NBER WORKING PAPER SERIES

THE LABOR MARKET AS AN EQUILIBRIUM NEWSVENDOR PROBLEM

Alexandre Mas

Working Paper 35148
http://www.nber.org/papers/w35148

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2026, Revised June 2026

I am grateful to Pat Kline, Enrico Moretti, and Carl Shapiro for helpful discussions, and to seminar participants at the Nebraska Labor Summit and Tufts. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Labor Market as an Equilibrium Newsvendor Problem
Alexandre Mas
NBER Working Paper No. 35148
April 2026, Revised June 2026
JEL No. J0, J2, J3

ABSTRACT

I study hiring 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 endogenously determined 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 the competitive equilibrium involves too much buffering. I evaluate the model in the market for registered nurses using the universe of daily time sheet records from skilled nursing facilities. The data support the model's general equilibrium predictions. Markets with thicker agency markets have less rationing conditional on total nursing hours, and compressed upper tails of permanent staffing with higher overall hours in the lower part of the distribution. Estimating a facility-level newsvendor model with heterogeneous spot market frictions and rationing costs I find that the nursing labor market was more efficient post-COVID, there are substantial welfare gains from bringing frictions down to those of the most efficient markets, and that the marginal external benefit for a firm that participates in the spot market instead of hiring an employee as buffer averages 12 percent of wages. This labor market underprovides flexibility because the value of availability is not fully priced.

Alexandre Mas
University of California, Berkeley
Haas School of Business
and NBER
amas@berkeley.edu

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.[1] This problem is similar to the one in the canonical newsvendor model for inventory decisions 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. 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 can be 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 when there are thick market externalities. A firm that relies more on spot hiring thickens the market others use but does not internalize that contribution, so competitive equilibrium may involve too much private buffer capacity and too little spot-market participation. Demand shocks can in principle improve market functioning indirectly by shifting firms toward the spot market.

The model has four basic elements: heterogeneous firm productivity, aggregate demand uncertainty, different employment modes, and endogenous spot market frictions. Prior to observing demand 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. From these primitives the model produces equilibrium

[1] Adjusting for workload fluctuations is the primary reason employers use flexible staffing arrangements (Houseman, 2001).

predictions for labor market outcomes.

I study these forces in the market for registered nurses at skilled nursing facilities. This is a useful setting empirically because Centers for Medicare and Medicaid Services (CMS) Payroll-Based Journal (PBJ) data record daily hours separately for employee and agency nurses, so the buffer and spot margins are both directly observed. Additionally, there is 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 understanding the implications of this theory. Health care workers experienced pronounced wage growth over this period accompanied by high quits and separation rates. As I document, nurse wage growth was almost entirely within facilities rather than reallocation towards higher paying facilities, and separation increases were symmetric across the facility wage distribution and largely a market-level phenomenon and not explained by employer characteristics. 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 from PBJ I present evidence supporting the general equilibrium predictions of the model. Markets with thicker pre-period agency markets have lower rationing conditional on total nursing hours and a compressed upper tail of permanent staffing with better overall coverage throughout the distribution.

Finally, I estimate the structural parameters of the newsvendor model 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. Together with the observed mean residual shortfall on agency-using days, these conditions jointly identify the agency frictional markup and heterogeneous demand.

I find that agency frictions fell after 2020 and producer surplus and marginal rationing costs rose. Using the model, I estimate that if spot market frictions were capped at the 25th percentile market, there would be a 45 percent reduction in expenditure on nurses hired for precautionary purposes, and a 15 percent reduction in the value of rationed care. There is a strong negative relationship between market agency share and estimated spot market frictions, and the implied market-thickening marginal external benefit averages 12 percent of the spot market wage.

This paper contributes to several literatures. The mechanism is related to thick-market search externalities (Diamond, 1982; Mortensen, 1982), strategic complementarity (Howitt and McAfee, 1992), and agglomeration (Helsley and Strange, 1990; Bleakley and Lin, 2012; Moretti and Yi, 2024), where market size is determined by firms' buffer-spot choices. This trade-off is related to labor hoarding (Oi, 1962; Fay and Medoff, 1985), state-independent wages (Baily, 1974; Azariadis, 1975), and use of contract and outsourced labor in place of direct employment (Goldschmidt and Schmieder, 2017). This paper is a firm-side complement to Chen et al. (2019) who document large worker valuations of real-time scheduling flexibility in ride-share.

