: Summary Statistics

	Mean	Std. Dev.	Min	Max
Panel A: Deficiency measures				
Mean scope-severity score (1–12)	4.63	0.88	2.00	12.00
Number of deficiencies	8.9	8.2	1	97
Pct. with actual harm or jeopardy	8.0	18.4		
Panel B: Staffing outcomes (HPRD)				
Total nurse staffing	3.86	0.88	0.01	15.93
Registered nurses (RN)	0.66	0.44		
Licensed practical nurses (LPN)	0.86	0.36		
Certified nursing assistants (CNA)	2.33	0.56		
Total nursing staff turnover (%)	46.9	14.6		
Panel C: Instrument				
LOO state stringency index	4.63	0.24	4.09	5.42
Panel D: Facility characteristics				
Certified beds	108.8	58.1		
Average daily residents	86.1	48.4		
For-profit (%)	75.7			
In chain (%)	71.7			

Notes: N = 11,427 Medicare/Medicaid-certified nursing homes surveyed in 2025. Staffing hours per resident per day (HPRD) from CMS Payroll-Based Journal submissions. Mean scope-severity score averages the CMS A–L grid (A = 1, L = 12) across all deficiency citations from the facility's most recent standard health inspection survey. Leave-one-out (LOO) state stringency is the mean severity score for all *other* facilities in the same state and year.

Table 1 reports summary statistics. The mean deficiency severity is 4.63 on the 1–12 scale, with 8 percent of citations at the actual harm or jeopardy level. Total nurse staffing averages 3.86 HPRD, with substantial variation (SD = 1.24). The LOO state stringency instrument ranges from 4.09 to 5.42 (cross-state SD = 0.29), providing meaningful variation for identification. Seventy-six percent of facilities are for-profit, and 72 percent are chain-affiliated.

4. Empirical Strategy

4.1 Identification

I estimate the effect of deficiency severity on staffing investment using two-stage least squares (2SLS):

First stage:

$$\text{Severity}_i = \alpha + \gamma Z_{s(i)} + X_i'\beta + \varepsilon_i \tag{1}$$

Second stage:

$$Y_i = \mu + \delta \widehat{\text{Severity}}_i + X_i'\lambda + u_i \tag{2}$$

where Y_i is a staffing outcome for facility i, Severity_i is the endogenous mean deficiency severity score, $Z_{s(i)}$ is the leave-one-out state stringency instrument, and X_i includes log certified beds and ownership type indicators. Standard errors are clustered at the state level to account for within-state correlation in the instrument.

The instrument $Z_{s(i)}$ assigns to each facility the average severity score of all other facilities surveyed by the same state agency in the same year:

$$Z_{s(i)} = \frac{1}{N_s - 1} \sum_{j \in s, j \neq i} \overline{\text{Severity}}_j \tag{3}$$

Identifying Assumptions. Instrument relevance requires that state agency stringency predicts facility-level deficiency severity. This is strongly satisfied: a one-unit increase in $Z_{s(i)}$ raises facility severity by 0.94 (first-stage $F = 4{,}764$). The exclusion restriction requires that state stringency affects staffing outcomes only through the deficiency citations it generates. The key threat is that state-level factors (Medicaid reimbursement rates, nurse labor markets, regulatory culture) may independently affect both stringency and staffing. The within-chain specification addresses this by absorbing all chain-level confounders; the remaining variation is purely geographic. I also report balance tests and weak-IV-robust Anderson-Rubin inference.

Estimand. The 2SLS coefficient δ estimates the average causal response of staffing to deficiency severity among facilities whose measured severity is shifted by state stringency—the population of facilities that would receive different citations under a different surveyor regime. With a continuous instrument and continuous treatment, this is an average causal response along the margin moved by the instrument, analogous to the LATE interpretation in binary settings (Angrist et al., 1996).

4.2 Threats to Validity

State-level confounders. The primary concern is that state characteristics correlated with survey stringency may independently affect staffing. Balance tests (Section 5) show that log bed count and for-profit status are uncorrelated with the instrument. The urban indicator shows some imbalance ($\hat{\beta} = -0.30$, SE = 0.12), suggesting that stricter states tend to have more rural facilities. I note this as a limitation; however, the within-chain specification, which compares facilities in different states within the same chain, substantially mitigates this concern.

