The Labor Market as an Equilibrium Newsvendor Problem

Alexandre Mas*

UC Berkeley, Haas School of Business

March 28, 2026

Abstract

This paper studies staffing under uncertainty when firms can either hold permanent labor as a buffer or hire from a spot market after demand is realized. I build a model where the spot market is an equilibrium object such that when one firm relies more on spot labor, it thickens the market that other firms use as well. Individual firms do not internalize that contribution, so competitive equilibrium involves too much private buffer capacity. I evaluate the model in the market for registered nurses at skilled nursing facilities, using daily payroll records for roughly 15,000 facilities from 2017 to 2023. The data support the model's general equilibrium predictions. Counties with thicker pre-period agency markets have less rationing conditional on total nursing hours, and thicker agency markets compress the upper tail of permanent staffing without reducing total coverage. Finally, I estimate a facility-level newsvendor model with spot market frictions and rationing costs and find that spot market frictions fell after 2020. A revealed-preference exercise shows that the value of an additional nurse in the spot market typically exceeds the cost of holding that nurse as a regular employee implying that this labor market underprovides flexibility even in a competitive setting because the value of availability is not fully priced.

*I am grateful to Pat Kline, Enrico Moretti, and Carl Shapiro for helpful discussions.

1 Introduction

How do firms staff when demand is uncertain and labor cannot be adjusted costlessly in real time? A firm can hire enough permanent workers to cover peak demand, accepting idle capacity in normal periods, or it can run lean and fill gaps through the spot market. Adjusting for workload fluctuations is the primary reason employers use flexible staffing arrangements (Houseman, 2001). This problem is similar to the one in the canonical newsvendor model for inventory whereby a firm pre-commits to a stock before demand is realized, trading off the cost of excess inventory against the cost of running short.[1] Typically, these optimization problems are studied for a single firm. But in an economy with many firms solving this problem, there will be equilibrium implications. In labor markets, for example, the underage option is hiring from a spot market, and the spot market wage and the efficiency of this market itself depends on how many firms and workers participate.

This paper studies the decision to employ a stable workforce versus hiring from a spot market given demand uncertainty. The spot market has features of a shared public good: a firm that relies more on spot hiring thickens the market others use but does not internalize that contribution, so competitive equilibrium involves too much private buffer capacity and too little spot-market participation. The strategic complementarity underlying this result can support multiple equilibria, a thin market in which firms buffer because spot hiring is unreliable and a thick market in which spot hiring is attractive because enough others participate. Demand shocks can improve market functioning indirectly by shifting firms toward the spot market.

The model has three basic elements: heterogeneous firm productivity, aggregate demand uncertainty, and different employment modes. Prior to observing demand

[1]The newsvendor model was developed for inventory management by Arrow, Harris, and Marschak (1951) and formalized by Scarf (1958).

states, firms decide whether to hire permanent workers or wait until states are observed and hire from the spot market. The frictional cost firms incur in the spot market is endogenous to the number of firms and workers participating in it. Through these primitives, the model yields predictions for wage levels, wage dispersion, job separation rates, unemployment, productivity, output, and impacts of both business cycle fluctuations and changes to expectations.

I study these forces in the market for registered nurses at skilled nursing facilities. CMS Payroll-Based Journal data report daily hours separately for employee and agency nurses, so the buffer and spot margins are both directly observed. There is also a well-defined spot market with considerable nurse turnover and cross-facility movement of agency and per-diem staff, making nursing well suited for asking whether the buffer-spot margin of adjustment and market thickening change the price and allocation of flexible labor.

The nursing labor market from 2020 onward is especially interesting for this purpose. Health care workers experienced pronounced wage growth over this period (Appendix Figure A1), accompanied by high quits and separation rates (Appendix Figure A2). Among nurses, wage growth was almost entirely within facilities rather than compositional, and separation increases were symmetric across the facility wage distribution and largely a market-level rather than employer-level phenomenon. These facts are reconciled by the model as wages rise through the equilibrium value of flexible labor and measured churn rises because firms increasingly hire on-demand rather than maintaining a buffer.

Using facility by worker level data on work hours for regular and contract employees from Centers for Medicare and Medicaid Services (CMS), I show evidence supporting the general equilibrium predicts of the model hold in the data. Counties with thicker pre-period agency markets have lower rationing conditional on total

nursing hours and a compressed upper tail of permanent staffing without reduced total coverage.

Finally, I estimate the structural parameters of the newsvendor model facility-by-facility using the daily payroll records. Each facility's staffing problem implies two first-order conditions: an ex-ante buffer choice and an ex-post agency hiring decision on days when the permanent buffer falls short. These two conditions jointly identify the agency frictional markup, the marginal rationing cost, and the facility's tolerated shortfall without imposing functional form assumptions on the demand distribution. I find that agency frictions fell after 2020, rationing costs rose, and the tolerated residual shortfall narrowed.

A revealed-preference exercise then uses the structural estimates to evaluate three welfare channels. First, the marginal buffer nurse creates more expected value in the county spot market than in private employment in 73 percent of county-years, providing direct evidence of a social-private divergence at the buffer margin. Second, there is a strong negative relationship between county agency share and estimated spot market frictions, and the implied market-thickening participation externality averages \$1,776 per nurse engagement. Third, a facility that switches from buffer employment to spot market participation captures only 17 percent of the social value it creates. The implied Pigouvian subsidy rate peaks at 5.9 percent of the agency engagement wage in 2022. Section 7 reports the results.

The paper connects to several literatures. It offers a theory of staffing, wages, and churn when flexible labor is procured from a shared market rather than through bilateral search alone, and the shared-market feature is relevant for macroeconomics because market thickness can amplify aggregate shocks by changing the effective cost of labor adjustment. The mechanism is closest in spirit to thick-market search externalities (Diamond, 1982; Mortensen, 1982) and strategic complementarity (Howitt

and McAfee, 1992), operating through the margin between private capacity and a shared spot market. The density channel also connects to the agglomeration literature: Helsley and Strange (1990) show that city size improves expected match quality by increasing pool density in characteristic space, and Moretti and Yi (2024) provide empirical support for matching advantages in larger markets. Here market size is endogenous to firms' buffer-spot choices rather than determined by geography. The paper speaks to labor hoarding (Oi, 1962; Fay and Medoff, 1985; Hall, 1991), state-independent wages (Baily, 1974; Azariadis, 1975), and use of contract and outsourced labor in place of direct employment (Goldschmidt and Schmieder, 2017), providing an equilibrium foundation for where firms locate on this margin when demand is uncertain. Houseman et al. (2003) document employer incentives to use agency labor in tight markets, emphasizing wage discrimination enabled by workers' switching costs. Here wages and employment arise from demand uncertainty and equilibrium spot market frictions rather than employer market power.

Section 2 develops the model. Section 3 derives the theoretical results. Section 4 presents the nursing labor market facts and shows how they fit the model's predictions. Section 5 presents reduced-form evidence on market thickness and coverage. Section 6 presents the structural estimation. Section 7 quantifies over-buffering. Section 8 concludes.

2 Model

2.1 Basic Elements

The model has four ingredients: (i) a choice between holding stable buffer staff versus hiring as-needed from a spot market; (ii) heterogeneity in position productivity;

(iii) demand uncertainty with discrete high and low states; and (iv) thick-market externalities in the shared spot market.

2.2 Environment

Positions and firms. The economy has a continuum of productive positions with mass 1. Each position has productivity θ drawn from a distribution F on $[0, 1]$ with continuous, single-peaked density f: there exists $\mu \in (0, 1)$ such that $f' > 0$ for $\theta < \mu$ and $f' < 0$ for $\theta > \mu$. A staffed position produces θ when active.[2]

Demand states. State $s \in \{H, L\}$ realizes each period with probability π_s ($\pi_H + \pi_L = 1$). In state s, each position independently requires capacity with probability δ_s, where $\delta_H > \delta_L$. The unconditional activation rate is $\bar{\delta} = \pi_H \delta_H + \pi_L \delta_L$. The interpretation of a state can depend on the context. If can be a business cycle fluctuation, seasonal demand, or high-frequency idiosyncratic shocks.

Workers. Total mass $N \in (0, 1)$ of risk-neutral workers with outside option $b \geq 0$, representing the value of the next-best activity foregone when engaged by a facility.

2.3 Staffing Modes and Timing

Before the state realizes, each position is staffed via one of two modes:

Buffer (permanent/stable). A worker is assigned at wage w, paid regardless of whether the position is active. The worker is committed and foregoes b.

Spot (as-needed). The position is filled after the state realizes, only if active.

[2]Firms are collections of positions with some firms having higher productivity positions. Since positions are staffed independently, the aggregate distribution F governs all equilibrium objects.

Timing: prior to observing the state, firms choose their buffer-spot allocation; workers simultaneously choose between buffer contracts and entering the spot market. After the state realizes, spot hirings occur. Buffer workers are retained and cannot renegotiate.[3]

2.4 Spot Market Friction

A spot position served in state s produces $\gamma_s \cdot \theta$ for $\gamma_s \leq 1$.[4] The term γ_s represents the cost of spot market frictions and depends on two objects:

- *Market density*: total pool mass L. A denser market reduces the cost of any given match quality; more workers in the characteristic space means more workers close to any given firm.
- *Congestion*: slack ratio $R_s = L/E_s$ where $E_s = \delta_s F(\theta^*)$ is the mass of active spot positions. When many positions hire simultaneously, close workers are bid away, worsening match quality.[5]

Let $g(L)$ be match quality a firm achieves as sole employer drawing from L workers; let $h(R_s)$ be the fraction of that potential realized when E_s positions draw simultaneously. A spot position served in state s produces $\gamma_s \cdot \theta$:

$$\gamma_s \equiv g(L) \cdot h(R_s), \tag{1}$$

[3]There is considerable evidence that workers have preferences for stability of hours, see eg. Mas and Pallais (2017) and Lachowska et al. (2023). It is straightforward to add a cost workers experience when in the spot market versus a buffer position and it does not change the main elements or predictions of the model. I do not add this cost in order to isolate the market friction element of the buffer/spot trade-off.

[4]Workers and firms can be understood as existing as points in a high-dimensional characteristic space capturing location, shift preference, credentials, and network proximity. A spot-hiring firm optimizes along a cost-of-quality schedule, choosing the match quality at which the marginal output gain equals the marginal acquisition cost. The optimal quality is below 1 because close matches are costly.

[5]Bergeaud et al. (2024) document that the choice between permanent and agency employment involves efficiency considerations beyond simple cost comparisons.

where $g' > 0$, $g'' \leq 0$, $g(0) = g_0 > 0$, $\lim_{L\to\infty} g(L) = 1$; $h' > 0$, $h'' \leq 0$, $\lim_{R\to\infty} h(R) = 1$. The density term $g(L)$ captures the matching externality in Helsley and Strange (1990): in a denser pool, workers are closer on average in the characteristic space and match quality rises. The congestion term $h(R_s)$ is analogous to the tightness effect in Diamond (1982) and Mortensen (1982): when many positions hire simultaneously, close workers are bid away and realized match quality falls.

Since $\delta_H > \delta_L$ implies $R_H < R_L$, we have $\gamma_H < \gamma_L$: the spot market is more congested in the high-demand state.

2.5 Spot Market Clearing

There is a cutoff equilibrium in which positions with $\theta \geq \theta^*$ are buffered and positions with $\theta < \theta^*$ use spot.[6] Buffered positions have mass $1 - F(\theta^*)$, each assigned one permanent worker. The remaining workers enter the spot pool:

$$L(\theta^*) = N - 1 + F(\theta^*), \qquad \frac{dL}{d\theta^*} = f(\theta^*). \tag{2}$$

The derivative $f(\theta^*)$ governs how responsive the pool is to threshold changes: near the mode, many positions cluster near the margin and a small shift in θ^* moves many workers; in the tails, the pool responds sluggishly. The slack ratio responds to the threshold value as:

$$\frac{dR_s}{d\theta^*} = \frac{f(\theta^*)(1-N)}{\delta_s F(\theta^*)^2} > 0, \tag{3}$$

positive because $N < 1$: each position switching from buffer to spot frees one dedicated worker but adds only $\delta_s < 1$ units of expected demand, so the market becomes less congested.

[6] This ordering follows because the output gain from avoiding friction, $(1 - \gamma_s)\theta$, is increasing in θ, while the idle-cost premium of buffering is independent of θ.

H tight, L slack Regime. The model has three possible equilibrium regimes: In Regime 1 the market is tight in both states, in Regime 2 the market is tight in H states and slack in L states, and in Regime 3 the market is slack in both states. I focus on Regime 2 where state H is tight ($R_H < 1$) and state L is slack ($R_L \geq 1$).[7] In the slack state, active spot positions hire at the outside option: $w_L = b$.

In the tight state, $\delta_H F(\theta^*)$ positions compete for $L < \delta_H F(\theta^*)$ workers, who are allocated to the highest-productivity active positions. The marginal hired position θ_m satisfies:

$$\delta_H \big[F(\theta^*) - F(\theta_m)\big] = L(\theta^*). \tag{4}$$

Active positions with $\theta \in [\theta_m, \theta^*)$ hire; positions with $\theta < \theta_m$ are rationed.

Scarcity wage. The spot wage in state H equals the marginal position's willingness to pay:

$$w_H = \gamma_H \cdot \theta_m(\theta^*) = g(L)\, h(R_H) \cdot \theta_m(\theta^*). \tag{5}$$

Substituting $L(\theta^*)$ and solving (4) for θ_m:

$$w_H(\theta^*) = g(N - 1 + F(\theta^*)) \cdot h\left(\frac{N - 1 + F(\theta^*)}{\delta_H F(\theta^*)}\right) \cdot F^{-1}\left(\frac{(\delta_H - 1)F(\theta^*) + 1 - N}{\delta_H}\right). \tag{6}$$

The scarcity wage depends on θ^* through three channels: density, congestion, and rationing (θ_m falls as fewer positions are shut out). As the market thickens, the first two channels raise w_H while the third lowers it. Figure A9 illustrates each channel separately.

[7] State H is tight iff $F(\theta^*) < (1 - N)/(1 - \delta_H)$; state L is slack iff $F(\theta^*) \geq (1 - N)/(1 - \delta_L)$. Since $\delta_H > \delta_L$, this is a nonempty interval. Appendix C analyzes Regimes 1 (both slack) and 3 (both tight).

Non-employment. Expected non-employment is:

$$u(\theta^*) = \pi_L\big[N - 1 + (1 - \delta_L)F(\theta^*)\big]. \tag{7}$$

Higher θ^* means more workers are in the spot pool and more are unhired in slack markets. But those workers earn the outside option b rather than sitting idle on permanent contracts, so higher measured non-employment reflects less commitment waste.

Output. Expected aggregate output is:

$$Y(\theta^*) = \bar{\delta}\int_{\theta^*}^{1}\theta\, dF + \pi_H\delta_H\gamma_H\int_{\theta_m(\theta^*)}^{\theta^*}\theta\, dF + \pi_L\delta_L\gamma_L\int_{0}^{\theta^*}\theta\, dF. \tag{8}$$

Output varies at the rationing margin. When there are more workers in buffer contracts, more positions are rationed. Workers in buffer contracts may not produce if their position does not have demand after the state is realized.

2.6 Sorting and Equilibrium

Buffer wage. Worker indifference requires:

$$w = \pi_H w_H + \pi_L b. \tag{9}$$

The buffer wage equals expected spot earnings.

Equilibrium condition. Setting the expected profit of a buffer position equal to that of the marginal spot position:

$$\underbrace{\pi_H \delta_H (1-\gamma_H)\theta^* + \pi_L \delta_L (1-\gamma_L)\theta^*}_{\text{friction cost of spot}} = \underbrace{\pi_H (1-\delta_H) w_H + \pi_L (1-\delta_L) b}_{\text{idle cost of buffering}}. \tag{10}$$

Equilibrium threshold. Solving (10) for the cutoff:

$$\theta^{BR}(\theta^*) = \frac{\pi_H (1-\delta_H)\, w_H(\theta^*) + \pi_L (1-\delta_L)\, b}{\pi_H \delta_H \big(1-\gamma_H(\theta^*)\big) + \pi_L \delta_L \big(1-\gamma_L(\theta^*)\big)}. \tag{11}$$

A position with $\theta > \theta^{BR}$ buffers; with $\theta < \theta^{BR}$ it uses spot.

A competitive equilibrium is a tuple (θ^*, w, w_H, L) satisfying the pool equation (2), the tight-state clearing condition (4), the buffer wage condition (9), and the indifference condition below. The equilibrium threshold is the fixed point $\theta^* = \theta^{BR}(\theta^*)$. The equilibrium threshold value determines the remaining endogenous variables in the model.

Figure 3 illustrates the equilibrium for a fixed set of parameter values.[8] The equilibrium is determined by the location of the downward sloping portion of the best-response curve crossing the 45 degree line.

This model has implications for a number of market characteristics. Permanent workers receive w regardless of the realized state, generating wage rigidity. Both within-state and cross-mode wage dispersion arise endogenously. When more positions use spot, more workers rotate across facilities and separation rates rise. Unhired spot workers in slack markets appear as non-employment, and output per worker is higher in thick markets with fewer idle workers. Treating states as booms and busts gives a business cycle interpretation, and changes in the anticipated frequency or

[8] Existence is established in Appendix A.

severity of high-demand states shift the equilibrium cutoff before any state realizes, allowing forward-looking responses to news.

3 Theoretical Results

3.1 Excess Buffering

When a position is filled with a stable worker as buffer, it withdraws one worker from the spot pool but adds only $\bar{\delta} < 1$ units of expected demand. The market thins, reducing pool density and worsening congestion, and match quality falls for every remaining spot user. The buffering firm does not internalize these costs.