The approach also has an analogue to financial markets. Caballero and Krishnamurthy (2008), for example, show that when intermediaries hoard liquidity to self-insure, the resulting withdrawal of liquidity from the market raises the cost of intermediation, so prudent precaution is collectively excessive. I show that a similar dynamic applies in real factor markets. Permanent staffing provides firms with private insurance against demand shocks and withdraws workers from the spot market affecting market liquidity. Because the efficiency depends on participation, this self-insurance reduces market depth and raises hiring frictions for other firms. In labor markets, this liquidity effect appears in wages, employment levels, and rationing.

2 Model

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) spot market frictions that depend on market size and density. While I study nursing, the basic mechanism in the paper applies to any market where firms choose between private buffer capacity and where these conditions apply. An example outside of labor markets is cloud computing, where firms face the choice between reserved instances, committed in advance, and spot instances drawn from a shared pool after demand is realized. The latter carries spot market friction via preemption risk.

2.1 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 satisfying $f(\theta) > 0$ for $\theta \in (0, 1)$: there exists $\mu \in (0, 1)$ such that $f' > 0$ for $\theta < \mu$ and $f' < 0$ for $\theta > \mu$.[2] A staffed position produces θ when active.[3]

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 of a job is $\bar{\delta} = \pi_H\delta_H + \pi_L\delta_L$. The interpretation of a state can depend on the context. It can be a business cycle fluctuation, seasonal demand, or high-frequency idiosyncratic shocks.

[2] Single-peaked density is not necessary for the main results of the model but helps for tractability.

[3] 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.

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 hired.

2.2 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.

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.[4]

2.3 Spot Market Friction

A spot position filled in state s produces $\gamma_s \cdot \theta$ for $\gamma_s \leq 1$. The term γ_s represents the cost of spot market frictions and depends on market density and congestion.[5] Let $g(L)$ be match quality a firm achieves as sole employer drawing from L workers. This term is a matching externality akin to Helsley and Strange (1990): in a denser pool, workers are closer on average in the characteristic space and match quality rises. Let $h(R_s)$ be the fraction of that potential realized, where $R_s = L/E_s$ is the slack-ratio, and E_s is the mass of active spot positions. This term represents congestion and corresponds to tightness in DMP models (Diamond, 1982; Mortensen, 1982; Pissarides, 1985).

[4] There is considerable evidence that workers have preferences for stability of hours, see eg. Mas and Pallais (2017) and Lachowska et al. (2026). 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.

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

When many positions hire simultaneously, close workers are bid away and realized match quality falls. I parameterize γ_s as $\gamma_s \equiv g(L) \cdot h(R_s)$, where $g' > 0$, $g'' \leq 0$, $g(0) = g_0 > 0$, $\lim_{L\to\infty} g(L) = 1$; $h' > 0$, $h'' \leq 0$, $\lim_{R\to\infty} h(R) = 1$. 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.

2.4 Spot Market Clearing

There is a cutoff equilibrium in which firms with positions with $\theta \geq \theta^*$ are buffered and positions with $\theta < \theta^*$ use spot. The output gain from avoiding spot-market friction is $(1-\gamma_s)\theta$, which is increasing in θ, while the idle-wage cost of keeping a permanent worker is independent of θ. Thus, the most productive positions buffer and lower-productivity positions use the spot market. The cutoff θ^* is the marginal position that is indifferent between the two employment modes. Buffered positions have mass $1 - F(\theta^*)$, each assigned one permanent worker. The remaining workers enter the spot market:

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

The derivative $f(\theta^*)$ governs how responsive the spot market 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 spot market wage responds sluggishly. The slack ratio responds to the threshold value as:

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

which is 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 as more firms enter the spot market.

To work through the implications parsimoniously, I focus on an interior equilibrium in which high-demand spot labor is scarce, low-demand spot labor is slack, the high-state wage is above the worker outside option, and some low-state spot hiring occurs.[6] Appendix A states the conditions for this equilibrium, and all subsequent equilibrium and welfare statements below correspond to this region of parameter values.