Monotonicity. The instrument shifts all facilities in the same direction: stricter state agencies issue higher severity scores across the board. This is supported by the near-unity first-stage coefficient (0.94), indicating a roughly one-for-one relationship between state stringency and facility severity.

5. Results

5.1 OLS and First Stage

Table 2: OLS Estimates and First Stage

	OLS: Total HPRD		First Stage: Mean Severity		
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Mean severity	-0.0480*	-0.0455*			
	(0.0261)	(0.0258)			
Panel B: First Stage					
LOO state stringency			0.9407***	0.9391***	0.8045***
			(0.0128)	(0.0136)	(0.0742)
Controls	No	Yes	No	Yes	Yes
Chain FE	No	No	No	No	Yes
Observations	11,427	11,427	11,427	11,427	8,192
First-stage F			5398.2	4763.6	117.7

Notes: Standard errors clustered at the state level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns 1–2 report OLS estimates of deficiency severity on total nurse staffing HPRD. Columns 3–5 report first-stage regressions of the LOO state stringency instrument on facility mean severity. Controls include log certified beds and ownership type indicators. Column 5 restricts to chain-affiliated facilities and includes chain fixed effects.

Table 2 presents OLS and first-stage results. Panel A shows that the naïve OLS relationship between deficiency severity and total HPRD is weakly negative (–0.048, SE = 0.026) and attenuated with controls (–0.046, SE = 0.027). This small OLS estimate is consistent with two offsetting forces: worse facilities have both lower staffing *and* more citations (negative selection), while citation-induced compliance investment raises staffing (positive causal effect). The IV estimate, by isolating the regulatory pressure channel, separates these forces.

Panel B reports the first stage. The instrument is powerful across all specifications. Unconditionally, a one-unit increase in LOO state stringency raises facility severity by 0.94 ($F = 5{,}398$). Adding controls barely changes the estimate (0.94, $F = 4{,}764$). Even within chains, where the identifying variation comes from cross-state differences in surveyor regimes

facing the same corporate management, the coefficient remains large (0.80, $F = 118$). These F-statistics far exceed conventional weak-instrument thresholds.

5.2 Main IV Results

Table 3: IV Estimates: Effect of Deficiency Severity on Nurse Staffing

	Total HPRD			RN HPRD	
	(1)	(2)	(3)	(4)	(5)
Mean severity (IV)	-0.7371**	-0.7315**	-0.5872***	-0.1960*	-0.0281
	(0.3107)	(0.3238)	(0.1984)	(0.1142)	(0.0981)
Controls	No	Yes	Yes	Yes	Yes
Chain FE	No	No	Yes	No	Yes
Sample	Full	Full	Chain	Full	Chain
Observations	11,427	11,427	8,192	11,427	8,192
First-stage F	5398.2	4763.6	117.7	4763.6	117.7

Notes: 2SLS estimates. The endogenous variable is mean facility deficiency severity (1–12 scale); the instrument is leave-one-out state mean severity in 2025. Standard errors clustered at the state level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Controls: log certified beds, ownership type indicators. Chain FE: chain-entity fixed effects (column 3, 5 restrict to the 8,192 chain-affiliated facilities).

Table 3 presents the core finding. A one-unit increase in instrumented deficiency severity reduces total nurse staffing by 0.74 HPRD (column 1, SE = 0.31, $p = 0.018$). Adding controls changes the estimate negligibly (–0.73, column 2). The within-chain specification (column 3) is the preferred estimate: among 8,192 chain-affiliated facilities, the IV estimate with chain fixed effects is –0.59 HPRD ($p = 0.003$). Since chain fixed effects absorb all corporate-level budget and management policies, this estimate compares facilities within the same chain that face different state surveyors—the closest feasible approximation to random surveyor assignment.