Since workers are risk-neutral and total labor N is exogenous, welfare equals expected output plus the outside-option value of non-employed workers:

$$W(\theta^*) = Y(\theta^*) + u(\theta^*) \cdot b. \tag{12}$$

At any equilibrium θ^*, $W'(\theta^*) > 0$ the planner prefers strictly less buffering and the competitive equilibrium has too much buffering relative to the social optimum $\theta^{**} > \theta^*$ (See Appendix E for proof).[9]

Competitive markets also underprovide spot-market participation. A worker who enters the spot market earns $\pi_H w_H + \pi_L b$ privately, yet her social contribution exceeds this because her presence thickens the pool and lowers congestion for every other hirer:

$$A^* = \pi_H \delta_H \frac{\partial \gamma_H}{\partial N} Q_H + \pi_L \delta_L \frac{\partial \gamma_L}{\partial N} Q_L > 0, \tag{13}$$

[9]The result follows because welfare is evaluated at the point where the private trade-off is exactly balanced: the equilibrium condition $\mathcal{D} \cdot \theta^* = \mathcal{N}$ means the marginal buffering position breaks even, so the private cost-benefit terms in W' cancel to zero and only the uninternalized market-thickening externality remains. It is because the private margin contributes zero net surplus at equilibrium that any positive externality is sufficient to sign the welfare derivative.

where Q_s is the aggregate productivity of served spot positions in state s. The derivation that $dW/dN = \pi_H w_H + \pi_L b + A^*$ and hence that the social return exceeds the private return by $A^* > 0$ is given in Appendix B. A^* captures the improvement in match quality that the entering worker delivers to everyone else through denser market conditions (g') and reduced congestion (h'), in both demand states and regardless of whether they are active or idle. It is largest when the market is thin, where unemployment is low but the social value of an additional worker in the spot market is highest.[10]

3.2 Demand Expectations Shocks and Equilibrium Outcomes

Regime 2, where the high-demand state is tight and the low-demand state is slack, is the empirically relevant case for the nursing and many other applications and supports a unique equilibrium for a wide range of parameters. The question is what happens to θ^* when a shock to demand expectations raises both the intensity (δ_H) and the frequency (π_H) of the high-demand state.[11]

Demand intensity expectations shock (higher δ_H). At fixed θ^*, higher δ_H worsens congestion in the spot market ($\partial R_H/\partial \delta_H < 0$) and raises the rationing threshold θ_m. These two channels push the scarcity wage in opposite directions: $\frac{\partial w_H}{\partial \delta_H} = \frac{\gamma_H}{\delta_H}\left[\frac{F(\theta^*)-F(\theta_m)}{f(\theta_m)} - \varepsilon_h(R_H)\,\theta_m\right]$, where $\varepsilon_h = h'R_H/h$ is the congestion elasticity. In Regime 2, rationed positions are low-productivity and θ_m lies in the left tail of F, where $f(\theta_m)$ is small. The rationing-relief channel therefore dominates: relieving the marginal rationed position is inframarginal to the congestion cost, so $\partial w_H/\partial \delta_H > 0$

[10] Standby fees, capacity reservation charges, and retainer contracts are partial market corrections for A^*, but none is efficient because employers who benefit from the thicker market are not charged the fee.

[11] Appendix Figure A11 confirms that this branch remains stable under the joint demand shifts considered.

and $\partial\theta^{BR}/\partial\delta_H > 0$. Higher demand intensity raises the value of flexibility and shifts the equilibrium toward spot.

Demand frequency expectations shock (higher π_H). A higher probability of the tight state raises the expected scarcity rent from spot participation and raises the idle cost of permanent buffering simultaneously. The net effect on is $\partial\theta^{BR}/\partial\pi_H > 0$ iff $\frac{w_H}{b} > \frac{\delta_H(1-\delta_L)(1-\gamma_H)}{\delta_L(1-\delta_H)(1-\gamma_L)}$ (See Appendix H for proof). This condition is satisfied when the scarcity premium is large relative to the outside option: when the tight state is worth substantially more to a spot worker than her slack-period alternative. When this holds, spot participation is sufficiently valuable that a higher probability of the tight state tips the equilibrium toward the spot market. In markets where spot wages substantially exceed permanent wages—a common feature of tight spot markets—the condition is met. When the condition holds, θ^* rises. Three outcomes follow directly from the equilibrium shift and require no further conditions: spot share in state H is $(N - 1 + F(\theta^*))/N$, strictly increasing in θ^*, separation rates rise as more workers cycling through the spot pool, in state H idle share $(1 - \delta_H)(1 - F(\theta^*))$, strictly decreasing in θ^*.

The scarcity wage w_H is non-monotone in θ^*: it rises as density and congestion-relief improvements dominate in thin markets, but eventually falls as the rationing channel weakens near the Regime 2/1 boundary (Figure A9). Appendix G shows that there exists $\underline{\theta}$ below the mode of F such that if $\theta^* < \underline{\theta}$, then a demand shock satisfying $\frac{w_H}{b} > \frac{\delta_H(1-\delta_L)(1-\gamma_H)}{\delta_L(1-\delta_H)(1-\gamma_L)}$ also delivers: (iv) wages rise in both modes; and (v) wage dispersion widens.[12]

[12]The sufficient condition is that the equilibrium is sufficiently far from the Regime 2/1 boundary. In a thin market, w_H is on the upward-sloping portion of the scarcity-wage curve, so the density and congestion-relief gains dominate as θ^* rises. For dispersion, the same thinness ensures $w_H - b$ is small, so the general-equilibrium wage response easily exceeds the direct compression from higher π_H at fixed θ^*. In the limit as the market approaches maximum thinness, $w_H \to b$ and the dispersion condition becomes trivial.

Figure 4 displays all five predictions, with the underlying best-response shifts shown in Appendix Figure A11, as (δ_H, π_H) increase jointly from $(0.55, 0.45)$ to $(0.76, 0.66)$, starting from a thin-market equilibrium.[13]

Multiple equilibria. The comparative statics above assume a unique equilibrium and follow without a coordination mechanism. However, the strategic complementarity in the model can also generate multiple equilibria. When a position switches from buffer to spot, the market thickens, both $g(L)$ and $h(R_s)$ improve, and spot becomes more attractive for every remaining position. Figure A10 illustrates the three-equilibrium structure for specific parameter values. The thin-market equilibrium θ_1^* lies in the rationing regime, where match quality γ_H is low, buffering is individually rational, and the equilibrium is stable because $f(\theta_1^*)$ is small in the left tail. The unstable tipping threshold θ_2^* sits near the mode of f, where complementarity is strongest and small perturbations amplify. [14] The thick-market equilibrium welfare-dominates the thin-market equilibrium: all components of spot output improve simultaneously as θ^* rises from θ_1^* to θ_3^*, and the freed workers earn the outside option in slack periods, partially offsetting the loss of frictionless buffer output.

[13]The COVID-19 shock likely raised the outside option b because heightened competition for nurses from hospitals increased the value of nurses' next-best alternative. Auerbach et al. (2024) document that total RN employment grew roughly 6 percent above pre-pandemic levels by 2022–2023, suggesting the rise in outside options reflected surging demand rather than a contraction in supply. An increase in b reinforces all five predictions: higher idle costs make buffering less attractive, shifting the equilibrium toward spot, while the direct effect on the buffer wage and the induced thickening of the market raise w_H; for wage dispersion, whether the spread widens depends on whether the market-thickening increase in w_H dominates the direct compression from a higher floor.

[14]The thick-market equilibrium θ_3^* is pinned at the regime boundary where rationing just vanishes ($R_H \approx 1$, $\theta_m \approx 0$): as $\theta_m \to 0$, the scarcity wage $w_H = \gamma_H \theta_m \to 0$, pulling θ^{BR} back below the 45-degree line and providing the stabilizing force. Near the boundary the only force sustaining spot participation is the value workers earn during idle slack periods, $\pi_L(1-\delta_L)b$; the reason no stable equilibrium can exist deeper inside the slack regime is established in Appendix F.

4 The Nursing Labor Market

Skilled nursing facilities face the basic staffing problem in this model. They must decide how many nurses to keep on permanent payroll before daily demand and attendance are known, and then decide whether to fill the remaining gaps with labor from the spot market.[15]

4.1 Data

Payroll-Based Journal (PBJ). CMS requires all certified SNFs to report daily staffing hours for every employee and every agency contractor. I use these records to construct, for each facility-day, hours worked by RNs in each staffing mode. The PBJ spans 2017–2023, covering approximately 15,000 facilities. The daily frequency provides simultaneous observation of the permanent buffer H and spot hiring A_t—the two main objects of the structural estimation.

The outcome of interest in this model is whether a facility is meeting needs through the RN staffing ratio, measured as hours of RN care per patient day (HPPD). HPPD varies for two reasons: patient counts fluctuate daily with admissions, discharges, and acuity, and employee call-outs and scheduling gaps generate supply-side shocks to effective permanent hours. Both are observed in the PBJ, though the vast majority of variation comes from scheduling.

HCRIS. I use Medicare Cost Reports to extract annual facility-level RN compensation from Worksheet S-3 Part II—both wages and fringe benefits—total agency expenditures, patient census, and facility characteristics. County-year agency wages w_a are

[15] Seo and Spetz (2013) show that hospitals have long used agency nurses to manage this trade-off, with temporary staffing rising when demand uncertainty and labor market tightness make maintaining a permanent buffer costly.

constructed as aggregate agency expenditures divided by aggregate agency hours from PBJ. Permanent labor cost is hourly wages w_e multiplied by a benefit loading from the HCRIS data and employer-side payroll tax rate of 0.1: $mc_e = w_e(1 + b_{\text{ben}} + 0.1)$.

Table 1 reports summary statistics for the main estimation sample of 38,798 facility-years from 2018 to 2024. The median facility has a daily census of 77 residents and a staffing need of roughly 30 RN hours per day. On average, employee supply falls short of need by 1.0 hours per day, and employee supply is volatile: the mean CV of employee hours is 0.37, far exceeding the CV of census at 0.07. As a result, facilities fall below 90 percent of their acuity-adjusted staffing minimum on 40 percent of days. Agency nurses account for only 5.8 percent of observed RN hours on average, well below the frictionless newsvendor optimum. The median agency wage premium is 1.61, meaning agency nurses cost 61 percent more per hour than permanent employees on a total-compensation basis. Panel C reports the structural parameters for the identified subsample; the median frictional markup is 0.31 and the median marginal rationing cost evaluated at the typical agency-day shortfall is $84 per hour.

4.2 The Post-2020 Nursing Boom

The nursing labor market tightened sharply beginning in 2020. Appendix Figure A1 shows annual mean RN wages rising from roughly $75,000$ in 2019 to more than $98,000$ by 2024, a roughly 30 percent increase over five years. Figure A3 shows the hospital aggregate weekly payroll index rising steeply after 2020, and Appendix Figure A2 shows the JOLTS quit rate for health care and social assistance spiking in 2021–2022. The nursing data do not fit the pattern predicted by job-ladder models (Burdett and Mortensen, 1998; Moscarini and Postel-Vinay, 2012) or related monopsony mechanisms (Autor et al., 2024), which imply that wage growth should be at least partly

compositional and separations should be disproportionately concentrated at lower-wage employers. A shift-share decomposition of RN wage growth from 2020 to 2024 shows that 98 percent of the increase occurred within facilities, not through reallocation toward higher-paying employers. Stratifying by pre-pandemic wage tercile, churn increases were symmetric across all three terciles (Appendix Figure A8), and facility wages explain essentially none of the cross-facility variation in separation changes whether entered alone or with controls (Appendix Figure A4). County fixed effects explain roughly one quarter of the variation. Separation changes are therefore primarily a market-level phenomenon, not concentrated among low-wage employers.

Several indicators point to the increased importance of the spot market. Figure 1 shows RN agency hours rising from roughly 3–4 percent of total RN hours before the pandemic to more than 10 percent by 2023 before partly receding, consistent with the national trend documented in Bowblis et al. (2024). The same shift appears in aggregate employment dynamics. The seasonal amplitude of the hospital payroll index nearly doubled after 2020, from 9.6 in the pre-pandemic period to 19.3 afterward, consistent with a staffing margin that became more demand-responsive (Figure A3).

COVID-19 increased both the level δ_H and volatility π_H of nursing demand, raising the value of as-needed spot coverage relative to holding a large permanent buffer. As more facilities turned to agency labor, the spot market thickened, frictions fell, and the scarcity value of flexible nurses rose, consistent with evidence that workers in thicker labor markets exhibit higher rates of churning (Bleakley and Lin, 2012). In the model that change pushes up wages at all facilities through the worker indifference condition rather than through worker reallocation across employers. Figure 2 shows a substantial hourly wage premium for contract workers relative to permanent employees in every major nursing occupation. The scarcity wage is higher than the buffer wage, as the model predicts. Appendix Figure A7 shows a post-2020 rise in

wage dispersion, also consistent with a shift towards the spot market.

5 General Equilibrium Evidence

I test the model's central general equilibrium prediction using two complementary approaches. First, I test whether thicker spot markets are associated with less rationing. Second, I test whether thicker spot markets experience less precautionary overstaffing and, in turn, whether there is a more even distribution of staff hours across facilities.

In Figure 5 counties are split by their baseline (pre-2020) agency share, and weekly RN rationing rates are plotted separately for each group after residualizing on log total nursing market hours. This residualization controls for the overall scale of nursing activity in the county, so differences across groups reflect coverage quality rather than total hours worked. High-agency counties exhibit persistently lower rationing throughout the sample: for the same number of total nursing hours, counties with a thicker agency market in the baseline period provide better coverage. After 2020, the gap widens. The increase in rationing that accompanied the COVID demand shock was more muted in counties that entered the pandemic with deeper spot market. When demand increases, a thicker market can absorb the shock more efficiently than a thin one.

Table 2 estimates the same relationship in a regression framework. The outcome is the fraction of days in the facility-year on which total RN staffing falls below 90 percent of the acuity-adjusted minimum, the newsvendor rationing measure used throughout the structural analysis. The regressor of interest is the leave-own-facility-out county agency share, which measures the thickness of the local spot market while excluding the facility's own hours to avoid mechanical correlation. Column (1) is the

bivariate regression, yielding a coefficient of -0.638 ($p < 0.01$): a ten-percentage-point increase in county agency share is associated with a 6.4-percentage-point reduction in the facility's rationing rate. Column (2) adds log county RN hours, which absorbs the scale of total nursing activity; the coefficient is essentially unchanged at -0.644, so the agency share effect reflects coverage quality conditional on total hours worked rather than a spurious correlation between market size and both agency density and rationing rates. Column (3) adds year fixed effects and the estimate rises slightly in magnitude to -0.791, suggesting that the cross-sectional relationship is not driven by secular trends in agency use. Column (4) adds county fixed effects, exploiting within-county variation across facilities over time; the coefficient falls to -0.236 but remains significant at the five percent level. The attenuation under county fixed effects suggests that counties that developed thicker agency markets may differ from thin-market counties on unobservables, or within-county variation in agency share may be contaminated by confounding time-varying shocks that are less of a concern in the cross-sectional comparison. With these caveats, these relationship suggest that thicker spot markets reduce rationing in labor markets.

Figure 6 examines the entire distribution of staffing outcomes. I run Firpo-Fortin-Lemieux unconditional quantile regressions (Firpo et al., 2009) of facility-level HPPD on county agency share. Two specifications are shown: a baseline controlling for log census, and series controlling for for-profit status and case mix index. The left panel shows employee (permanent) HPPD only. The coefficient on county agency share is near zero at low quantiles and turns increasingly negative at the top, around P90 and above. In counties with a thicker spot market, heavy-buffering facilities at the top of the permanent staffing distribution hold fewer permanent workers, while facilities at the bottom, which presumably were not over-buffering to begin with, are unaffected. The right panel adds agency hours to get total HPPD. The coefficient on total hours

is positive across most of the distribution and close to zero at the upper quantiles. Counties with higher agency share have more hours per patient day over most of the distribution except at the top. Facilities at the top of the distribution that shed permanent workers show no significant decline in total staffing as they replaced idle permanent buffers with flexible spot coverage. The interpretation is that the thick market raised total employment over most of the distribution, improving coverage for facilities that previously fell short, while the high-buffering facilities at the top achieved the same total coverage more efficiently.

6 Structural Estimation

The reduced-form evidence in Section 5 supports the model's general equilibrium predictions but does not speak to the magnitude of spot market frictions, the cost of unmet care, or how either has changed over time. This section takes the model to the data structurally. I recover, for each facility, the three parameters that govern the buffer-spot trade-off: α, the frictional markup paid above the posted spot market wage; θ, the marginal cost of unmet nursing need; and r, the tolerated residual shortfall of staffing. With these estimates in hand, I examine how frictions changed across the 2020 demand shock, test whether the change was larger where pre-pandemic markets were thinner, and ask in Section 7 whether the estimated valuations imply systematic over-buffering.

Applications of the newsvendor model are largely within operations management which takes cost parameters as known and solves for optimal procurement rules. The structural estimation here proceeds conditional on the observed equilibrium, recovering parameters from observed choices rather than solving for an equilibrium. Olivares et al. (2008) use the newsvendor first-order condition as a moment condition to recover

cost parameters from observed decisions in the context of operating room scheduling, estimating implicit underage and overage costs from staffing choices in a setting structurally similar to the one here. I follow that approach, by the simultaneous observability of two distinct margins: the ex-ante buffer level and the ex-post spot hiring decision on days when the permanent workforce falls short. Those two margins jointly identify the frictional markup α and the rationing cost θ from a quadratic cost specification without imposing parametric assumptions on the demand distribution. The gap distribution entering the buffer first-order condition is taken directly from the empirical distribution of daily shortfalls. The recovered parameters are inputs into the welfare calculations in Section 7.

Section 6.1 develops the setup and derives the basic objective function. Section 6.2 extends it to the full cost specification and derives the identifying moments.

6.1 Newsvendor Setup

Each SNF chooses permanent RN hours H (the buffer) to minimize expected daily staffing cost. In a canonical newsvendor formulation, permanent nurses cost w_e per hour and are paid whether or not they are needed; agency nurses cost $w_a > w_e$ per hour and are hired only after demand D_t is realized. The gap $G_t = D_t - H$ is the staffing shortfall on day t. On days when $G_t > 0$, the shortfall is filled by agency nurses. The facility's problem is:

$$\min_{H \geq 0} \quad w_e H + w_a \, \mathbb{E}\big[\max(G_t,\, 0)\big]. \tag{14}$$

The first term is the certain cost of the stable employee; the second is the expected cost of covering any shortfall with agency labor. The first-order condition equates the marginal cost of an additional permanent nurse to the expected savings from

displacing agency labor:

$$w_e = w_a \, P(G_t > 0). \tag{15}$$

The textbook solution to the newsvendor problem is to set the optimal buffer at the $(1 - w_e/w_a)$ quantile of the gap distribution.[16] Because we observe the gap empirically for every facility, the optimal buffer can be calculated.[17]

6.2 Structural Parameters

In practice, this simplified model misses two features of labor markets. An employer might tolerate rationing at some cost. And, as in the theoretical model, there may be frictional costs from participating in the spot market that are not reflected in the wage. To allow for these possibilities, I write the period-t variable cost of agency hiring, for a given level of buffer, as:

$$c_t(A_t) = w_a(1 + \alpha)A_t + \frac{\theta}{2}(S_t - A_t)^2, \tag{16}$$

where α is the agency friction (markup from search, placement fees, travel premiums, and scheduling frictions) and $\theta \geq 0$ is the marginal rationing cost (quadratic penalty for unmet need: care quality degradation, regulatory risk, and the cost of care gaps). A_t denotes agency hiring in period t and S_t is the shortfall from target hours in period t.