In the slack state, active spot positions hire at the outside option: $w_L = b$. Because non-hired spot workers take the outside option, every spot-market participant receives payoff b in the slack state, either as a wage or as outside-option value. In the tight state, $E_s = \delta_H F(\theta^*)$ 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:

$$\delta_H \big[F(\theta^*) - F(\theta_m) \big] = L(\theta^*). \tag{3}$$

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

In the slack state, a firm with a position with productivity θ earns $\gamma_L \theta - b$ from hiring and chooses to hire only when this is non-negative; the break-even productivity is $\theta^L_{\min}(\theta^*) \equiv b/\gamma_L(\theta^*)$. Firms with active spot positions with $\theta < \theta^L_{\min}$ do not hire; their would-be workers remain unemployed at the outside option.

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

$$w_H = \gamma_H \cdot \theta_m(\theta^*) = g(L)\, h(R_H) \cdot \theta_m(\theta^*). \tag{4}$$

[6] Two other configurations are possible: one in which both states are slack ($R_H \geq 1$ and $R_L \geq 1$), so spot wages equal the outside option b in both states; and one in which both states are tight ($R_H < 1$ and $R_L < 1$), so both states feature rationing and scarcity wages. Figure A2 illustrates both regimes.

Substituting $L(\theta^*)$ and solving (3) for θ_m:

$$w_H(\theta^*) = g(N-1+F(\theta^*)) \cdot h\left(\frac{N-1+F(\theta^*)}{\delta_H F(\theta^*)}\right) \cdot F^{-1}\left(\frac{(\delta_H-1)F(\theta^*)+1-N}{\delta_H}\right). \tag{5}$$

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 A1 illustrates each channel separately and the combined non-monotone relationship between scarcity wage and the threshold firm.

Non-employment. In this regime, there is full employment in the high state. Expected non-employment is the slack-state gap between workers in the spot market and spot hires weighted by the slack-state probability. There are fewer hires than spot workers because only a fraction δ_L of spot positions activate, and among those, only ones with $\theta \geq \theta^L_{\min}$ find it profitable to hire:

$$u(\theta^*) = \pi_L\big[N-1+(1-\delta_L)F(\theta^*)+\delta_L F(\theta^L_{\min}(\theta^*))\big]. \tag{6}$$

Higher θ^* has two effects on employment that run counter to each other: more workers are in the spot market and so more are unhired in slack states (less hoarding), but the larger market size also raises matching efficiency, which lowers the break-even threshold $\theta^L_{\min}$ and more low-productivity positions are filled in the slack state.

Output. Expected aggregate output is:

$$Y(\theta^*) = \bar{\delta}\int_{\theta^*}^{1}\theta\,dF + \pi_H\delta_H\gamma_H\int_{\theta_m(\theta^*)}^{\theta^*}\theta\,dF + \pi_L\delta_L\gamma_L\int_{\theta^L_{\min}(\theta^*)}^{\theta^*}\theta\,dF. \tag{7}$$

Higher θ^* thus has two opposing effects on output: shifting one position from buffer to spot forgoes its expected guaranteed contribution $\bar{\delta}\,\theta^*$, while the thicker spot market raises matching efficiency γ_s, relieves rationing in the tight state by lowering θ_m, and activates more positions in the slack state by lowering $\theta^L_{\min}$.

2.5 Sorting and Equilibrium

In equilibrium, workers receive equal expected utility across modes, the buffer wage matches expected spot earnings, and no firm can raise profits by switching modes for positions.

Worker indifference requires the buffer wage equals expected spot earnings:

$$w = \pi_H w_H + \pi_L b. \tag{8}$$

A buffered marginal position earns expected profit $\Pi^B \equiv \bar{\delta}\,\theta^* - w$, while a marginal spot position earns $\Pi^S \equiv \pi_H \delta_H(\gamma_H \theta^* - w_H) + \pi_L \delta_L(\gamma_L \theta^* - b)$, where the low-state wage is b. Using worker indifference, and equating $\Pi^B = \Pi^S$ gives

$$\underbrace{\pi_H \delta_H (1-\gamma_H)\theta^* + \pi_L \delta_L (1-\gamma_L)\theta^*}_{\text{friction cost of spot}} = \underbrace{\pi_H (1-\delta_H) w_H + \pi_L (1-\delta_L) b}_{\text{idle cost of buffering}}. \tag{9}$$