To translate to policy-relevant magnitudes: a one-standard-deviation increase in state stringency (0.29 units) shifts facility severity by $0.29 \times 0.94 = 0.27$ points, reducing total HPRD by $0.27 \times 0.73 = 0.20$ hours per resident per day, or 5.2 percent of the mean (3.86). For a facility with 80 residents, this represents 16 fewer nursing hours per day—approximately two full-time nurse positions.

Columns 4–5 show that RN staffing declines by 0.20 HPRD in the full sample and 0.13 in the chain subsample, though the latter is imprecisely estimated.

5.3 Mechanisms

Table 4: Mechanisms: Staff-Type Decomposition and Turnover

	Staffing (HPRD)				Turnover (%)	
	Total	RN	LPN	CNA	Total	RN
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Staff-type decomposition						
Mean severity (IV)	-0.7315**	-0.1960*	-0.2522	-0.2833	12.70***	6.90**
	(0.3238)	(0.1142)	(0.1876)	(0.2265)	(4.04)	(3.12)
Panel B: Heterogeneity by ownership						
For-profit	-0.7291***					
	(0.2794)					
Non-profit	-0.4429					
	(0.3632)					
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11,427	11,427	11,427	11,427	10,727	10,055

Notes: 2SLS estimates. All columns instrument mean facility deficiency severity with leave-one-out state mean severity. Standard errors clustered at the state level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel A decomposes total HPRD into registered nurses (RN), licensed practical nurses (LPN), and certified nursing assistants (CNA). Turnover is the annual percentage of staff who left. Panel B splits the sample by ownership type for the total HPRD outcome.

Table 4 decomposes the staffing reduction and examines turnover. Panel A reveals that all three staff categories decline: RN (–0.20 HPRD), LPN (–0.25), and CNA (–0.28). The three components sum exactly to the total (–0.73), confirming internal consistency. The largest absolute reduction falls on CNAs, the lowest-cost and most numerous caregiving staff—consistent with facilities reducing the cheapest-to-cut positions first.

The turnover results strengthen the compliance crowding interpretation. Total nursing staff turnover rises by 12.7 percentage points per unit of instrumented severity ($p = 0.002$),

and RN turnover rises by 6.9 points ($p = 0.027$). Strict enforcement does not merely cause facilities to hire fewer nurses; it drives existing nurses away.

Panel B shows that the effect is concentrated in for-profit facilities (–0.73, SE = 0.28), with a smaller and imprecise effect among non-profits (–0.44, SE = 0.36). This heterogeneity is consistent with the compliance crowding mechanism: for-profit facilities, with tighter margin constraints, are more likely to redirect resources from staffing to administrative compliance when faced with regulatory pressure.

5.4 Robustness

Table 5: Robustness Checks

	Coefficient	SE	N	First-stage F
Panel A: Main specification				
Baseline (Table 3, col. 2)	-0.7315**	0.3238	11,427	4763.6
Panel B: Alternative specifications				
Max severity (instead of mean)	-0.4051**	0.1888	11,427	
Chain facilities only	-0.7074***	0.2704	8,192	
Independent facilities only	-0.7216*	0.3702	3,235	
Chain FE (Table 3, col. 3)	-0.5872***	0.1984	8,192	117.7
Panel C: Leave-one-state-out jackknife				
Mean	-0.7297			
Range	[-0.8769, -0.4164]			
Panel D: Weak-IV-robust inference				
Anderson-Rubin F-stat	5.12 ($p = 0.028$)			

Notes: All estimates use 2SLS with state-clustered standard errors unless noted. Panel B varies the endogenous variable or sample while maintaining the LOO state stringency instrument. Panel C reports summary statistics from 52 regressions, each dropping one state. Panel D reports the Anderson-Rubin F-statistic, which is valid regardless of instrument strength.

Table 5 reports robustness checks. The main finding is stable across specifications (Panel B): using maximum severity rather than mean (–0.41), restricting to chain or independent facilities separately (–0.71 and –0.72), and including chain fixed effects (–0.59). The leave-one-state-out jackknife (Panel C) shows remarkable stability: the mean across 52 regressions is –0.73, with a range of [–0.88, –0.42]. No single state drives the result.