The employer then solves a two-stage problem. In stage 1, before demand is realized, it chooses permanent buffer H. In stage 2, after observing the realized gap $G_t = D_t - H$, it chooses agency hours $A_t \geq 0$. Let $S_t = \max(G_t, 0)$ denote the total

[16] The optimality condition follows from differentiating $w_e H + w_a \, \mathbb{E}[\max(D - H, 0)]$ with respect to H, yielding $w_e = w_a \, P(D > H)$, so H^* satisfies $F_D(H^*) = 1 - w_e/w_a$.

[17] The cost-minimization framing abstracts from insurance and bargaining because workers are assumed risk-neutral and wages are market-determined.

shortfall. The facility minimizes expected total cost:

$$\min_{H \geq 0} \; w_e H + \mathbb{E}\left[\min_{A_t \geq 0} \; c_t(A_t)\right]. \tag{17}$$

The parameter α is the empirical counterpart to the theory's match-quality discount $1 - \gamma_H$. When the spot market thickens, the density (g) and congestion-relief (h) improvements reduce the effective frictional markup and α falls.

Solving the inner minimization, the ex-post FOC for A_t on days when agency is used ($G_t > r$) gives $w_a(1 + \alpha) = \theta(S_t - A_t)$, so the facility purchases agency up to the point where the marginal cost of an additional hire equals the marginal benefit of reducing unmet need, leaving a residual shortfall $r \equiv w_a(1 + \alpha)/\theta$. On days $0 < G_t \leq r$ the corner solution binds and no agency is used such that the full shortfall is rationed. On days $G_t \leq 0$ no shortfall exists. The quantity r is therefore observable as the mean of $S_t - A_t$ on agency-using days, which is moment 1.

Moment 1 (ex post, days spot labor is used). The interior FOC for A_t on agency days is:

$$w_a(1 + \alpha) = \theta \cdot r_i, \tag{18}$$

where r_i is the mean residual shortfall on agency days at facility i (the average of $S_t - A_t$ when $A_t > 0$, observable directly from PBJ). This identifies α/θ.

Moment 2 (ex ante, buffer choice). The FOC for the permanent buffer H, obtained by differentiating (17) with the optimal A_t^* substituted in, is:

$$w_e = w_a(1 + \alpha)\left[P(G > r) + \frac{\mathbb{E}[G \cdot \mathbf{1}_{\{0 < G \leq r\}}]}{r}\right]. \tag{19}$$

Given observed r_i, w_e, w_a, and the empirical gap CDF, this pins down α without functional form assumptions on the demand distribution. With α identified, $\theta =$

$w_a(1+\alpha)/r_i$ follows immediately from (18).

Moment 1 draws on the ex-post spot-hiring pattern on high-demand days; moment 2 draws on the ex-ante permanent staffing level relative to the demand distribution. Together they identify $\hat{\alpha}_i$ and $\hat{\theta}_i$ separately for each facility, with $\hat{\theta}_i$ capturing the facility's marginal cost of a care gap, and neither requires functional form assumptions on the demand distribution.

6.3 Defining the Staffing Target

The gap and shortfall measures entering the two moments require a definition of daily staffing need at each facility. A natural target is the regulatory minimum RN HPPD implied by each facility's CMS staffing star rating and resident acuity. CMS publishes expected staffing thresholds for each staffing domain star level and scales them for resident acuity using the ratio of the facility's RUG-IV case-mix index to the national average:

$$h^*_{jt} = \text{RN_thresh}[\text{star}_{jt}] \times \frac{\text{CM_RN}_{jt}}{\overline{\text{CM_RN}_t}}, \tag{20}$$

where star_{jt} is the staffing domain star rating from the November NHC snapshot for year t and $\overline{\text{CM_RN}_t}$ is the national average case-mix index. Daily staffing need is defined as $\text{need}_{jt} = h^*_{jt} \times \text{census}_{jt}$, and the daily gap is $G_{jt} = \text{need}_{jt} - \text{total}_{jt}$.[18] I restrict the estimation sample to facility-years in which employee supply falls short of the acuity-adjusted minimum on at least 5 percent of days, ensuring that the constraint is operative.[19]

[18] Star ratings and case-mix indices vary annually; 2017–2018 observations use 2019 values and 2024 observations use 2023 values to avoid the PDPM methodology break in July 2024. The target is thus identified from both cross-sectional differences in star rating and acuity and within-facility year-to-year changes in either.

[19] A facility that is almost never below the minimum is not making a meaningful buffer choice relative to this floor; the 5 percent threshold retains facilities for which the newsvendor trade-off is economically active. Results are robust to thresholds ranging from 5 to 20 percent, or to dropping

I conduct two analyses to confirm that this target is relevant and that actual staffing tracks h^*_{jt}. The left panel of Figure 10 shows raw employee HPPD distributions by star rating. The HPPD distributions shift rightward almost one-for-one with the group median CMS minimum (within star classification there is variation in minimums due to different acuity), from roughly 0.22 HPPD for 1-star facilities to 0.46 for 5-star facilities.[20] The right panel re-centers the x-axis by each facility's own h^*_{jt} and plots the histogram of HPPD relative to distance to the CMS minimum. The mass of HPPD days is approximately centered at the minimum across all star groups, providing visual confirmation that facilities treat the acuity-adjusted minimum as the relevant target. The normalized distributions also reveal that higher-star facilities hold more excess staffing above their minimum, consistent with the model's prediction that higher rationing costs θ induce more precautionary buffering.

An event study confirms that facility HPPD levels move when minimums change. The design is a stacked difference-in-differences around cohorts experiencing a shift in h^*_{jt} in 2020, 2021, and 2022, with never-treated facilities as controls, facility-by-cohort and year-by-cohort fixed effects, and county-clustered standard errors.[21] The treatment is the realized shift in h^*_{jt}, normalized to one standard deviation. Figure 11 shows the estimates. Employee HPPD rises sharply with an increase in minimums, by about 8 log points at event year zero ($\hat{\beta} = 0.080$, $t = 15.1$) following a one standard deviation increase in the minimum. There is no evidence that an increase in minimums results in a higher hourly wage of staff RNs, inconsistent with the monopsony prediction following an expansion of employment. This finding tracks with

the criterion entirely.

[20] This figure is broken-down by star rating, but within a star rating there is variation in minimums due to varying facility accuity.

[21] The sample covers $N = 46{,}598$ facility-cohort-year observations across $G = 1{,}240$ treated cohort-facility groups with a ± 2 year window around the event. The event window is balanced within each cohort.

the results in Matsudaira (2014) who found no wage response to employment gains for state-level minimums. Agency share falls by roughly 8 log points ($\hat{\beta} = -0.077$, $t = -3.0$) following an increase in the minimum, consistent with the theoretical model's sorting prediction because a higher regulatory minimum raises the effective rationing cost θ, shifting facilities toward the permanent buffer margin and away from ex-post spot coverage.[22]

6.4 Empirical Implementation

The two identifying moments map directly to PBJ observables once the staffing target is defined. For each facility i and day t, employee hours H_{it} and agency hours A_{it} are observed separately in the PBJ records. The daily gap is $G_{it} = \text{need}_{it} - H_{it}$, where $\text{need}_{it} = h^*_{it} \times \text{census}_{it}$ from equation (20). The total shortfall is $S_{it} = \max(G_{it}, 0)$, and the residual shortfall on days when agency is used is $S_{it} - A_{it}$ when $A_{it} > 0$.

Moment 1. The residual shortfall r_i in equation (18) is estimated as the mean of $S_{it} - A_{it}$ over all agency-using days within facility i's sample period. This is directly observable as the average of the gap r_i on days when a facility calls in agency nurses but still ends the day below its target.[23]

Moment 2. Given r_i, the quantities $P(G > r_i)$ and $\mathbb{E}[G \cdot \mathbf{1}_{\{0<G\leq r_i\}}]$ in equation (19) are computed from the empirical distribution of daily gaps $\{G_{it}\}$ within the facility-

[22]The reduction in agency share is also contrary to the monopsony prediction since a facility with an upward sloping labor supply curve might respond to a higher minimum through agency staffing without raising wages on the remaining workforce. Such a response is akin to the diseconomies of scale in recruitment in Manning (2006) .

[23]Facilities must have at least 10 agency-using days per year to enter the structural sample. This restriction ensures r_i is estimated from a sufficient number of ex-post observations. The requirement that $r_i > 0$—i.e., that some residual shortfall remains even on agency days—is the identifying signature of the quadratic cost model. If the linear model held exactly, agency hiring would always close the gap completely and $r_i = 0$. In the data, 37.8 percent of agency days have a residual shortfall exceeding half an hour, consistent with the quadratic specification.

year. The gap CDF is estimated nonparametrically from the observed daily shortfalls. $P(G > r_i)$ is the fraction of days on which the employee shortfall exceeds r_i; $\mathbb{E}[G \cdot \mathbf{1}_{\{0<G\leq r_i\}}]$ is the mean gap on days with a moderate shortfall between zero and r_i, that is, days where the model predicts no agency will be called. Both are estimated by their sample analogues over the facility's annual record.

Wages. Employee hourly wage is the HCRIS-derived hourly compensation. The county-year agency wage w_a is constructed as aggregate agency expenditure divided by aggregate agency hours from PBJ, so it reflects the market rate actually paid by facilities in the county.[24] The ratio w_e/w_a determines the frictionless newsvendor quantile; identification of α comes from the deviation of the observed buffer from that frictionless benchmark after conditioning on the empirical gap distribution.

Estimation sequence. For each facility-year: (i) compute r_i from Moment 1; (ii) evaluate the right-hand side of Moment 2 at r_i using the empirical gap CDF; (iii) solve for α_i from equation (19); (iv) recover $\theta_i = w_a(1 + \alpha_i)/r_i$ from equation (18). The procedure is closed-form and requires no numerical optimization.[25]

For this analysis I use data covering 9,548 SNFs matched to HCRIS wages and CMS data that are inputs to calculate minimum thresholds.

County-level estimation of α. The friction parameter α is a market-level object. In the theory, $\alpha = 1/\gamma_H - 1$ measures the equilibrium quality discount prevailing in the county's spot market, common to all facilities drawing from the same pool. Averaging the facility-level estimates $\hat{\alpha}_i$ within county-years is feasible but introduces substantial

[24] Using the county-year average rather than facility-specific agency costs absorbs idiosyncratic contracting arrangements and aligns with the theory, where w_a is a market-clearing price common to all spot hirers in the county.

[25] The bracketed term in (19) is monotone in α through $r = w_a(1 + \alpha)/\theta$, so equation (19) has a unique solution given observed wages and the empirical gap moments.

sampling error: each $\hat{\alpha}_i$ is identified from only 50–150 facility-days. Applying the two structural moments directly to the pooled county-day gap distribution yields more precise estimates at the relevant unit of analysis.

For each county-year, I pool all facility-days from identified facilities and construct county-level analogs of Moments 1 and 2.[26] Moment 1 yields a county-level residual shortfall $\hat{r}_{ct}$ from the mean of $S_{it} - A_{it}$ across all agency-using days in the county-year. Moment 2 applies the buffer FOC to the aggregate county-day gap CDF:

$$\bar{w}_{e,ct} = \hat{w}_{a,\text{eff},ct} \left[P(G > \hat{r}_{ct}) + \frac{\mathbb{E}[G \cdot \mathbf{1}_{\{0 < G \leq \hat{r}_{ct}\}}]}{\hat{r}_{ct}} \right], \tag{21}$$

where $\bar{w}_{e,ct}$ is the hours-weighted mean employee wage in the county-year. Solving gives $\hat{w}_{a,\text{eff},ct}$ and hence $\hat{\alpha}_{ct} = \hat{w}_{a,\text{eff},ct}/w_{a,ct} - 1$.[27] All market-level analyses that involve α as an outcome use this county-year pooled measure $\hat{\alpha}_{ct}$.

6.5 Results

Frictionless benchmark With no search or placement markup ($\alpha = 0$ in (17)) agency labor is available at the posted wage w_a, the residual shortfall $r = w_a/\theta$ is determined entirely by the rationing cost θ, and the facility sets its buffer at the quantile of the gap distribution where the marginal cost of permanent labor equals the expected agency cost avoided.[28]

[26] Identified facilities require two conditions: (1) the facility must have at least 10 agency days to estimate the mean residual shortfall $\hat{r}_i$ from the first structural moment, (2) the effective agency cost $w_{a,\text{eff}}$ must be estimable, which requires the facility to be operating at an interior solution — sometimes short-staffed and sometimes using agency — so that both the rationing cost θ and the buffer level r can be identified from the empirical gap distribution.

[27] The county-level pooled estimator does not require θ directly. The rationing cost θ_i is estimated in the facility-level stage alongside the tolerated shortfall threshold $\hat{r}_i$; these facility-level estimates are embedded in the construction of the daily residual shortfall $\text{resid}_{it} = \max(G_{it} - \hat{r}_i - A_{it},\ 0)$.

[28] The canonical newsvendor sets $\alpha = 0$ and $\theta \to \infty$, forcing $r \to 0$ so the facility covers the gap completely with agency labor on every high-demand day. Under those two restrictions, (17) collapses to the basic formulation in (14).

The newsvendor model predicts that facilities with more volatile permanent staffing should use more agency. A facility with a more dispersed gap distribution has a higher gap at any given quantile, so H^* is larger in levels but covers a smaller fraction of realized gaps. Table 4 shows this relationship (Panel A) along with the implied frictionless-optimal agency share assuming facilities staff optimally (Panel B). Regressing observed agency share on the CV of employee RN hours yields a bivariate coefficient of 0.249 (column 1), highly significant and essentially unchanged after adding controls, year fixed effects, or county fixed effects, indicating the relationship is not driven by sorting across markets. The optimal-share gradient on employee-hour volatility is smaller (0.097–0.120), while the gradient on census volatility is larger in Panel B than in Panel A. The results confirm the directional prediction of the model.

Figure 7 plots observed and frictionless-optimal ($\alpha = 0$) agency share over time. In every year facilities use substantially less agency than the frictionless benchmark prescribes, confirming that spot market frictions are empirically relevant. The frictional markup α reconciles what facilities "should" do given the relative wages of regular and agency versus what we observe them to do.

With frictional markups Panel C of Table 1 summarizes the three structural parameters recovered from the newsvendor moments. The median frictional markup is $\hat{\alpha} = 0.31$, meaning the effective cost of an agency nurse exceeds the posted wage by roughly 31 percent once search, placement, and scheduling frictions are accounted for; the distribution is right-skewed, with a mean of 0.36 and a 90th-percentile value of 0.84. The median marginal rationing cost, evaluated at each facility's typical agency-day shortfall, is $\hat{\theta} \times r = \$84$ per hour, rising to \$115 at the 90th percentile, reflecting substantial heterogeneity in the cost of unmet nursing need across facilities. The median tolerated residual shortfall is 2.1 hours per day on agency-using days—

roughly 2.6 percent of the acuity-adjusted staffing target—implying that even when facilities call in agency nurses they leave meaningful unmet need at the margin.

Figure 8 shows the time series of $\hat{\alpha}_{ct}$ from 2018 to 2024 and Table 3 reports the annual distribution of the county-year frictional markup, along with significance tests against the 2018 baseline and year-on-year changes. The mean markup is stable at roughly 0.30 in 2018–2019, then falls sharply: by 2021 it has declined to 0.178, a drop of 0.126 relative to the 2018 mean of 0.304 that is significant at the one-percent level ($t = -4.72$). The 2022 mean is essentially unchanged at 0.179, also highly significant relative to 2018 ($t = -4.58$). The markup then reverses, rising to 0.286 in 2023—statistically indistinguishable from 2018—and 0.370 in 2024, which is significantly *above* the 2018 baseline ($t = 2.50$, $p = 0.012$). Both permanent and agency wages rose sharply over the 2020–2022 period, so the decline in $\hat{\alpha}$ reflects a compression of frictional markups above posted wages rather than tracking wages, consistent with a thicker, more competitive spot market. Agency share and agency wages nonetheless remain well above their 2018 levels as $\hat{\alpha}$ reverts, sustained by rationing costs $\hat{\theta}$ that remained elevated.

7 Welfare

This section uses the structural estimates to ask whether firms hold too much labor in private employment relative to the shared spot market. I first ask whether the marginal buffer nurse would create more expected value in the spot market than as a regular employee, then I estimate the participation externality from moving a nurse from buffer to pool.

The value of the marginal buffer nurse in the spot market At the buffer first-order condition, the marginal employee is worth exactly w_e to the facility holding her. If that same nurse would create more expected value in the county pool, the competitive equilibrium involves too much buffering. For each facility i on each day t, the shadow value of one additional nurse-hour is:

$$v_{it} = \begin{cases} 0 & \text{if } G_{it} \leq 0 \quad \text{(fully staffed)} \\ \hat{\theta}_i \cdot G_{it} & \text{if } 0 < G_{it} \leq \hat{r}_i \quad \text{(rationing band)} \\ w_a(1+\hat{\alpha}_i) \equiv w_{a,\text{eff},i} & \text{if } G_{it} > \hat{r}_i \quad \text{(agency-using days).} \end{cases} \tag{22}$$

On agency-using days the shadow value (the marginal benefit of avoiding an additional agency hire at the interior FOC) equals $w_{a,\text{eff},i}$. On rationing-band days it equals the marginal rationing cost times the gap. On fully staffed days it is zero.

Define the maximum county shadow value in day t as $V_{ct} = \arg\max_j v_{jt}$. For each i, I compare the county's maximum shadow value excluding its own demand to w_e:

$$V_{-i,ct} = \begin{cases} V_{ct}^{(2)} & \text{if } i = \arg\max_j v_{jt} \text{ and } v_{it} > 0 \\ V_{ct} & \text{otherwise,} \end{cases} \tag{23}$$

where $V_{ct}^{(2)}$ is the second-highest shadow value in county c on day t.

Define $\rho_i = \frac{\mathbb{E}[V_{-i,ct}]}{w_{e,i}}$, averaged over all facility-days in the year. A value $\rho > 1$ means the marginal buffer nurse would create more expected value in the county pool than she costs in private employment. Because the private FOC pins the marginal value at exactly w_e, any pool value above w_e is evidence of a social-private gap.

Figure 9 summarizes the evidence. The left panel plots $\mathbb{E}[V_{-i,ct}]$ against mc_e facility-by-facility; the majority of points lie above the 45-degree parity line. The

right panel shows the pool value ratio $\mathbb{E}[V_{-i}]/mc_e$ over time; it exceeds one in 72.5 percent of county-years and is well above parity in every year of the sample. When $\rho > 1$ there is utilization inefficiency: buffer nurses are idle roughly half the time at their home facility while other county facilities have unmet need on those same days, and there is no price that compensates the holding facility for releasing a nurse.