Solving (9) for the cutoff:

$$\theta^I(\theta^*) = \frac{\pi_H (1-\delta_H)\, w_H(\theta^*) + \pi_L (1-\delta_L)\, b}{\pi_H \delta_H \big(1-\gamma_H(\theta^*)\big) + \pi_L \delta_L \big(1-\gamma_L(\theta^*)\big)}. \tag{10}$$

A position with $\theta > \theta^I$ buffers; with $\theta < \theta^I$ it uses spot. A competitive equilibrium is a tuple (θ^*, w, w_H, L) satisfying equations (1)–(8) and the fixed point $\theta^* = \theta^I(\theta^*)$. Because L, w_H, and w are all functions of θ^* alone, the fixed point determines all other objects in the equilibrium: the spot market $L(\theta^*)$, market tightness R_s, spot frictions γ_s, the rationing threshold θ_m, the scarcity wage w_H, and the buffer wage w. Figure 1 illustrates the equilibrium for a specific set of parameter values.[7] The equilibrium

[7] Sufficient conditions for an interior fixed point are given in Appendix B.

is determined by the downward-sloping portion of the indifference-threshold curve crossing the 45-degree line.

Multiple equilibria. Strategic complementarity can also generate multiple equilibria. When a position switches from buffer to spot, the market thickens, $g(L)$ and $h(R_s)$ both improve, and spot becomes more attractive to remaining positions, thickening the market further. Figure A2 illustrates multiple-equilibria. The thin-market equilibrium is self-reinforcing because low hiring efficiency makes buffering individually rational, which keeps the market thin, which keeps frictional costs high.

2.6 Excess Buffering

When a position is filled with a stable worker as buffer, it withdraws one worker from the spot market but withdraws only $\bar{\delta} < 1$ units of expected demand. The market thins, worsening match quality 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:

$$W(\theta^*) = Y(\theta^*) + u(\theta^*) \cdot b. \tag{11}$$

The buffering externality B^* arises from the lower frictions that all spot users experience when one position switches from buffer to spot. Differentiating W with respect to θ^*, after the private friction and idle costs cancel at the equilibrium margin, yields

$$B^* \equiv \frac{W'(\theta^*)}{f(\theta^*)} = \pi_H \delta_H \Gamma_H Q_H + \pi_L \delta_L \Gamma_L Q_L > 0, \tag{12}$$

where $\Gamma_s \equiv d\gamma_s/d\theta^*/f(\theta^*)$ is the hiring efficiency gradient described in Appendix C and Q_s is the aggregate productivity of filled spot positions in state s (Appendix D).

Since $B^* > 0$ at any interior equilibrium with positive spot activity, a marginal reduction in buffering at the competitive equilibrium raises welfare.[8]

3 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.[9] While the spot market need not consist exclusively of agency workers—firms can also hire and cycle through regular employees on a just-in-time basis—agency hiring is the natural empirical parallel to the model's spot margin. It is directly and consistently observed in the PBJ data, contractually distinct from permanent employment, and constitutes a well-defined market whose participants move fluidly across facilities. I focus on demand for Registered Nurses (RNs).[10]

3.1 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

[8] By the same logic there is a labor participation externality: when a net new worker enters the market from outside they bring no demand of their own, so the full congestion-relief benefit accrues to other hirers and the per-worker externality strictly exceeds B^*.

[9] 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.

[10] The model's assumption of a fixed worker mass N in the short run also fits the RN labor market as there are certification barriers to entry. Agency RNs are not a distinct labor pool from regular employees as staffing firms generally require two years of bedside experience before placement (Gaines, 2025; Health Carousel Travel Nursing, n.d.), so the agency workforce is drawn from nurses who started as permanent employees, and movement between the two modes is common. Foreign-trained RNs entering on employment-based visas are a partial exception, but the EB-3 visa process typically runs one to three years (Scott D. Pollock and Associates, 2026), so these flows respond on a multi-year horizon. Foreign edicated nurses consitute approximately 5-8 percent of the nursing workforce (Olanrewaju et al., 2022).

two choice variables of the structural estimation. The PBJ spans 2017Q4–2024Q4 , covering 15,933 distinct facilities and and 37.5 million RN facility-day records. The empirical counterpart to the theoretical model's rationing outcome is whether a facility is meeting needs through the RN staffing ratio, measured as hours of RN care per patient day (HPPD).