Balance tests. The instrument is uncorrelated with log bed count (0.025, SE = 0.14) and for-profit status (–0.009, SE = 0.10). The urban indicator shows some correlation (–0.30, SE = 0.12), a limitation I acknowledge. However, controlling for urban status does not meaningfully change the IV estimates, and the within-chain specification absorbs state-level urban-rural composition.

Alternative treatment measures. Using maximum severity rather than mean severity as the endogenous variable yields an IV coefficient of –0.41 (SE = 0.19), smaller in magnitude but the same sign. Using the *count* of deficiencies produces a positive coefficient (+0.14, SE = 0.06), consistent with the interpretation that it is the *severity* of regulatory action—not merely its volume—that drives the compliance crowding response.

Weak-IV inference. The Anderson-Rubin F-statistic is 5.12 ($p = 0.028$), confirming that the reduced-form relationship is significant under weak-IV-robust inference. Given the extremely large first-stage F (over 4,700), weak instruments are not a concern in this setting, but the AR test provides a model-free confirmation.

6. Discussion

The central finding—that strict regulatory enforcement *reduces* nursing home staffing—challenges the conventional model of health care regulation as a quality-promoting force. The compliance crowding mechanism offers a unified explanation: deficiency citations impose real administrative costs (plans of correction, resurvey preparation, documentation) that consume the same nursing hours they are intended to protect. The turnover result adds a dynamic dimension: the regulatory burden is not merely a static reallocation but an ongoing friction that erodes the nursing workforce at cited facilities.

Three caveats temper these conclusions. First, the exclusion restriction may be violated if state-level characteristics—particularly Medicaid reimbursement rates, state-specific staffing mandates, and nurse labor market conditions—independently affect both surveyor stringency and facility staffing (Hackmann, 2019). The within-chain estimate controls for corporate management but not for the fact that a chain facility in Utah faces Utah's Medicaid rates

while one in Texas faces Texas's. I note, however, that the balance tests show no correlation between the instrument and observable facility characteristics (beds, ownership), and the jackknife confirms no single state drives the result. Second, the 2SLS coefficient identifies a local average causal response for facilities whose measured severity is shifted by state stringency. These "compliers" may have systematically lower baseline staffing than the full population, limiting external validity for uniform policy reforms. Third, the cross-sectional design cannot establish temporal ordering: if annual staffing is measured concurrently with the annual survey, reverse causality (low staffing causing citations) remains a concern despite the IV. Longitudinal evidence on staffing changes *after* specific survey events would strengthen the causal interpretation.

These results speak to a broader principle in regulatory design: the cost of compliance is itself a binding constraint, and optimal enforcement must account for the resources that citations consume. In nursing homes, where care quality is measured in nursing minutes per resident, the compliance burden falls directly on the margin it is supposed to improve.

7. Conclusion

State survey agency stringency creates a regulatory lottery for nursing homes. Facilities that draw strict surveyors receive more and worse deficiency citations, but rather than responding with quality investment, they shed nursing staff and experience higher turnover. The mechanism is compliance crowding: each citation generates administrative demands that crowd out clinical care. For the 1.3 million Americans living in nursing homes, the inspection lottery is not just unfair—it is counterproductive.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers

Contributors: @olafdrw

First Contributor: https://github.com/olafdrw

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Bardach, Eugene and Robert A. Kagan, "Going by the Book: The Problem of Regulatory Unreasonableness," *Transaction Publishers*, 2002. Revised edition.

Bowblis, John R., "Ownership Conversion and Closure in the Nursing Home Industry," *Health Economics*, 2011, *20* (6), 631–644.

Castle, Nicholas G., Lisa M. Wagner, Jamie C. Ferguson, and Steven M. Handler, "Nursing Home Deficiency Citations for Safety," *Journal of Aging and Social Policy*, 2011, *23* (1), 34–57.

Centers for Medicare and Medicaid Services, "State Operations Manual: Chapter 7—Survey and Enforcement Process for Skilled Nursing Facilities and Nursing Facilities," Technical Report, CMS 2017.