The market-thickening participation externality I now estimate how much a nurse moving from buffer to pool improves the county spot market for other facilities, the channel in equation (13).

The county-level pooled $\hat{\alpha}_{ct}$ estimated in Section 6 is the building block for quantifying the externality. Since $\hat{\alpha} = w_{a,\text{eff}}/w_a - 1$ and $\gamma_H = 1/(1+\alpha)$, the matching function gradient is

$$\frac{\partial \gamma_H}{\partial(\text{agency share})} = \frac{|\beta_\alpha|}{(1+\alpha)^2}, \tag{24}$$

where $\beta_\alpha = \partial\alpha/\partial(\text{agency share})$ is estimated from a county-year regression of $\hat{\alpha}_{ct}$ on the county leave-out agency share. Table 3 reports the alpha time series from 2018 to 2024. Table 5 regresses $\hat{\alpha}_{ct}$ on the county leave-out agency share, yielding $\hat{\beta}_\alpha = -0.4553$ ($t = -4.95$), the primary estimate entering the A^* calculation.

For each county-year, the participation externality per nurse engagement is

$$A^*_{ct} = P(\text{agency day}) \cdot \mathbb{E}[Q_H \mid \text{agency day}] \cdot \frac{|\hat{\beta}_\alpha|}{(1+\hat{\alpha}_{ct})^2} \cdot \frac{1}{H_{\text{county},ct}}, \tag{25}$$

where $Q_{H,ct} = \sum_i w_{a,\text{eff},i} \cdot \text{ctr_hours}_{it}$ is the aggregate revealed-preference value of agency labor deployed on tight days, $H_{\text{county},ct}$ is the total nurse-hours in the county pool, and $P(\text{agency day})$ serves as the empirical analog of $\pi_H \delta_H$. [29]

[29] Because the structural estimates cover identified facilities only, A^* is scaled by the county-year coverage factor $n_{\text{all}}/n_{\text{id}}$ under the assumption that unidentified facilities have proportional rationing costs.

Table 6 reports the results. The coverage-corrected A^* rises from \$896 per engagement in 2018 to \$2,444 in 2022, averaging \$1,776 across the sample. The increase reflects Q_H growing faster than the matching function flattens as α falls. The 2024 decline reflects both lower Q_H and a higher α that reduces the matching function gradient.

These estimates imply a Pigouvian subsidy. When a facility releases one nurse from its private buffer into the spot market, it saves w_e per hour and expects to pay $w_a(1+\alpha)$ per hour when it needs her back via spot. By the buffer FOC these cancel exactly at the margin: the facility was indifferent, so the expected spot cost equals the wage saving. What remains is the externality A^*_{corr}, the county-level gain that flows to all n_{all} county facilities proportionally through the thicker pool. The switching facility creates A^*_{corr} in social value but captures only $A^*_{\text{corr}}/n_{\text{all}}$, its own share of the improved market access. The remainder $A^*_{\text{corr}} \times \frac{n_{\text{all}}-1}{n_{\text{all}}}$ flows to other facilities and is not internalized. The required Pigouvian subsidy per engagement is $A^*_{\text{corr}} \times \frac{n_{\text{all}}-1}{n_{\text{all}}}$, or expressed as a rate relative to the agency wage: $\frac{n_{\text{all}}-1}{n_{\text{all}}} \times \frac{A^*_{\text{corr}}}{w_a \times \text{hrs}}$.

Table 7 reports the estimated optimal subsidies. On average, a switching facility creates \$1,756 per engagement in social value but captures only \$295 (17 percent). The Pigouvian subsidy rate rises from 2.5 percent of the engagement wage pre-pandemic, peaked at 5.9 percent in 2022 as A^* grew with market thickening, then partially reverses to 4.8 percent in 2024. With approximately 23 facilities per county on average, each individual facility's private incentive to release a buffer worker is small.

8 Conclusion and Discussion

This paper studies the choice between holding labor as private buffer capacity and obtaining it from a shared spot market after demand is realized. Treating the spot market as a shared institution has general equilibrium implications for the standard staffing problem. A firm's staffing choice affects not only its own exposure to uncertainty but also the thickness and quality of the market that other firms rely on. Competitive equilibrium need not allocate flexible labor efficiently: firms hold too much private buffer capacity because they do not internalize the value their spot-market participation creates for others. The same complementarity can make market structure state-dependent, as a demand shock raises the value of flexibility directly and may also improve market functioning by thickening the shared market. Wages and churn can rise together without a job ladder, with wages rising because the value of flexible labor increases and churn rising because more workers are employed through short-duration matches. The same structure appears in any market where firms choose between private buffer capacity and a shared adjustment margin.

The welfare results show that the marginal buffer nurse creates more expected value in the county pool than in private employment. A decentralized price does not correct this. The Pigouvian subsidy rate estimated here relies on a tax-and-transfer apparatus that does not exist in most of these markets. However, in labor markets there are institutions that can be seen as solving the problem of missing spot markets. An example is craft union hiring halls. In the building trades, longshore work, and similar industries, workers obtain employment through a union-operated dispatch system. Workers register with the hall, employers request labor through the hall, and dispatch follows seniority so that workers with longer tenure receive priority. The seniority queue creates an asset, the position in line, that workers

accumulate through continued attachment to the hall. A worker who drifts away to outside employment loses seniority standing and must rebuild from the back of the queue. This binding mechanism creates spot market liquidity. Workers remain attached to the hall even during slack periods because their accumulated seniority has value. They respond when called because refusing dispatch damages their standing. The union controls entry through apprenticeship requirements and licensing, ensuring that seniority is not diluted by unlimited entry. The resulting workforce has the characteristic that defines a thick market: workers who remain attached and respond reliably when demand materializes. In terms of the model, the hall sustains a thick market in both the high and low demand states, so the frictional markup facing any individual employer is low and rationing in the tight state is limited. This reliability is the product the competitive spot market does not supply: the uncompensated externality is internalized through the institutional structure rather than through prices.

References

Autor, David, Arindrajit Dube, and Annie McGrew, "The Unexpected Compression: Competition at Work in the Low Wage Labor Market," *Quarterly Journal of Economics*, 2024. forthcoming.

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, 1975, *83* (6), 1183–1202.

Baily, Martin Neil, "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, 1974, *41* (1), 37–50.

Bergeaud, Antonin, Pierre Cahuc, Clément Malgouyres, Sara Signorelli, and Thomas Zuber, "The wage of temporary agency workers," Discussion Paper DP19261, Centre for Economic Policy Research 2024.

Bleakley, Hoyt and Jeffrey Lin, "Thick-market effects and churning in the labor market: Evidence from US cities," *Journal of urban economics*, 2012, *72* (2-3), 87–103.

Bowblis, John R, Christopher S Brunt, Huiwen Xu, Robert Applebaum, and David C Grabowski, "Nursing homes increasingly rely on staffing agencies for direct care nursing: study examines nursing home staffing," *Health Affairs*, 2024, *43* (3), 327–335.

Burdett, Kenneth and Dale T. Mortensen, "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, 1998, *39* (2), 257–273.

Diamond, Peter A., "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, 1982, *90* (5), 881–894.

Fay, Jon A. and James L. Medoff, "Labor and Output over the Business Cycle: Some Direct Evidence," *American Economic Review*, 1985, *75* (4), 638–655.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux, "Unconditional Quantile Regressions," *Econometrica*, 2009, *77* (3), 953–973.

Goldschmidt, Deborah and Johannes F Schmieder, "The rise of domestic outsourcing and the evolution of the German wage structure," *The Quarterly Journal of Economics*, 2017, *132* (3), 1165–1217.

Hall, Robert E., "Labor Demand, Labor Supply, and Employment Volatility," *NBER Macroeconomics Annual*, 1991, *6*, 17–47.

Helsley, Robert W and William C Strange, "Matching and agglomeration economies in a system of cities," *Regional Science and Urban Economics*, 1990, *20* (2), 189–212.

Houseman, Susan N, "Why employers use flexible staffing arrangements: Evidence from an establishment survey," *Ilr Review*, 2001, *55* (1), 149–170.

_ , Arne L Kalleberg, and George A Erickcek, "The role of temporary agency employment in tight labor markets," *ILR Review*, 2003, *57* (1), 105–127.

Howitt, Peter and R. Preston McAfee, "Animal Spirits," *American Economic Review*, 1992, *82* (3), 493–507.

Lachowska, Marta, Alexandre Mas, Raffaele Saggio, and Stephen A Woodbury, "Work hours mismatch," Technical Report, National Bureau of Economic Research 2023.

Manning, Alan, "A generalised model of monopsony," *The Economic Journal*, 2006, *116* (508), 84–100.

Mas, Alexandre and Amanda Pallais, "Valuing alternative work arrangements," *American Economic Review*, 2017, *107* (12), 3722–3759.

Matsudaira, Jordan D, "Monopsony in the low-wage labor market? Evidence from minimum nurse staffing regulations," *Review of Economics and Statistics*, 2014, *96* (1), 92–102.

Moretti, Enrico and Moises Yi, "Size matters: Matching externalities and the advantages of large labor markets," Technical Report, National Bureau of Economic Research 2024.

Mortensen, Dale T., "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, 1982, *72* (5), 968–979.

Moscarini, Giuseppe and Fabien Postel-Vinay, "The contribution of large and small employers to job creation in times of high and low unemployment," *American Economic Review*, 2012, *102* (6), 2509–2539.

Oi, Walter Y., "Labor as a Quasi-Fixed Factor," *Journal of Political Economy*, 1962, *70* (6), 538–556.

Olivares, Marcelo, Christian Terwiesch, and Lydia Cassorla, "Structural Estimation of the Newsvendor Model: An Application to Reserving Operating Room Time," *Management Science*, 2008, *54* (1), 41–55.

Seo, Sukyong and Joanne Spetz, "Demand for temporary agency nurses and nursing shortages," *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 2013, *50* (3), 216–228.

Figures

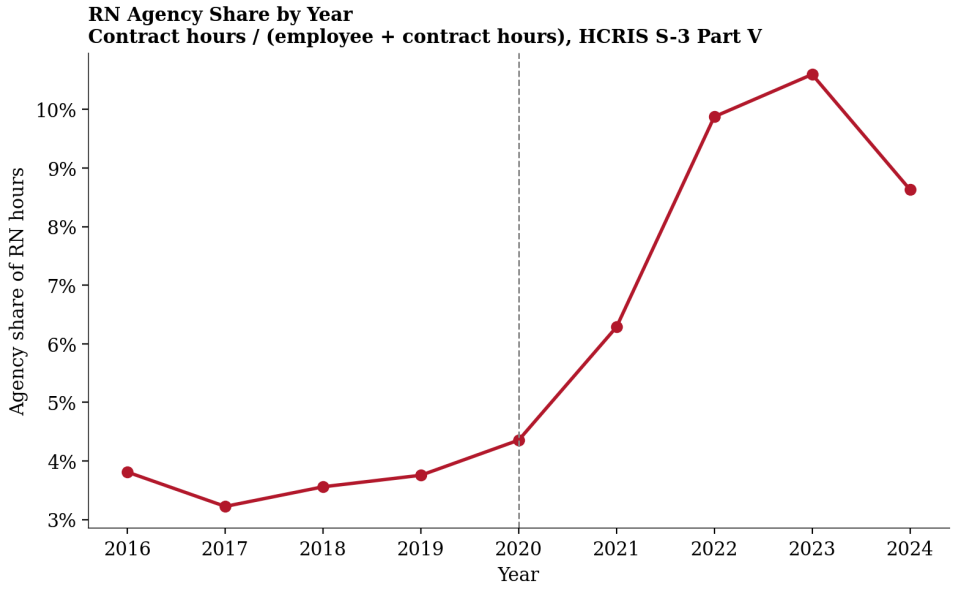


Figure 1: **RN agency share by year, SNFs.** Contract hours as a fraction of total RN hours, HCRIS Worksheet S-3 Part V, unbalanced panel 2016–2024.

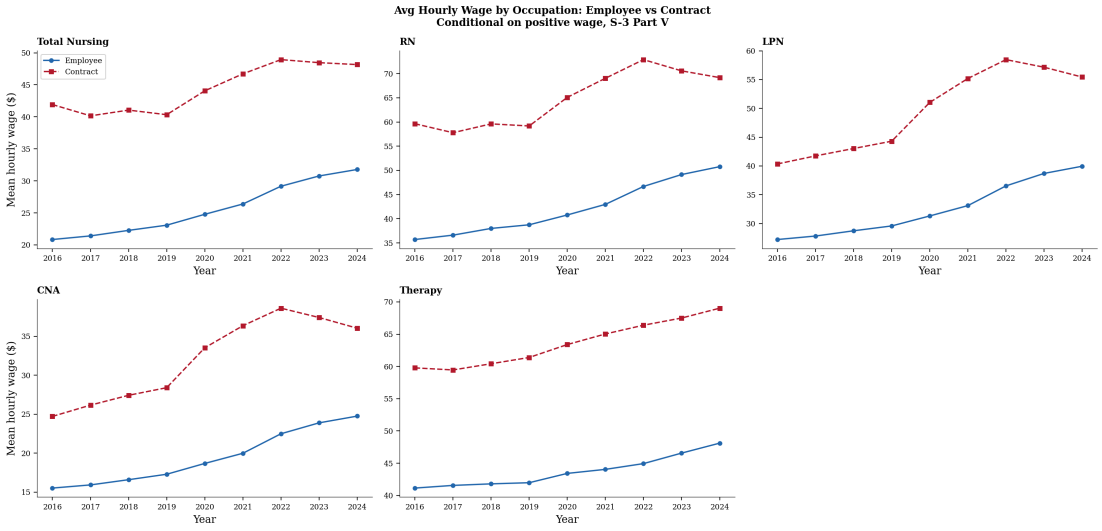


Figure 2: **Employee vs. contract wages by occupation, 2016–2024.** Mean hourly wage for employee (permanent, solid blue) and contract (dashed red) staff by occupation, conditional on positive reported wage, HCRIS Worksheet S-3 Part V. Panels show Total Nursing, RN, LPN, CNA, and Therapy. Sample: SNFs with positive reported wages in both staffing modes, unbalanced panel 2016–2024.

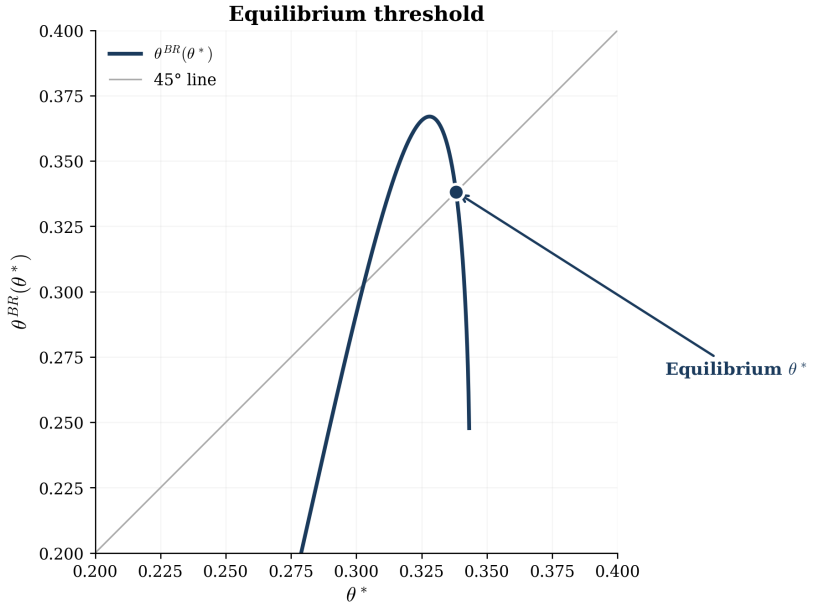


Figure 3: **Equilibrium threshold.** The best-response curve $\theta^{BR}(\theta^*)$ from (11) and the 45-degree line. The equilibrium cutoff θ^* is labeled at the intersection. Parameters: $F = \text{Beta}(2,2)$, $N = 0.85$, $\delta_H = 0.45$, $\delta_L = 0.05$, $\pi_H = 0.30$, $b = 0.015$.

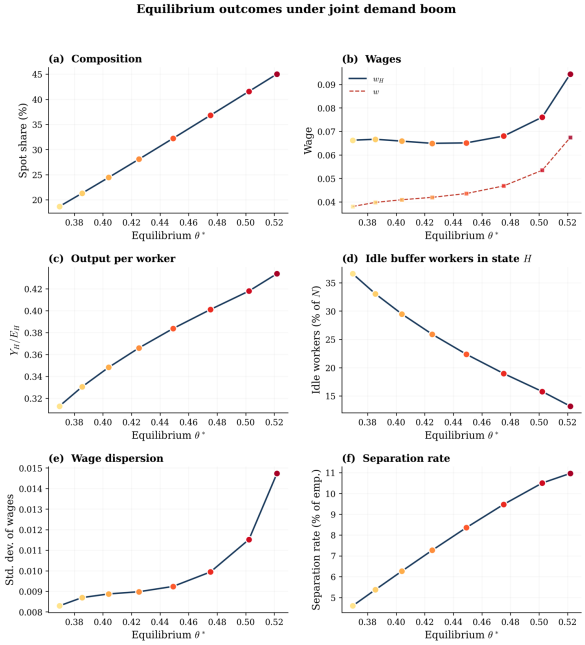


Figure 4: **Equilibrium outcomes under joint demand boom.** Panel (a): spot share (composition). Panel (b): scarcity wage w_H and buffer wage $\bar{w}$. Panel (c): output per worker (Y_H/E_H). Panel (d): idle buffer workers in state H (as % of N). Panel (e): wage dispersion (std. dev. of wages). Panel (f): separation rate. Parameters vary as (δ_H, π_H) increase jointly from $(0.55, 0.45)$ to $(0.76, 0.66)$, starting from a thin-market equilibrium.