_ , "Design for Nursing Home Compare Five-Star Quality Rating System: Technical Users' Guide," Technical Report, CMS 2019.

Doyle, Joseph J., "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *American Economic Review*, 2007, *97* (5), 1583–1610.

Government Accountability Office, "Nursing Homes: Federal Monitoring Surveys Demonstrate Continued Understatement of Serious Care Problems and CMS Oversight Weaknesses," Report GAO-08-517, U.S. Government Accountability Office 2008.

Grabowski, David C., "A Longitudinal Study of Medicaid Payment, Private-Pay Price and Nursing Home Quality," *International Journal of Health Care Finance and Economics*, 2004, *4* (1), 5–26.

Hackmann, Martin B., "Incentivizing Better Quality of Care: The Role of Medicaid and Competition in the Nursing Home Industry," *American Economic Review*, 2019, *109* (5), 1684–1716.

Harrington, Charlene, Barbara Olney, Helen Carrillo, and Taewoon Kang, "Nurse Staffing and Deficiencies in the Largest For-Profit Nursing Home Chains and Chains Owned by Private Equity Companies," *Health Services Research*, 2012, *47* (1pt1), 106–128.

_ , David Zimmerman, Sarita L. Karon, James Robinson, and Patricia Beutel, "Nursing Home Staffing and Its Relationship to Deficiencies," *Journals of Gerontology: Social Sciences*, 2004, *59B* (5), S278–S287.

Johnson, Matthew S., "Regulation by Shaming: Deterrence Effects of Publicizing Violations of Workplace Safety and Health Laws," *American Economic Review*, 2020, *110* (6), 1866–1904.

Kling, Jeffrey R., "Incarceration Length, Employment, and Earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Konetzka, R. Tamara, Sally C. Stearns, and Jeongyoung Park, "The Staffing–Outcomes Relationship in Nursing Homes," *Health Services Research*, 2008, *43* (3), 1025–1042.

Mukamel, Dana B., David L. Weimer, Charlene Harrington, William D. Spector, Helen Ladd, and Yue Li, "The Effect of State Regulatory Stringency on Nursing Home Quality," *Health Services Research*, 2012, *47* (5), 1791–1813.

Parker, Christine and Vibeke Lehmann Nielsen, "The Challenge of Empirical Research on Business Compliance in Regulatory Capitalism," *Annual Review of Law and Social Science*, 2009, *5*, 45–70.

Short, Jodi L. and Michael W. Toffel, "Making Self-Regulation More Than Merely Symbolic: The Critical Role of the Legal Environment," *Administrative Science Quarterly*, 2010, *55* (3), 361–396.

A. Data Appendix

Data Sources and Access. All data come from the CMS Provider Data Catalog (https://data.cms.gov/provider-data/), accessed March 30, 2026. No authentication is required. Four datasets were used:

1. **Health Deficiencies** (dataset ID: `r5ix-sfxw`): 418,972 individual deficiency citations from standard health inspections and complaint surveys, covering the three most recent inspection cycles per facility (survey dates from March 2017 to February 2026). Each record contains the facility CCN, survey date, deficiency tag number, scope-severity code (A–L), and whether the deficiency was corrected.

2. **Provider Information** (dataset ID: `4pq5-n9py`): 14,703 Medicare/Medicaid-certified nursing facilities. Key fields include certified bed count, average daily resident count, ownership type, chain name and ID, urban/rural classification, reported staffing HPRD by staff type (from Payroll-Based Journal), case-mix-adjusted staffing, nursing staff turnover rates, and CMS Five-Star ratings.

3. **Quality Measures — MDS** (dataset ID: `djen-97ju`): 249,951 quality measure records derived from the Minimum Data Set, covering 14 long-stay measures (e.g., falls, pressure ulcers, UTIs, physical restraints) reported as four-quarter average scores.

4. **Penalties**: Penalty counts and amounts were obtained from the Provider Information dataset (fields: Number of Fines, Total Amount of Fines in Dollars, Total Number of Penalties).