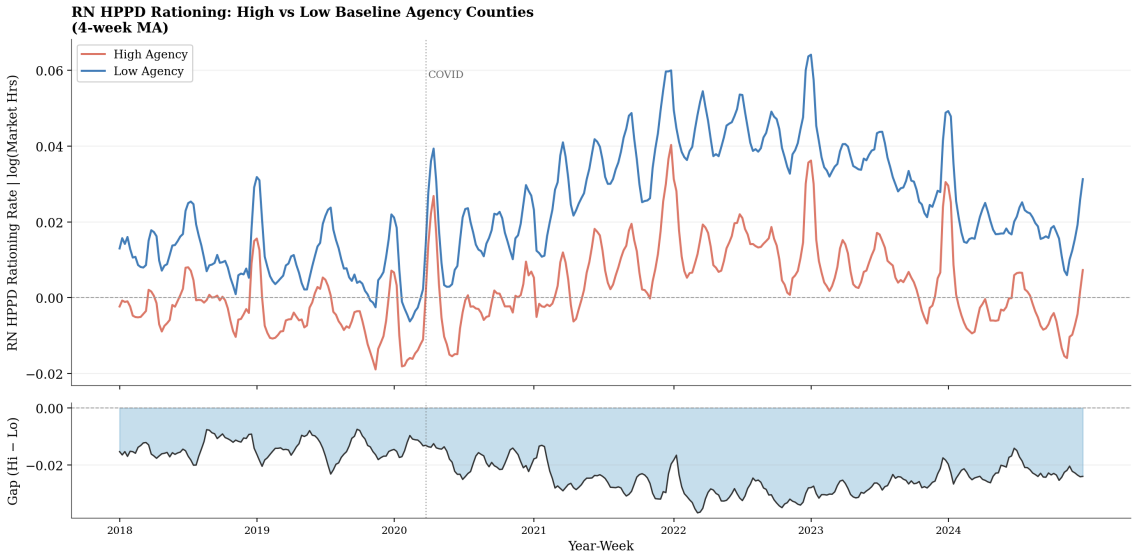


Figure 5: **RN rationing rate: high vs. low baseline agency counties.** Weekly RN HPPD rationing rate residualized on log total nursing market hours in the county-week, plotted as a 4-week moving average, 2018–2024. Counties are split at the median pre-2020 RN agency share. Top panel: residualized rationing rate for high-agency (orange) and low-agency (blue) county groups. Bottom panel: gap between the two groups (high minus low). Vertical dotted line marks COVID (2020).

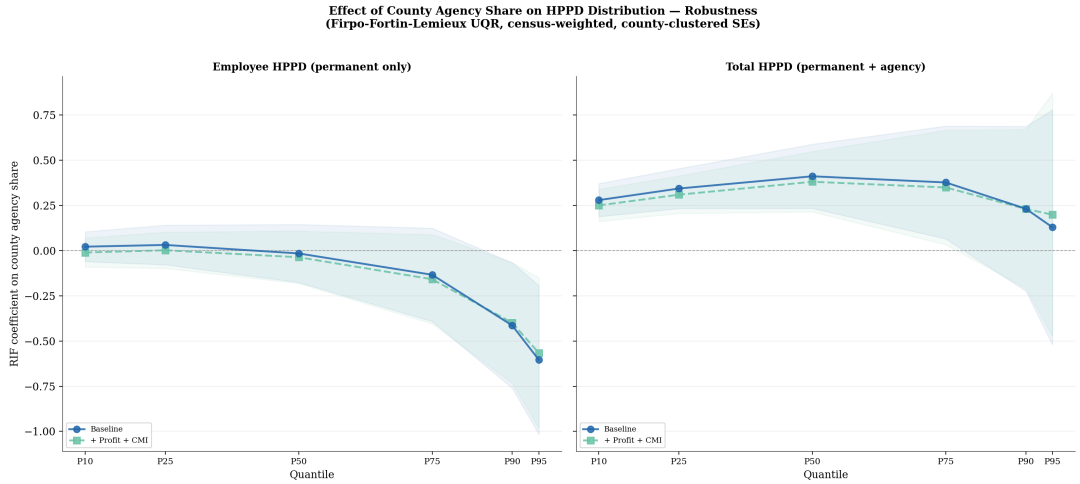


Figure 6: **Effect of county agency share on the HPPD distribution, pre-COVID (2018–2019).** Firpo-Fortin-Lemieux unconditional quantile regression (UQR) coefficients on county leave-own-facility-out RN agency share, 2018–2019. Census-weighted; standard errors clustered by county; shaded bands show 95% confidence intervals. Left panel: dependent variable is employee (permanent) RN HPPD. Right panel: dependent variable is total RN HPPD (employee plus agency). Two specifications: baseline controlling for log census (solid blue) and adding for-profit status and case mix index (dashed green, "+ Profit + CMI").

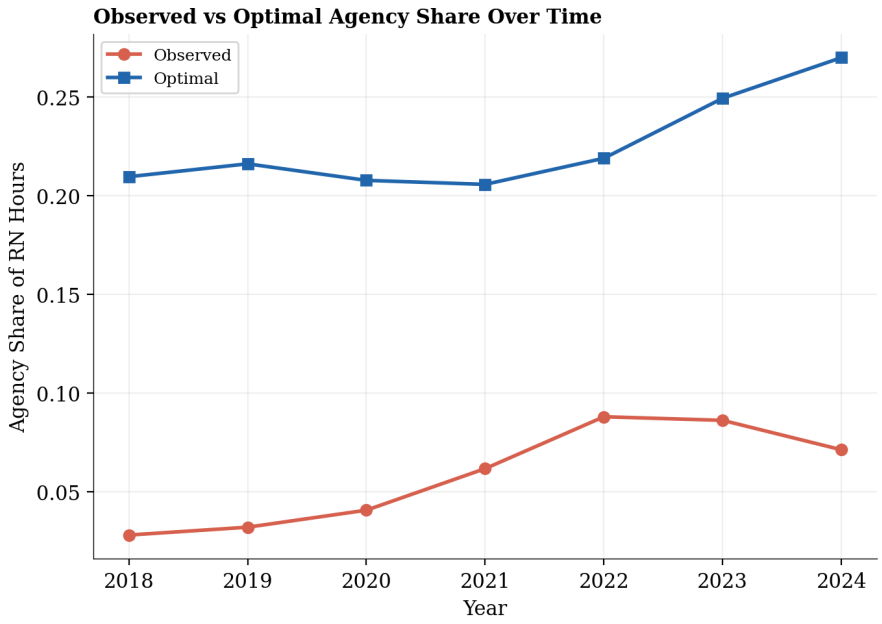


Figure 7: **Observed vs. frictionless-optimal agency use, 2018–2024.** Observed (red) and frictionless newsvendor optimum ($\alpha = 0$, blue) agency share of RN hours by year, averaged across facilities in the estimation sample. The frictionless optimum sets the permanent buffer at the $(1 - w_e/w_a)$ quantile of the empirical gap distribution.

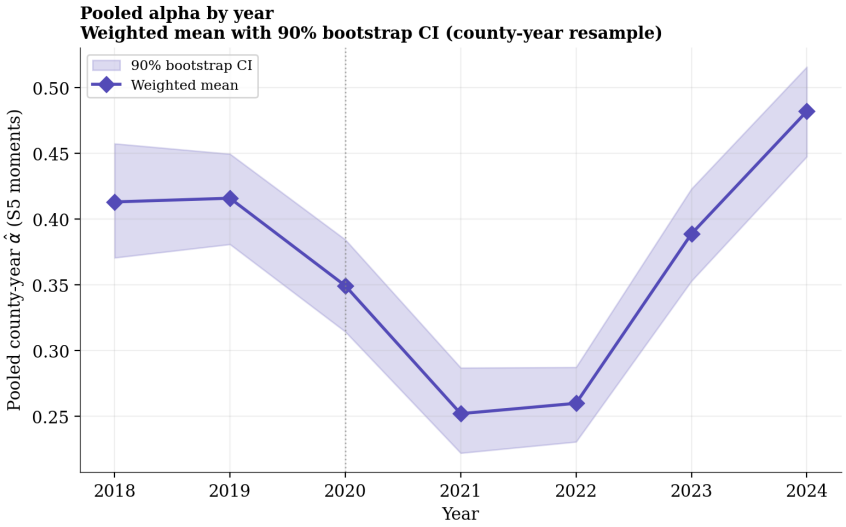


Figure 8: **Pooled county-year $\hat{\alpha}$ by year.** Weighted mean of the county-year pooled frictional markup $\hat{\alpha}_{ct}$, estimated by applying the two structural FOC moments to the pooled county-day gap distribution (S5 moments). Shading shows the 90% bootstrap confidence interval from a county-year resample. Dotted vertical line marks 2020. Sample: 696 county-years across identified facilities, 2018–2024.

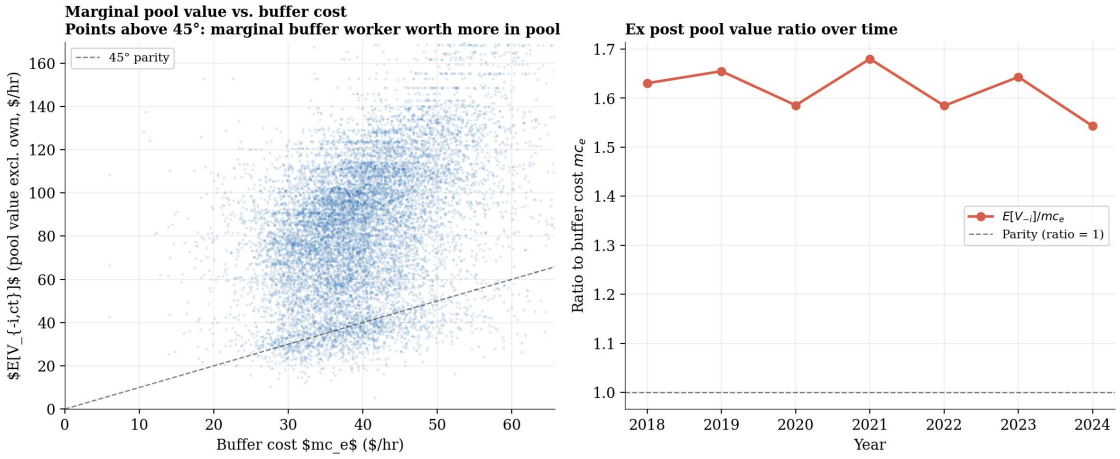


Figure 9: **Pool value versus private buffer cost.** Left panel: scatter of $\mathbb{E}[V_{-i,ct}]$ (expected shadow value of the marginal buffer nurse in the county pool, excluding own facility) against private buffer cost mc_e ($/hr); dashed line is the 45-degree parity line. Right panel: ex post pool value ratio $\mathbb{E}[V_{-i}]/mc_e$ by year; dashed horizontal line marks parity (ratio = 1). Sample: identified facility-years, 2018–2024.

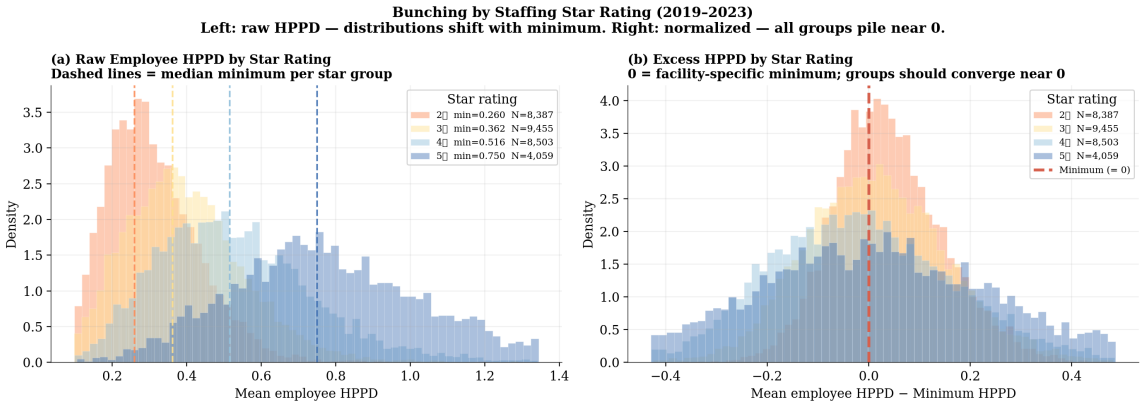


Figure 10: **Employee HPPD distributions by staffing star rating, 2019–2023.** Panel (a): raw mean employee HPPD by star rating (2–5 stars); dashed lines mark the median acuity-adjusted minimum h^*_{jt} for each group (2 star: 0.260, 3 star: 0.362, 4 star: 0.516, 5 star: 0.750). Panel (b): excess HPPD, defined as mean employee HPPD minus the facility-specific minimum h^*_{jt}. Sample sizes: 2 star: $N = 8{,}387$; 3 star: $N = 9{,}455$; 4 star: $N = 8{,}503$; 5 star: $N = 4{,}059$.

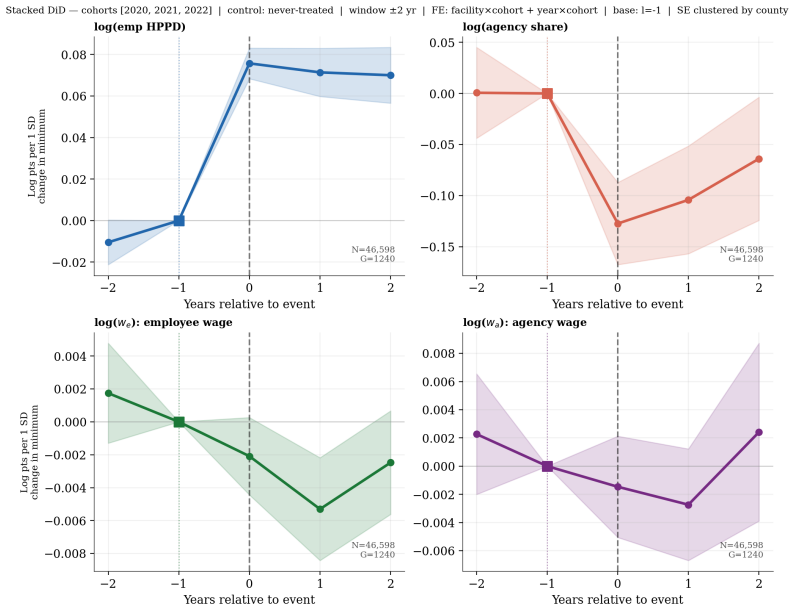


Figure 11: **Stacked difference-in-differences: facility response to shifts in the acuity-adjusted minimum.** Cohorts defined by facilities experiencing a shift in h^*_{jt} in 2020, 2021, and 2022; control group is never-treated facilities. Dose is the realized shift in the acuity-adjusted minimum (normalized to one standard deviation). Fixed effects: facility-by-cohort and year-by-cohort. Standard errors clustered by county; shaded bands show 95% confidence intervals. Base period: $\ell = -1$. Panels show log employee HPPD (blue), log agency share (red), log employee wage w_e (green), and log county agency wage w_a (purple). First stage shown in Appendix Figure A12. $N = 46{,}598$ facility-cohort-year observations; $G = 1{,}240$ treated cohort-facility groups.

Tables

Table 1: Summary Statistics. Sample: 38,798 facility-years, 2018–2024.

	N	Mean	SD	p10	p25	p50	p90
A. Demand, supply, and coverage							
Daily census (residents)	38,798	82.8	43.4	37.6	52.4	77.3	132.1
Daily staffing need (hrs)	38,798	35.3	23.1	15.0	20.6	29.7	61.6
Daily employee supply (hrs)	38,798	34.3	22.8	13.1	18.9	28.8	61.8
Daily gap: need minus supply (hrs)	38,798	1.040	13.2	-13.2	-6.767	-0.0013	15.8
CV of employee hours	38,798	0.367	0.203	0.209	0.264	0.347	0.525
CV of daily census	38,798	0.0686	0.0477	0.0294	0.0393	0.0557	0.121
Fraction of days total staffing $< 90\%$ of need	38,798	0.401	0.294	0.0538	0.142	0.342	0.863
Observed agency share of RN hours	38,798	0.0583	0.120	0.0000	0.0000	0.0020	0.194
Agency days per facility-year	38,798	21.5	35.2	0.0000	0.0000	1.000	80.0
B. Wages and cost ratios							
Employee RN wage w_e (\$/hr)	38,798	39.1	8.480	29.3	33.2	38.2	50.4
County agency wage w_a (\$/hr)	38,798	63.8	14.1	47.5	54.8	62.6	80.3
Agency wage premium w_a/w_e	38,798	1.688	0.485	1.201	1.377	1.612	2.212
C. Structural parameters							
Residual shortfall r (hrs/day on agency days)	14,596	2.980	3.308	0.297	0.904	2.079	6.649
Residual share r / pre-2020 census	14,596	0.0370	0.0421	0.0039	0.0116	0.0261	0.0804
Frictional cost $\hat{\alpha}$	13,855	0.356	0.378	-0.0717	0.102	0.311	0.842
Rationing cost $\hat{\theta} \times r$ (\$/hr)	13,855	86.9	22.5	62.0	72.0	84.3	114.8

Notes: Staffing need is the acuity-adjusted CMS minimum $h^*_{jt} \times \text{census}_{jt}$. Employee supply is daily RN hours from permanent staff. Wages are from HCRIS; county agency wage is aggregate agency expenditure divided by aggregate agency hours within the county-year. Panel C reports the structural parameters identified from the newsvendor moments; the smaller sample requires ≥ 10 agency-using days per facility-year ($N = 14,596$) and a valid buffer FOC estimate ($N = 13,855$). $\hat{\alpha}$ is the frictional markup above the posted agency wage, benefits-corrected. $\hat{\theta} \times r$ is the marginal rationing cost evaluated at the facility's typical agency-day shortfall r.