Variable Construction. The scope-severity code maps to a numeric 1–12 scale following the CMS classification matrix. Let d_{ij} denote the severity score for deficiency citation j at facility i in the most recent survey year. The facility-level mean severity is $\bar{S}_i = \frac{1}{J_i}\sum_j d_{ij}$, where J_i is the number of citations. The leave-one-out instrument is $Z_{s(i)} = \frac{\sum_{k \in s, k \neq i} \bar{S}_k}{N_s - 1}$, where N_s is the number of facilities in state s.

Sample Construction. The main restriction is the requirement of a completed 2025 survey (drops 2,874 facilities without recent inspections, primarily small or newly certified homes). A further 400 facilities are missing PBJ staffing data, often the smallest or worst-performing facilities that fail to submit complete payroll records. While this selection is non-random, it biases *against* finding a negative staffing effect: if the lowest-staffing facilities are excluded, the estimated compliance crowding effect is a lower bound.

Step	Facilities
All certified nursing homes	14,703
With completed 2025 survey	11,829
Non-missing staffing data	11,429
Non-missing instrument and beds	11,427
Chain subsample	8,192

B. Identification Appendix

Instrument Balance. The LOO state stringency instrument is uncorrelated with log bed count ($\hat{\beta} = 0.025$, SE = 0.143) and for-profit ownership ($\hat{\beta} = -0.009$, SE = 0.096). The urban indicator shows some correlation ($\hat{\beta} = -0.297$, SE = 0.115), which is a limitation of the cross-sectional design.

Leave-One-State-Out Stability. Dropping each state in turn produces 52 IV point estimates ranging from –0.88 to –0.42 (mean –0.73, SD 0.06), confirming that no individual state drives the main result.

C. Robustness Appendix

See Table 5 in the main text for all robustness specifications. The Anderson-Rubin F-statistic (5.12, $p = 0.028$) confirms significance under weak-IV-robust inference, though the first-stage F of 4,764 makes weak instruments a non-issue in this setting.

D. Standardized Effect Sizes

Table 6: Standardized Effect Sizes for Main Outcomes

Outcome	Specification	$\hat{\beta}$	SD(X)	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled							
Total HPRD	IV (controls)	-0.731	0.882	0.884	-0.730	0.323	Large negative
RN HPRD	IV (controls)	-0.196	0.882	0.439	-0.394	0.230	Large negative
Total turnover	IV (controls)	12.704	0.882	14.555	0.770	0.245	Large positive
Panel B: Heterogeneous (by ownership type)							
Total HPRD (for-profit)	IV	-0.729	0.862	0.722	-0.871	0.334	Large negative
Total HPRD (non-profit)	IV	-0.443	0.944	1.036	-0.404	0.331	Large negative

Notes: **Country:** United States. **Research question:** Does stricter state regulatory inspection of nursing homes cause facilities to invest more in clinical staffing, or does it crowd out care through compliance costs? **Policy mechanism:** CMS delegates nursing home certification surveys to state agencies, which assign deficiency citations on a scope-severity grid (A–L). Stricter state agencies issue more and worse citations for identical underlying conditions, triggering mandatory plans of correction, resurveys, and potential financial penalties that impose compliance costs on facilities. **Outcome definition:** Total nurse staffing hours per resident per day (HPRD) from CMS Payroll-Based Journal submissions, covering RNs, LPNs, and CNAs; nursing staff turnover is the annual percentage of nursing staff who left the facility. **Treatment:** Continuous — facility mean deficiency severity score (1–12), instrumented by leave-one-out state mean severity. **Data:** CMS Provider Data Catalog, 2025 survey year, facility-level cross-section, N = 11,427 Medicare/Medicaid-certified nursing homes across 52 states/territories. **Method:** 2SLS with leave-one-out state stringency as instrument; standard errors clustered at the state level. **Sample:** All Medicare/Medicaid-certified nursing homes with completed standard health inspections in 2025 and non-missing staffing data. SDE = $\hat{\beta} \times \text{SD}(X)/\text{SD}(Y)$ where SD(X) is the standard deviation of mean severity and SD(Y) is the pre-treatment standard deviation of the outcome. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).