Table 2: County RN Agency Penetration and Facility Rationing

	(1)	(2)	(3)	(4)
County agency share (LOO)	-0.6379^{***}	-0.6441^{***}	-0.7908^{***}	-0.2355^{**}
	(0.0934)	(0.0879)	(0.1274)	(0.1021)
log(county RN hours)		−0.0078	−0.0063	-0.0496^{**}
		(0.0186)	(0.0180)	(0.0211)
Year FE	No	No	Yes	Yes
County FE	No	No	No	Yes
N	38,289	38,289	38,289	38,289
Counties	889	889	889	889
R^2	0.018	0.018	0.026	0.199

Notes: Dependent variable is the fraction of days in the facility-year on which total RN staffing falls below 90% of the acuity-implied minimum (the newsvendor rationing measure). All specifications estimated by WLS with weights equal to facility mean census. Standard errors clustered on county in parentheses. County agency share is the leave-own-facility-out fraction of annual RN hours supplied by agency staff in the county-year, weighted by facility total RN hours. log(county RN hours) is the log of total annual county RN hours. Specifications (1)–(3) include year fixed effects as dummy variables where indicated; R^2 is overall. Specification (4) absorbs county and year fixed effects via iterative demeaning; R^2 is within, exploiting within-county variation across facilities over time. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Pooled-Moment County-Year $\hat{\alpha}$: Annual Distribution and Significance Tests

	Distribution						vs. 2018		Year-on-year	
Year	Mean	SE	Median	P25	P75	N	t	p	t	p
2018	0.304	0.023	0.246	0.065	0.472	401	—	—	—	—
2019	0.298	0.015	0.250	0.092	0.458	440	-0.20	0.844	-0.20	0.844
2020	0.256	0.016	0.201	0.059	0.375	325	-1.71	0.088*	-1.94*	0.052
2021	0.178	0.014	0.130	0.007	0.280	439	-4.72	0.000***	-3.64***	0.000
2022	0.179	0.015	0.146	-0.022	0.299	447	-4.58	0.000***	0.06	0.951
2023	0.286	0.016	0.240	0.080	0.402	417	-0.63	0.531	4.85***	0.000
2024	0.370	0.013	0.331	0.175	0.502	382	2.50	0.012**	3.96***	0.000

Notes: $\hat{\alpha}_{ct}$ is the S5 pooled-moment county-year frictional markup, winsorized at the 1st and 99th percentiles. For each county-year, the two structural FOC moments are applied to the pooled county-day gap distribution, giving one estimate from ∼500–2,000 facility-days. Distribution statistics are unweighted across county-years. SE is the bootstrap standard error of the annual weighted mean (500 resamples with replacement). Tests vs. 2018 and year-on-year use two-sample z-tests on annual means. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: **Agency share and employee supply volatility.**

	(1) Bivariate	(2) + Controls, Year FE	(3) County FE
Panel A: Observed agency share			
CV employee hours ($\sigma_H/\bar{H}$)	0.249***	0.253***	0.270***
	(0.030)	(0.032)	(0.038)
CV census ($\sigma_N/\bar{N}$)		0.084***	0.106***
		(0.021)	(0.018)
Log(mean census)		0.026***	0.025***
		(0.0040)	(0.0032)
N	40,728	40,728	40,728
R^2	0.179	0.214	0.356
Panel B: Optimal agency share			
CV employee hours ($\sigma_H/\bar{H}$)	0.120***	0.091***	0.097***
	(0.027)	(0.024)	(0.028)
CV census ($\sigma_N/\bar{N}$)		0.266***	0.182***
		(0.039)	(0.031)
Log(mean census)		−0.035***	−0.040***
		(0.0034)	(0.0032)
N	40,728	40,728	40,728
R^2	0.019	0.056	0.216
Year FE	No	Yes	Yes
County FE	No	No	Yes

Notes. Panel A reports OLS regressions of observed RN agency share on employee supply volatility. Panel B uses the newsvendor-optimal agency share as the dependent variable; because optimal agency share is determined mechanically from the empirical gap distribution and wages, the Panel B coefficients represent the model prediction, and the gap between panels measures what the newsvendor framework accounts for. *CV employee hours* is the facility-year coefficient of variation of daily RN employee hours ($\sigma_H/\bar{H}$). *CV census* is the analogous measure for daily resident census. *Log(mean census)* is the log of mean daily census. Standard errors clustered by county in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table 5: County Agency Share and Frictional Markup

	(1)	(2)	(3)	(4)
County agency share	-0.4571^{***}	-0.4600^{***}	-0.4179^{***}	-0.4553^{***}
	(0.0969)	(0.0928)	(0.0990)	(0.0919)
log(county RN hours)		0.0423^{***}	0.0495^{***}	-0.0250
		(0.0139)	(0.0137)	(0.0152)
Year FE	No	No	Yes	Yes
County FE	No	No	No	Yes
N	2,851	2,851	2,851	2,851
Counties	924	924	924	924
R^2	0.019	0.054	0.144	0.024

Notes: Dependent variable is the county-year S5 pooled-moment frictional markup $\hat{\alpha}_{ct}$, estimated by applying the two structural FOC moments directly to the pooled county-day gap distribution (equations 18–19). All specifications estimated by WLS with weights equal to county-year total employee RN hours. Standard errors clustered on county in parentheses. County agency share is the fraction of total annual RN hours supplied by agency staff in the county-year. log(county RN hours) is the log of total annual county RN hours. Specification (4) absorbs county and year fixed effects via iterative demeaning; R^2 is overall. A negative coefficient on county agency share indicates that thicker spot markets have lower frictional markups, consistent with the market-thickening externality A^* in equation (13). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: Participation Externality per Agency Nurse Engagement

Year	$\hat{\alpha}_{ct}$	P(ag)	$\partial\gamma/\partial s$	$E[Q_H]$ \$/cty-day	A^* (id.) \$/eng	A^* (corr.) \$/eng	Corr. factor n_{all}/n_{id}
2018	0.266	0.527	0.298	\$1,117	\$377	\$896	2.61
2019	0.269	0.559	0.296	\$1,261	\$427	\$961	2.53
2020	0.228	0.538	0.310	\$1,497	\$697	\$1,533	3.38
2021	0.175	0.581	0.331	\$1,827	\$1,074	\$1,920	2.64
2022	0.183	0.594	0.326	\$2,107	\$1,561	\$2,444	2.12
2023	0.246	0.588	0.303	\$2,383	\$1,449	\$2,350	2.43
2024	0.300	0.606	0.281	\$2,075	\$1,410	\$2,324	2.39
Mean	0.238	0.571	0.306	\$1,752	\$999	\$1,776	2.59

Notes: Unit of observation is county-year. $\hat{\alpha}_{ct}$ is the S5 pooled-moment county-year frictional markup (annual unweighted mean). $\partial\gamma/\partial s = |\hat{\beta}_{\text{FE}}|/(1+\hat{\alpha}_{ct})^2$ where $\hat{\beta}_{\text{FE}} = -0.455$ (county + year FE, $t = -4.96$). $E[Q_H]$ is mean aggregate shadow value of agency deployed on agency days (\$/county-day). A^* (id.) uses identified facilities only; A^* (corr.) scales by the year-specific coverage factor n_{all}/n_{id}. Engagement = 500 hours (13-week travel nurse contract). A^* isolates the market-thickening (γ_H) channel; the utilization channel is measured separately in the pool value test.

Table 7: Pigouvian Subsidy Estimates

Year	A^*_{corr} \$/eng	n_{all} facilities	A^*/n_{all} \$/eng	Uncaptured \$/eng	Pigouvian rate % of $w_a \times h$	Facilities
2018	\$807	24.1	\$121	\$686	2.5%	1296
2019	\$881	23.9	\$129	\$752	2.6%	1448
2020	\$1,455	24.6	\$208	\$1,247	4.0%	1768
2021	\$1,911	23.3	\$268	\$1,642	4.9%	2607
2022	\$2,577	20.4	\$449	\$2,128	5.9%	2435
2023	\$2,501	22.1	\$418	\$2,083	5.8%	2428
2024	\$2,161	20.2	\$474	\$1,687	4.8%	1873
Mean	\$1,756	22.7	\$295	\$1,461	4.4%	1979

Notes: A^*_{corr} is the coverage-corrected theory-consistent participation externality per 500-hour travel nurse engagement from Table 6. n_{all} is the mean number of all SNF facilities per county-year. A^*/n_{all} is facility F's private share of the social gain it creates. By the buffer first-order condition, F's private net gain from switching is zero; it creates A^* and captures only A^*/n_{all}. Uncaptured externality $= (n_{all} - 1)/n_{all} \times A^*$ is the required Pigouvian subsidy per engagement. The Pigouvian rate expresses this as a percentage of the engagement wage $w_a \times 500$ hours.

A Existence of Equilibrium

Lemma 1. *Let $\theta^*_{\min}$ be the smallest cutoff consistent with $L(\theta^*) \geq 0$. If $\theta^{BR}(\theta^*_{\min}) > \theta^*_{\min}$ and $\theta^{BR}(1) < 1$, then at least one equilibrium exists.*

Proof. Define $g_{\text{eq}}(\theta^*) \equiv \theta^{BR}(\theta^*) - \theta^*$. The boundary conditions give $g_{\text{eq}}(\theta^*_{\min}) > 0$ and $g_{\text{eq}}(1) < 0$.

Case $b = 0$. Within each demand regime, $\theta^{BR}(\theta^*)$ is continuous: $L(\theta^*) = N - 1 + F(\theta^*)$ is continuous; $g(L)$ and $h(R_s)$ are continuous by assumption; in tight states $\theta_m(\theta^*) = F^{-1}\big(F(\theta^*)(1 - 1/\delta_s) + (1 - N)/\delta_s\big)$ is continuous because F is strictly increasing. When $b = 0$ the stitched map is also continuous at regime boundaries: as $R_H \to 1$ from below, $\theta_m \to 0$ and $w_H = \gamma_H \theta_m \to 0$, matching the slack-side wage $b = 0$. Hence g_{eq} is continuous on $[\theta^*_{\min}, 1]$, and the intermediate value theorem delivers a zero.

Case $b > 0$. Within each regime θ^{BR} is continuous (the same argument as above), but at regime boundaries θ^{BR} has upward jumps of order b (Appendix C). Upward jumps in θ^{BR} are upward jumps in $g_{\text{eq}} = \theta^{BR} - \theta^*$ of the same magnitude. The critical observation is that an upward jump *cannot* convert a sign change from positive to negative: if g_{eq} is positive just to the left of a regime boundary, an upward jump leaves it positive (or more positive) just to the right. Therefore every transition of g_{eq} from positive to negative must occur strictly inside a regime, where θ^{BR} is continuous.

Since $g_{\text{eq}}(\theta^*_{\min}) > 0$ and $g_{\text{eq}}(1) < 0$, there exists at least one regime in which g_{eq} transitions from positive to negative. The intermediate value theorem applies within that regime and delivers a zero, i.e., a fixed point. □

Boundary conditions. The first boundary condition holds when nearly all positions are buffered: the pool is so thin that severe rationing drives the scarcity wage

high, making the idle cost of buffering large relative to the friction cost of spot. The second holds when all positions use spot: even with g and h near 1, the residual quality discount $1 - \gamma_s$ exceeds the cost of paying an idle worker at rate b.

B Derivation of the Pool Participation Externality

Total welfare when N workers participate (the rest earn b outside) is $W_{\text{total}} = W(\theta^*; N) + (N_{\text{total}} - N) \cdot b$. At fixed θ^*:

$$\left.\frac{dW}{dN}\right|_{\theta^*} = \frac{\partial Y}{\partial N} + b\,\frac{\partial u}{\partial N}.$$

The output derivative $\partial Y/\partial N$ has three components.

(i) Match-quality improvement, state H. A unit increase in N raises $L = N - 1 + F(\theta^*)$ at fixed θ^*, improving $\gamma_H = g(L)h(R_H)$ through both the density channel ($g' > 0$) and the congestion-relief channel ($h' > 0$, since $R_H = L/(\delta_H F(\theta^*))$ rises). The resulting increase in expected output from served spot positions in state H is

$$\pi_H \delta_H \frac{\partial \gamma_H}{\partial N}\, Q_H, \quad Q_H = \int_{\theta_m}^{\theta^*} \theta \, dF.$$

(ii) Rationing relief, state H. Differentiate the tight-state clearing condition $\delta_H(F(\theta^*) - F(\theta_m)) = L$ with respect to N at fixed θ^*:

$$-\delta_H f(\theta_m)\,\frac{\partial \theta_m}{\partial N} = 1 \implies \frac{\partial \theta_m}{\partial N} = -\frac{1}{\delta_H f(\theta_m)} < 0.$$

The entering worker lowers the rationing threshold, admitting the marginal position, which produces $\gamma_H \theta_m = w_H$. This contributes $\pi_H w_H$ to $\partial Y/\partial N$.

(iii) Match-quality improvement, state L. By the same reasoning as (i), the contri-

bution is $\pi_L \delta_L (\partial \gamma_L / \partial N)\, Q_L$ where $Q_L = \int_0^{\theta^*} \theta \, dF$.

Non-employment term. From equation (7), $u(\theta^*) = \pi_L(N - 1 + (1 - \delta_L)F(\theta^*))$, so $\partial u / \partial N = \pi_L$ at fixed θ^*. The contribution to welfare is $\pi_L b$.

Summing all four components:

$$\left.\frac{dW}{dN}\right|_{\theta^*} = \underbrace{\pi_H w_H + \pi_L b}_{\text{private return}} + A^*,$$

where A^* is defined in equation (13). The worker's private return from pool participation is $\pi_H w_H + \pi_L b$ and her opportunity cost is b, so the net private gain is $\pi_H(w_H - b) \geq 0$. The social return exceeds the private return by $A^* > 0$, establishing the participation externality.

C Regime Analysis

Regime 1 (both states slack, $R_H \geq 1$). No rationing occurs in either state; all active spot positions hire at b. The best response reduces to

$$\theta^{BR}(\theta^*) = \frac{(1 - \bar{\delta})\, b}{\pi_H \delta_H \big(1 - g(L) h(R_H)\big) + \pi_L \delta_L \big(1 - g(L) h(R_L)\big)}. \tag{26}$$

The numerator is constant in θ^*; dependence enters entirely through the denominator via $g(L(\theta^*))$ and $h(R_s(\theta^*))$. The denominator is strictly decreasing in θ^* (both γ_H and γ_L rise toward 1, shrinking $1 - \gamma_s$), so θ^{BR} is strictly increasing. This regime applies to markets with modest demand fluctuations where aggregate supply always suffices.

Regime 3 (both states tight, $R_L < 1$). Both states feature rationing and scarcity wages. Define $\theta_{m,s}$ by $\delta_s(F(\theta^*) - F(\theta_{m,s})) = L$ for $s \in \{H, L\}$. The scarcity wages

are $w_s = g(L)h(R_s)\theta_{m,s}$, the buffer wage is $w = \pi_H w_H + \pi_L w_L$, and the best response is

$$\theta^{BR}(\theta^*) = \frac{\pi_H(1-\delta_H)\, g\, h(R_H)\, \theta_{m,H} + \pi_L(1-\delta_L)\, g\, h(R_L)\, \theta_{m,L}}{\pi_H \delta_H(1 - g\, h(R_H)) + \pi_L \delta_L(1 - g\, h(R_L))}. \tag{27}$$

The key difference from Regime 2 is the absence of a stable slack-state wage floor: idle costs in both states are endogenous. This regime applies to markets with chronic capacity shortages.

Wage discontinuity at boundaries. At the Regime 2/Regime 1 boundary ($R_H = 1$), $\theta_m = 0$ and $w_H = \gamma_H \theta_m = 0$. For $b > 0$, the tight-state spot wage drops discontinuously from b to 0 as the economy crosses from Regime 2 into Regime 1. This is the upward jump of order b in θ^{BR} referenced in Appendix A. As θ^* increases further into Regime 2 and the pool thins, θ_m rises and w_H eventually exceeds b. For $b = 0$ the wage is continuous at the boundary.

D Best-Response Slope Derivation

We derive the slope $d\theta^{BR}/d\theta^*$ in Regime 2 and give conditions under which $H(\theta^*) > 0$ and the product $H(\theta^*)f(\theta^*)$ is unimodal.

State-specific friction derivatives

Each $\gamma_s = g(L) \cdot h(R_s)$ depends on θ^* through both $L = N - 1 + F(\theta^*)$ and $R_s = L/(\delta_s F(\theta^*))$. Using $dL/d\theta^* = f(\theta^*)$ and equation (3), the chain rule gives

$$\frac{d\gamma_s}{d\theta^*} = f(\theta^*)\, \Gamma_s, \qquad \Gamma_s \;\equiv\; g'(L)\, h(R_s) + g(L)\, h'(R_s) \cdot \frac{1-N}{\delta_s F(\theta^*)^2}. \tag{28}$$

The first term of Γ_s is the *density channel*: more workers populate the characteristic space, lowering the cost of close matches equally in both states. The second term is the *congestion-relief channel*: thicker pools improve R_s, benefiting state s at a rate proportional to $h'(R_s)$ and the demand level $\delta_s F(\theta^*)$. Since $R_H < R_L$ and h is concave, $h'(R_H) \geq h'(R_L)$: congestion relief is weakly stronger in the tight state, where marginal improvements in slack yield larger quality gains.

Marginal-position derivative

Differentiate the tight-state clearing condition $\delta_H(F(\theta^*) - F(\theta_m)) = L(\theta^*)$ with respect to θ^*:

$$\delta_H \left(f(\theta^*) - f(\theta_m) \frac{d\theta_m}{d\theta^*} \right) = f(\theta^*),$$

which gives

$$\frac{d\theta_m}{d\theta^*} = \frac{(\delta_H - 1)\, f(\theta^*)}{\delta_H\, f(\theta_m)}. \tag{29}$$

Since $\delta_H < 1$, this is negative: raising θ^* lowers the rationing threshold. Each position switching from buffer to spot frees one worker (supply $+1$) but adds only $\delta_H < 1$ units of expected demand, so net supply rises by $1 - \delta_H$ per switching position.

Scarcity wage derivative

Since $w_H = \gamma_H \cdot \theta_m$ and both factors depend on θ^*, the product rule and equations (28)–(29) give

$$\begin{aligned} \frac{dw_H}{d\theta^*} &= \frac{d\gamma_H}{d\theta^*} \cdot \theta_m + \gamma_H \cdot \frac{d\theta_m}{d\theta^*} \\ &= f(\theta^*) \left[\Gamma_H\, \theta_m + \gamma_H \cdot \frac{\delta_H - 1}{\delta_H f(\theta_m)} \right]. \end{aligned} \tag{30}$$

The first term (positive) is the quality-improvement channel: a thicker pool improves the match on the marginal position, raising its willingness to pay. The second term (negative) is the rationing-relief channel: a thicker pool lowers θ_m, reducing the marginal hirer's willingness to pay and hence the spot wage.

Assembling $H(\theta^*)$

Write $\theta^{BR} = \mathcal{N}/\mathcal{D}$ where

$$\mathcal{N} = \pi_H(1-\delta_H)w_H + \pi_L(1-\delta_L)b, \qquad \mathcal{D} = \pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L).$$

By the quotient rule, using equations (28) and (30) to factor $f(\theta^*)$ from all derivatives:

$$\frac{d\theta^{BR}}{d\theta^*} = f(\theta^*) \cdot H(\theta^*),$$

where

$$H(\theta^*) = \frac{1}{\mathcal{D}^2}\left[\pi_H(1-\delta_H)\left(\Gamma_H\theta_m + \frac{\gamma_H(\delta_H-1)}{\delta_H f(\theta_m)}\right)\mathcal{D} + \mathcal{N}(\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L)\right], \tag{31}$$

where Γ_s is defined in (28).

Sign of H

The term $\mathcal{N}(\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L)$ is strictly positive: this is the *denominator channel*, through which thicker pools reduce $1-\gamma_s$ in both states, making spot more attractive regardless of wages.

The term involving $\mathcal{D}$ is ambiguous. The quality-improvement component $\Gamma_H\theta_m > 0$ raises w_H and hence $\mathcal{N}$; the rationing-relief component $\gamma_H(\delta_H-1)/(\delta_H f(\theta_m)) < 0$

lowers w_H. Hence $H > 0$ when the denominator channel and quality-improvement channel together outweigh the rationing-relief channel. This holds whenever Γ_H is not too small (density and congestion-relief returns are still operative) and $f(\theta_m)$ is bounded away from zero (the rationing margin is not in the extreme left tail of F).

Unimodality of $H \cdot f$

Multiple equilibria require the slope $H(\theta^*)f(\theta^*)$ to exceed 1 over some interior interval. Since $d\theta^{BR}/d\theta^* = f(\theta^*)H(\theta^*)$, the slope inherits the hump shape of f when H varies slowly relative to f. At the mode μ of f, $f'(\mu) = 0$, so

$$\frac{d}{d\theta^*}[H \cdot f]\Big|_{\mu} = H'(\mu)\, f(\mu),$$

and $H \cdot f$ peaks near μ if $|H'/H| \ll |f'/f|$ away from the mode. This holds under four conditions:

(i) g is mildly concave (g'' small), so g' varies slowly with L;

(ii) h is mildly concave (h'' small), so h' varies slowly with R_s;

(iii) θ_m varies smoothly with L, i.e., $f(\theta_m)$ is bounded away from zero;

(iv) the mode μ lies in the interior of the Regime 2 parameter range.

Under these conditions $H \cdot f$ crosses 1 exactly twice within Regime 2, yielding exactly three equilibria when all crossings lie within that regime.

E Proof of Excess Buffering Result

Proof. Differentiate $W(\theta^*) = Y(\theta^*) + u(\theta^*) \cdot b$ with respect to θ^*. From equation (7), $u'(\theta^*) = \pi_L(1 - \delta_L)f(\theta^*)$. For $Y'(\theta^*)$, apply the Leibniz rule to each integral in (8),

using $d\gamma_s/d\theta^* = f(\theta^*)\Gamma_s$ from (28) and $d\theta_m/d\theta^* = (\delta_H - 1)f(\theta^*)/(\delta_H f(\theta_m))$ from (29):

$$Y'(\theta^*) = f(\theta^*)\Big\{ -\bar{\delta}\,\theta^* + \pi_H\delta_H\Gamma_H Q_H + \pi_H\delta_H\gamma_H\theta^* - \pi_H\delta_H\gamma_H \cdot \frac{(\delta_H - 1)\theta_m}{\delta_H} + \pi_L\delta_L\Gamma_L Q_L + \pi_L\delta_L\gamma_L\theta^*\Big\}, \tag{32}$$

where $Q_s = \int_{\theta_m}^{\theta^*} \theta\, dF$ in state H and $Q_s = \int_0^{\theta^*} \theta\, dF$ in state L is the aggregate productivity of served spot positions in state s, as in (13). Since $w_H = \gamma_H\theta_m$, the fourth term simplifies: $-\pi_H\delta_H\gamma_H(\delta_H - 1)\theta_m/\delta_H = \pi_H(1 - \delta_H)w_H$. Grouping the θ^* terms using $\bar{\delta} = \pi_H\delta_H + \pi_L\delta_L$ and adding $b\,u'(\theta^*) = \pi_L(1 - \delta_L)b \cdot f(\theta^*)$:

$$W'(\theta^*) = f(\theta^*)\Big\{ -\underbrace{\big[\pi_H\delta_H(1 - \gamma_H) + \pi_L\delta_L(1 - \gamma_L)\big]}_{\mathcal{D}}\theta^* + \underbrace{\pi_H(1 - \delta_H)w_H + \pi_L(1 - \delta_L)b}_{\mathcal{N}} + \pi_H\delta_H\,\Gamma_H\,Q_H + \pi_L \tag{33}$$

At any equilibrium, $\theta^* = \theta^{BR}(\theta^*) = \mathcal{N}/\mathcal{D}$ by (11), so $\mathcal{D}\cdot\theta^* = \mathcal{N}$ and the first two groups cancel exactly, leaving:

$$W'(\theta^*) = f(\theta^*)\big[\pi_H\delta_H\,\Gamma_H\,Q_H + \pi_L\delta_L\,\Gamma_L\,Q_L\big]. \tag{34}$$

Each factor is strictly positive: $f(\theta^*) > 0$ on the interior of $[0, 1]$ by the full-support assumption on F; $\Gamma_s > 0$ because both the density channel ($g' > 0$) and congestion-relief channel ($h' > 0$) in (28) are positive; and $Q_s > 0$ at any interior equilibrium $\theta^* > 0$. Hence $W'(\theta^*) > 0$ at every equilibrium, so the social optimum satisfies $\theta^{**} > \theta^*$.

The economic content of the cancellation is transparent. The equilibrium condition $\mathcal{D}\cdot\theta^* = \mathcal{N}$ is precisely the statement that the marginal buffering position breaks even: friction costs equal idle costs. The private margin therefore contributes zero

net surplus. What remains in (34) is the market-thickening externality A^* from (13): the match-quality improvement each switching position delivers to every remaining spot user through denser and less congested market conditions. Competitive firms ignore this term entirely. If $g' = h' = 0$ so that no externality exists, then $W'(\theta^*) = 0$ and the proposition fails, as it should. □

F No Stable Equilibrium in Regime 1

Lemma 2. *Suppose $b > 0$. Then the best-response map has no stable fixed point in the interior of Regime 1. Any fixed point in Regime 1 is unstable.*

Proof. In Regime 1, $\theta^{BR}(\theta^*) = \mathcal{N}_1/\mathcal{D}_1$ where $\mathcal{N}_1 = (1-\bar{\delta})b > 0$ is constant and $\mathcal{D}_1(\theta^*) = \pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L) > 0$ is strictly decreasing in θ^* (both γ_H and γ_L are strictly increasing, so their weighted sum $\mathcal{D}_1$ is strictly decreasing). Therefore θ^{BR} is strictly increasing.

Consider any fixed point θ_0^* in Regime 1, so $\theta^{BR}(\theta_0^*) = \theta_0^*$, i.e., $(1-\bar{\delta})b = \theta_0^*\,\mathcal{D}_1(\theta_0^*)$. The slope of the best response at θ_0^* is

$$\left.\frac{d\theta^{BR}}{d\theta^*}\right|_{\theta_0^*} = -\frac{\mathcal{N}_1\,\mathcal{D}_1'(\theta_0^*)}{\mathcal{D}_1(\theta_0^*)^2} = -\frac{\theta_0^*\,\mathcal{D}_1'(\theta_0^*)}{\mathcal{D}_1(\theta_0^*)}.$$

Since $\mathcal{D}_1' < 0$, this slope is strictly positive. Using equation (28):

$$-\mathcal{D}_1'(\theta_0^*) = f(\theta_0^*)\big[\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L\big]\big|_{\theta_0^*}.$$

Substituting and using the fixed-point condition $\mathcal{D}_1(\theta_0^*) = (1-\bar{\delta})b/\theta_0^*$:

$$\left.\frac{d\theta^{BR}}{d\theta^*}\right|_{\theta_0^*} = \frac{f(\theta_0^*)\,(\theta_0^*)^2\big[\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L\big]}{(1-\bar{\delta})b}. \tag{35}$$

For stability we need this slope to be less than 1, i.e.,

$$f(\theta_0^*)(\theta_0^*)^2\big[\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L\big] < (1-\bar{\delta})b.$$

Substituting the fixed-point condition $(1-\bar{\delta})b = \theta_0^*[\pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L)]$, this becomes

$$f(\theta_0^*)\,\theta_0^*\big[\pi_H\delta_H\,\Gamma_H + \pi_L\delta_L\,\Gamma_L\big] < \pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L).$$

In Regime 1, as θ^* increases, $\gamma_s \to 1$ so the right-hand side falls monotonically toward zero, while the left-hand side does not—f and Γ_s remain bounded away from zero on the interior of Regime 1 under conditions (i)–(iv) of Appendix D. The inequality is therefore violated at any interior fixed point of Regime 1, so every such fixed point is unstable. □

The economic content is that without a scarcity wage there is no checking force. In Regime 2, as the pool thickens, $w_H = \gamma_H\theta_m \to 0$ and $\mathcal{N}$ shrinks, pulling θ^{BR} back below the 45-degree line; this is the force that pins the thick-market equilibrium to the regime boundary. In Regime 1 the numerator is the flat value $(1-\bar{\delta})b$ with no analogous mechanism, so the best response rises without bound.

G Proof of Thin-Market Comparative Statics

Proof. From equation (30), $dw_H/d\theta^* > 0$ if and only if the quality-improvement channel dominates the rationing-relief channel:

$$\Gamma_H(\theta^*)\,\theta_m(\theta^*) > \frac{\gamma_H(\theta^*)(1-\delta_H)}{\delta_H\,f(\theta_m(\theta^*))}. \tag{36}$$

At the lower boundary of Regime 2, rationing is severe, θ_m is bounded away from zero, and both channels of Γ_H in (28) are positive, so (36) holds strictly. Near the upper Regime 2 boundary, $\theta_m \to 0$ so the left side vanishes while the right side remains bounded away from zero; (36) fails. By continuity there is a nonempty open interval at the base of Regime 2 on which (36) holds. For parameters supporting multiple equilibria, the lower boundary of Regime 2 lies below the mode μ—otherwise the best-response slope $f(\theta^*)H(\theta^*)$ could not exceed one near μ as required (Appendix D)—so $\underline{\theta}$ can be taken as the right endpoint of the largest sub-interval of $\{\theta^* : (36) \text{ holds}\}$ that lies below μ.

For $\theta^* < \underline{\theta}$, $dw_H/d\theta^* > 0$. The demand frequency result (Appendix H) implies that a shock satisfying condition (37) raises θ^*, so w_H rises. The buffer wage $w = \pi_H w_H + \pi_L b$ rises as well since b is fixed, establishing (iv). The spot premium $w_H - w = \pi_L(w_H - b)$ strictly increases because w_H rises and b is fixed, widening wage dispersion across employment modes, establishing (v). □

H Proof of Demand Frequency Result: Complete Algebra

We provide the complete cross-multiplication step for the demand frequency comparative static in Section 3.

Proof. Write $\theta^{BR} = \mathcal{N}/\mathcal{D}$ with $\mathcal{N} = \pi_H(1-\delta_H)w_H + \pi_L(1-\delta_L)b$ and $\mathcal{D} = \pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L)$. At fixed θ^*, an increase in π_H does not affect w_H, γ_H, or γ_L, so using $\pi_L = 1 - \pi_H$:

$$\frac{\partial \mathcal{N}}{\partial \pi_H} = (1-\delta_H)w_H - (1-\delta_L)b, \qquad \frac{\partial \mathcal{D}}{\partial \pi_H} = \delta_H(1-\gamma_H) - \delta_L(1-\gamma_L).$$

Since $\mathcal{N} > 0$ and $\mathcal{D} > 0$:

$$\frac{\partial \theta^{BR}}{\partial \pi_H} > 0 \iff \frac{\partial \mathcal{N}}{\partial \pi_H} \cdot \mathcal{D} > \frac{\partial \mathcal{D}}{\partial \pi_H} \cdot \mathcal{N}.$$

Expanding both sides:

$$\text{LHS} = \big[(1-\delta_H)w_H - (1-\delta_L)b\big]\big[\pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L)\big],$$
$$\text{RHS} = \big[\delta_H(1-\gamma_H) - \delta_L(1-\gamma_L)\big]\big[\pi_H(1-\delta_H)w_H + \pi_L(1-\delta_L)b\big].$$

Expanding each product into four terms and computing LHS − RHS, the $\pi_H \cdot \pi_H$ and $\pi_L \cdot \pi_L$ cross-terms cancel, leaving:

$$\begin{aligned}\text{LHS} - \text{RHS} = \; & (1-\delta_H)w_H \cdot \pi_L\delta_L(1-\gamma_L) - (1-\delta_L)b \cdot \pi_H\delta_H(1-\gamma_H) \\ & - \delta_H(1-\gamma_H) \cdot \pi_L(1-\delta_L)b + \delta_L(1-\gamma_L) \cdot \pi_H(1-\delta_H)w_H.\end{aligned}$$

Collecting the w_H terms and the b terms separately, and using $\pi_H + \pi_L = 1$:

$$\begin{aligned}w_H \text{ terms:} \quad & (1-\delta_H)(1-\gamma_L)\, w_H\, [\pi_L\delta_L + \pi_H\delta_L] = \delta_L(1-\delta_H)(1-\gamma_L)\, w_H, \\ b \text{ terms:} \quad & -(1-\delta_L)(1-\gamma_H)\, b\, [\pi_H\delta_H + \pi_L\delta_H] = -\delta_H(1-\delta_L)(1-\gamma_H)\, b.\end{aligned}$$

Hence

$$\text{LHS} - \text{RHS} = \delta_L(1-\delta_H)(1-\gamma_L)\, w_H - \delta_H(1-\delta_L)(1-\gamma_H)\, b,$$

which is positive if and only if

$$\frac{w_H}{b} > \frac{\delta_H(1-\delta_L)(1-\gamma_H)}{\delta_L(1-\delta_H)(1-\gamma_L)}, \tag{37}$$

which is the condition stated in Section 3. □ □

Appendix Figures

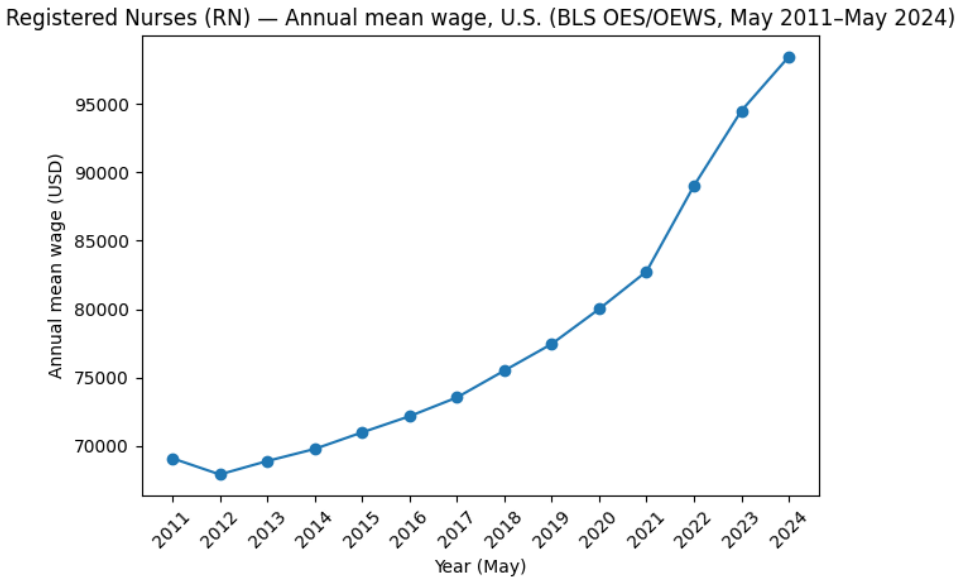


Figure A1: **RN wages, 2011–2024.** Annual mean wage for registered nurses (RN), U.S., BLS Occupational Employment and Wage Statistics (OES/OEWS), May 2011–May 2024.

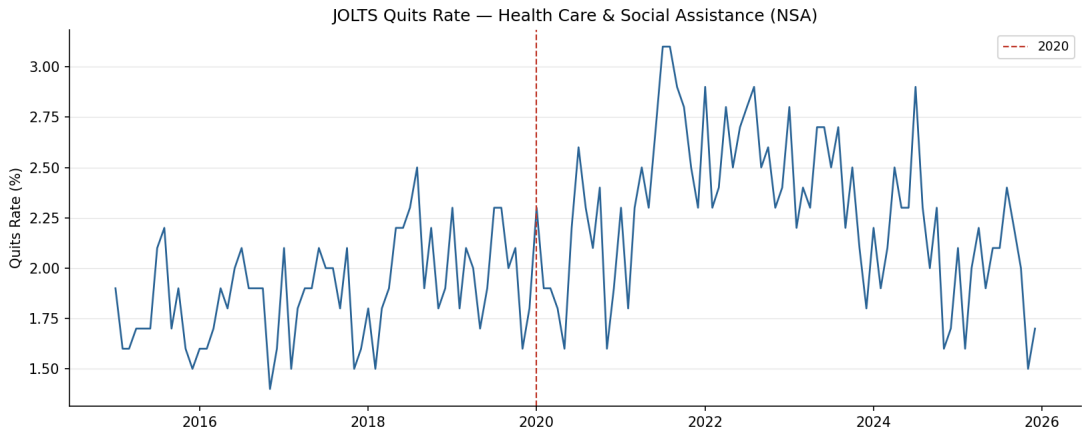


Figure A2: **Quits after 2020.** Monthly quit rate for health care and social assistance, JOLTS (not seasonally adjusted), 2014–2026. Red dashed line marks 2020.

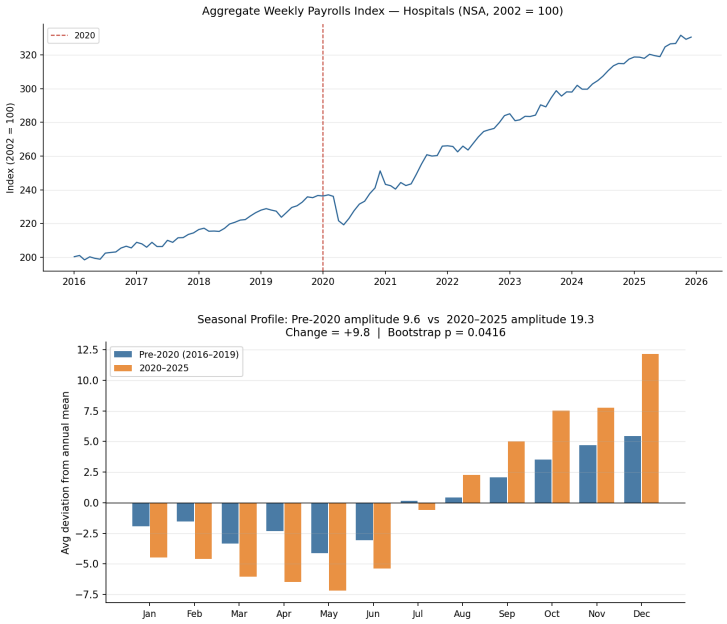


Figure A3: **Hospital employment after 2020.** Top panel: aggregate weekly payroll index, hospitals, BLS CES (not seasonally adjusted, 2002 = 100), 2016–2026. Red dashed line marks 2020. Bottom panel: seasonal profile of the payroll index, pre-2020 (2016–2019) vs. 2020–2025. Pre-2020 amplitude is 9.6; post-2020 amplitude is 19.3 (change = +9.8, bootstrap $p = 0.0416$).

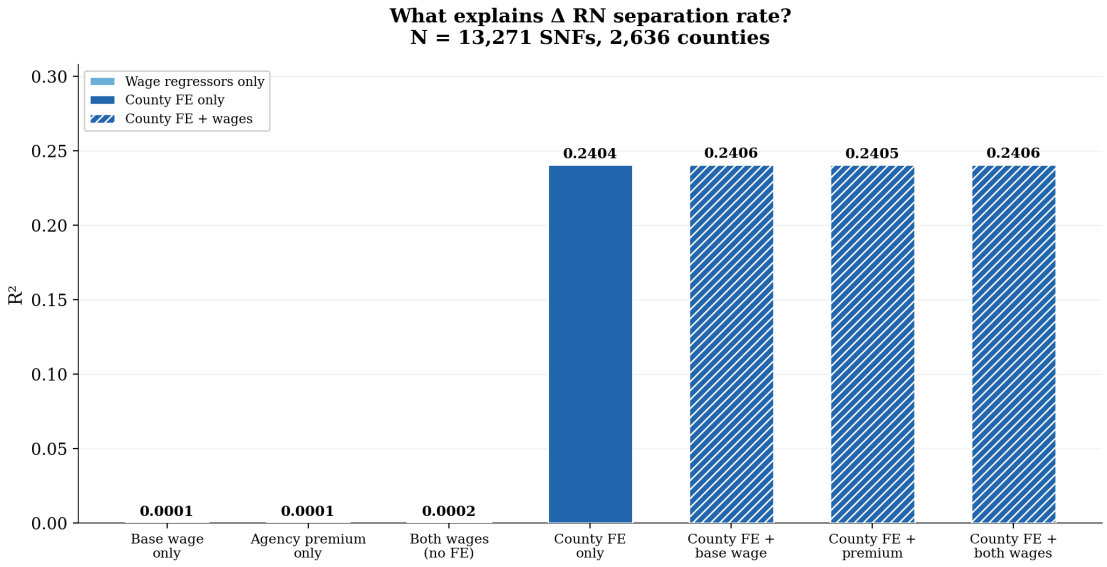


Figure A4: **What explains the change in RN separation rates?** R^2 from OLS regressions of the 2020–2022 facility-level change in RN separation rate on alternative regressor sets: base wage only, agency premium only, both wages without fixed effects, county fixed effects only, county fixed effects plus base wage, county fixed effects plus agency premium, and county fixed effects plus both wages. Sample: $N = 13{,}271$ SNFs across 2,636 counties.

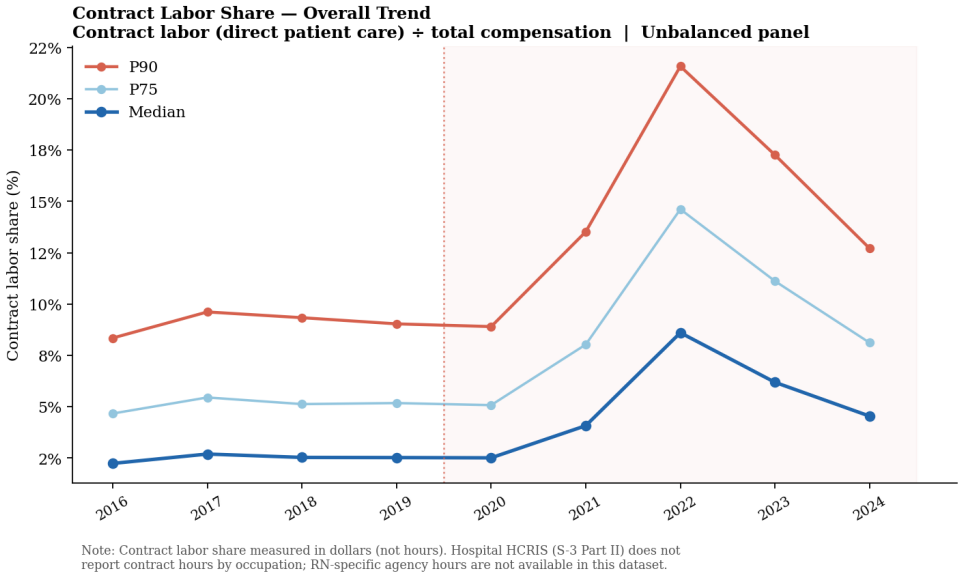


Figure A5: **Contract labor share, hospitals.** Contract labor expenditure (direct patient care) as a fraction of total compensation, HCRIS Worksheet S-3 Part II, unbalanced panel of hospitals, 2016–2024. Lines show the P90, P75, and median of the cross-facility distribution by year. Contract labor share is measured in dollars (not hours); RN-specific agency hours are not available in this dataset. Shaded region indicates post-2020 period.

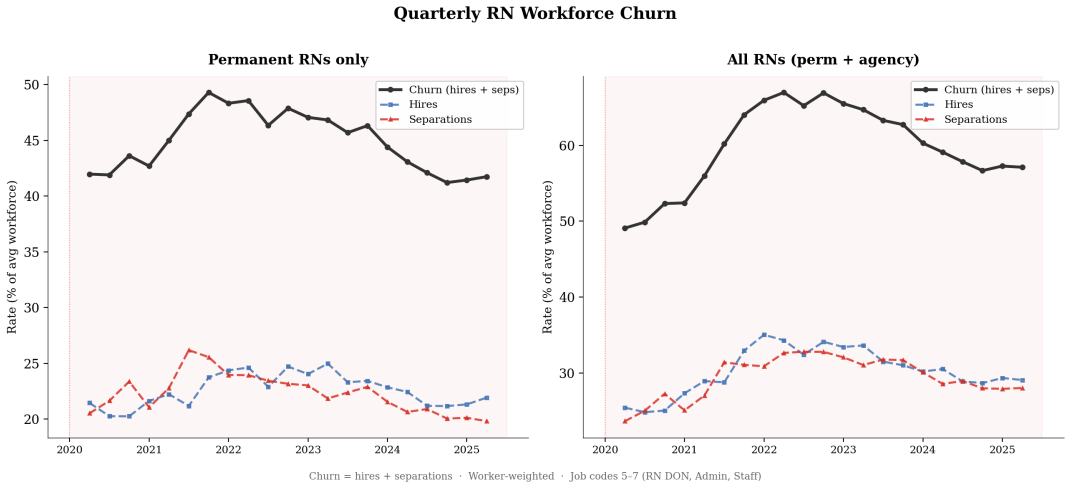


Figure A6: **Quarterly RN workforce churn, 2020–2025.** Hires, separations, and total churn (hires plus separations) as a share of average quarterly workforce, computed from individual-level PBJ employment records using worker identifiers across quarters, PBJ job codes 5–7 (RN DON, Admin, Staff). Worker-weighted. Left panel: permanent (employee) RNs only. Right panel: all RNs including agency workers. Shaded region indicates post-2020 period.

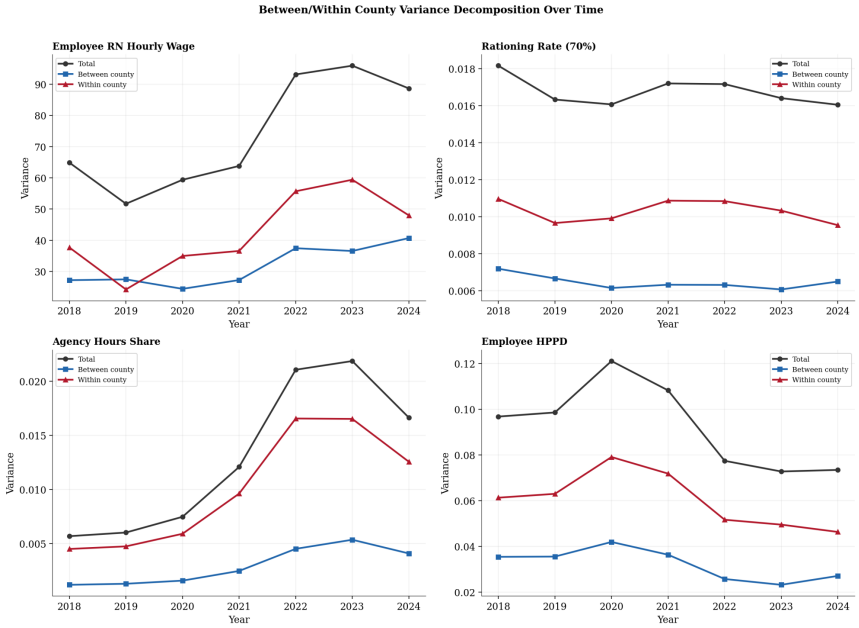


Figure A7: **Between/within county variance decomposition over time.** Total, between-county, and within-county variance by year for four outcomes: employee RN hourly wage, rationing rate (70% threshold), agency hours share, and employee hours per patient-day (HPPD). Computed from facility-year observations in the PBJ/HCRIS merged sample, 2018–2024.

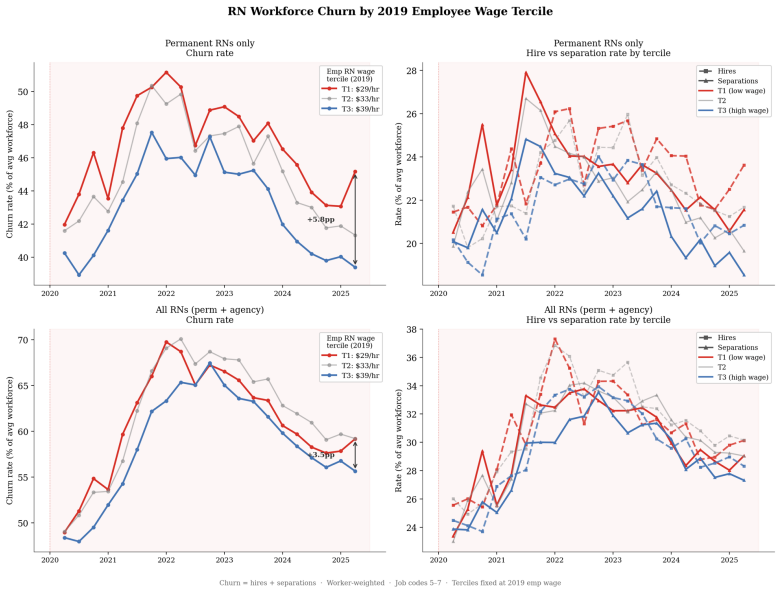


Figure A8: **RN workforce churn by 2019 employee wage tercile, 2020–2025.** Churn rate (hires plus separations as a share of average quarterly workforce) and hire/separation rates by pre-pandemic employee RN wage tercile (T1: $29/hr, T2: $33/hr, T3: $39/hr), PBJ job codes 5–7. Top row: permanent (employee) RNs only; bottom row: all RNs including agency workers. Left column: churn rate by tercile. Right column: hire and separation rates by tercile (solid = hires, dashed = separations). Worker-weighted; terciles fixed at 2019 employee wage. Shaded region indicates post-2020 period.

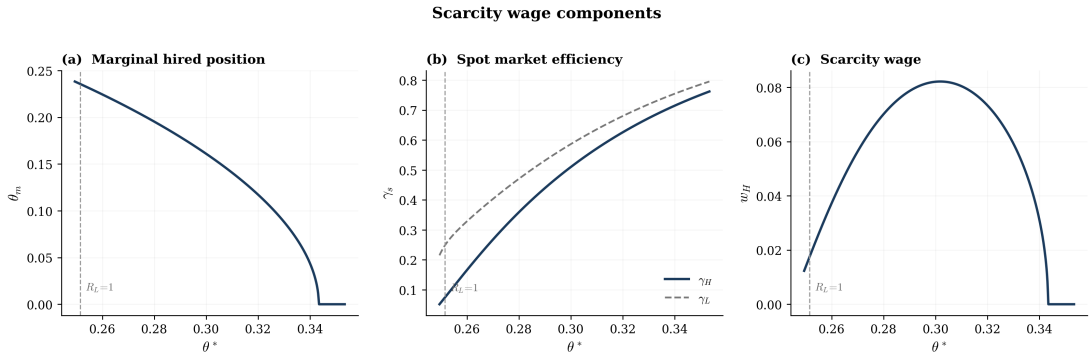


Figure A9: **Scarcity wage components.** Panel (a): marginal hired position θ_m as a function of θ^*. Panel (b): spot market efficiency $\gamma_s = g(L)h(R_s)$ for states H and L. Panel (c): scarcity wage w_H as a function of θ^*. Dashed vertical line marks the Regime 2/1 boundary ($R_L = 1$). Parameters as in Figure 3.

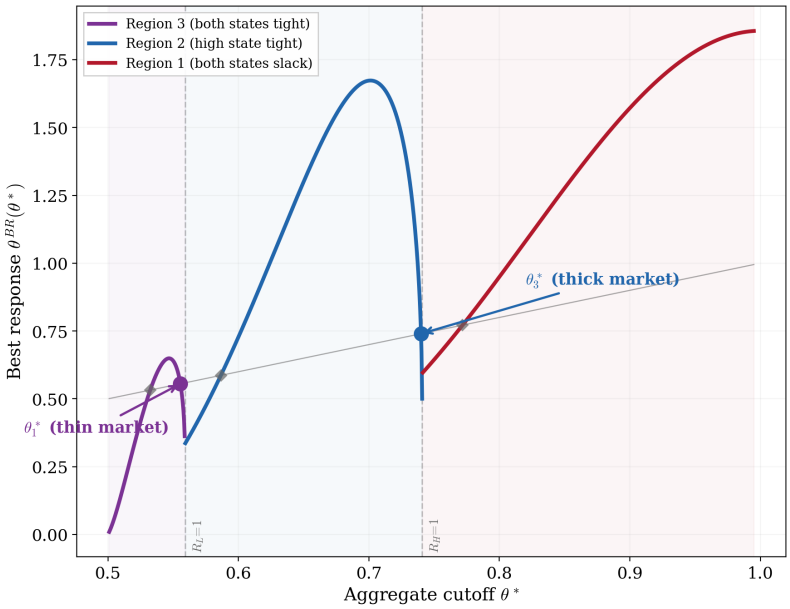


Figure A10: **Multiple equilibria across regimes.** Best-response curve $\theta^{BR}(\theta^*)$ plotted across all three regimes: Region 3 (both states tight, purple), Region 2 (high state tight, blue), and Region 1 (both states slack, red). Shaded backgrounds indicate regime regions; dashed vertical lines mark the regime boundaries $R_L = 1$ and $R_H = 1$. Filled circles: stable equilibria (θ_1^*, θ_3^*); open circle: unstable equilibrium. Parameters: $F = \text{Beta}(2, 2)$, $N = 0.50$, $\delta_H = 0.40$, $\delta_L = 0.15$, $\pi_H = 0.30$, $b = 0.01$.

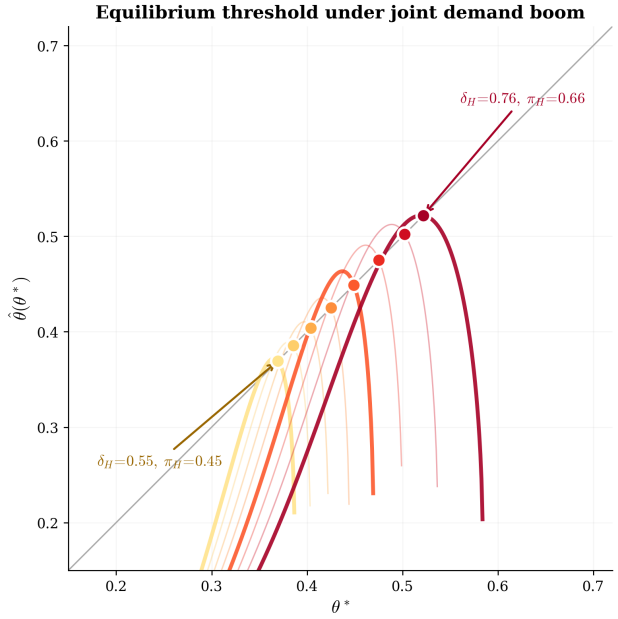


Figure A11: **Equilibrium threshold under joint demand boom.** Each curve plots $\theta^{BR}(\theta^*)$ for (δ_H, π_H) increasing jointly from $(0.55, 0.45)$ (yellow) to $(0.76, 0.66)$ (dark red), holding all other parameters at their baseline values. Filled dots mark stable equilibria.

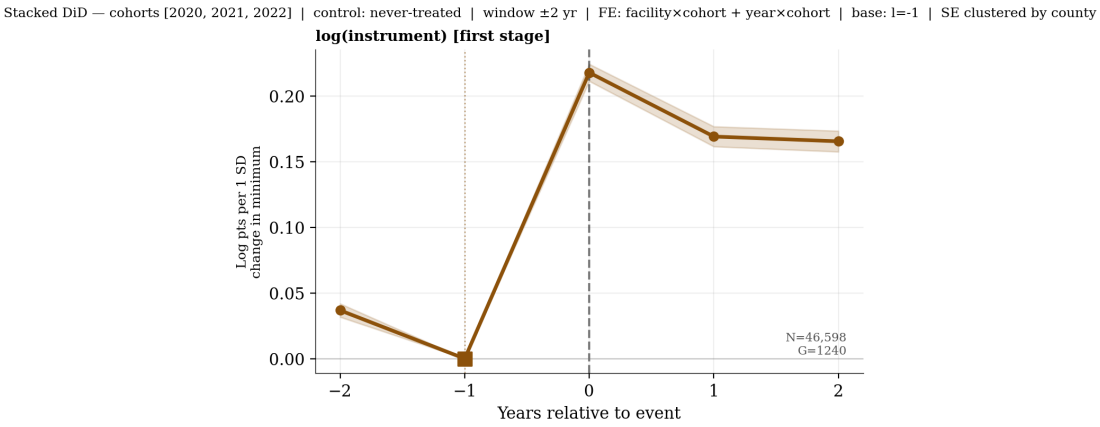


Figure A12: **First stage: stacked difference-in-differences.** Log instrument (acuity-adjusted minimum) response to shifts in h^*_{jt}, same stacked DiD design as Figure 11. Cohorts defined by facilities experiencing a shift in h^*_{jt} in 2020, 2021, and 2022; control group is never-treated facilities. Fixed effects: facility-by-cohort and year-by-cohort. Standard errors clustered by county; shaded bands show 95% confidence intervals. Base period: $\ell = -1$. $N = 46{,}598$ facility-cohort-year observations; $G = 1{,}240$ treated cohort-facility groups.