HCRIS. I use Medicare Cost Reports for freestanding SNFs (Form CMS-2540, Worksheet S-3 Part V) to extract annual facility-level RN compensation, along with contract RN hourly rates, patient census, and facility characteristics.[11] Core-Based Statistical Area (CBSA) -year agency wages w_a are the hours-weighted mean of facility-level contract RN hourly rates inclusive of agency fees within the CBSA-year. Permanent labor cost is total hourly compensation mc_e (salary plus fringe benefits per paid hour) from the same worksheet.

Table 1 reports summary statistics for the main sample of 56,381 facility-years from 2018 to 2024. Panels A and B cover this full sample; Panel C reports structural parameters for the restricted sub-sample of facilities that meet the conditions for the newsvendor model.

3.2 The Post-2020 Nursing Boom

The nursing labor market tightened sharply beginning in 2020. Appendix Figure A3 shows annual mean RN wages rising 30 percent from 2019 to 2024. Appendix Figure A4 shows the hospital aggregate weekly payroll index rising steeply after 2020, and Appendix Figure A5 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 monopsony (Autor et al.,

[11] Hospital-based SNFs, which file under Form CMS-2552 with the host hospital and account for roughly 3 percent of U.S. SNFs, are not included.

2024), which imply that wage growth should be at least partly through reallocation and separation increases should be disproportionately concentrated at lower-wage employers. Table 2 shows a between-within establishment decomposition of RN wage growth relative to 2019 (see Appendix G for details of the construction). For employee RNs (Panel A), the within-facility component accounts for more than 110 percent of total wage growth in every year from 2019 through 2023, while the between-facility reallocation component represents only 3 percent of the cumulative increase by 2023.[12] Stratifying by pre-pandemic wage tercile, churn increases were symmetric across all three wage terciles (Appendix Figure A6), and facility wages explain essentially none of the cross-facility variation in separation changes whether entered alone or with controls (Appendix Figure A7). 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 2 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); Appendix Figure A8 shows that the same pattern is observed in hospitals, where contract labor's share of total compensation rose sharply after 2020 across the distribution of facilities. Appendix Figure A9 confirms that this pattern reflects a broader shift toward spot-market employment. Overall RN workforce churn, defined as hires plus separations as a share of the quarterly workforce, increased sharply after 2020 for both permanent and agency workers, not just within contract staffing. 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 (Appendix Figure A4).

[12] The within components summing to more than 100 percent reflects a modest negative covariance between changes in facility wages and changes in facility employment shares—facilities that raised wages most tended to contract slightly.

In terms of the model, COVID-19 increased both the intensity (δ_H) and frequency (π_H) of the high-demand state. A shock that raises δ_H shifts the equilibrium toward spot when the marginal hire is in a sufficiently thin part of the productivity distribution and δ_H is not too close to 1, so that a buffer worker still faces a meaningful chance of being idle in the H state.[13] A shock that raises π_H has the same qualitative effect provided the scarcity premium is large relative to the outside option:

$$\frac{w_H}{b} > \frac{\delta_H(1-\delta_L)(1-\gamma_H)}{\delta_L(1-\delta_H)(1-\gamma_L)} \tag{13}$$

(Appendix F). When this holds, θ^* rises, spot share increases, separation rates rise, and idle share falls. If the equilibrium starts in the thin-market region, wages rise in both modes and wage dispersion widens (Appendix G). Figure 3 displays the five predictions as (δ_H, π_H) increase jointly for specific parameter values; Appendix Figure A10 shows the underlying shift in the indifference-threshold curve $\theta^I(\theta^*)$ that determines these outcomes.[14]

These predictions match the nursing market over the period: as more facilities turned to agency labor after 2020, the spot market thickened, frictions fell, and the option value of spot participation rose. The buffer wage is set by worker indifference so it equals the expected spot earnings a worker forgoes by taking a permanent position. When the spot market thickens and w_H rises, the buffer wage rises at every facility, which is consistent with the within-facility wage growth in Table 2.

[13] A rise in δ_H activates more H-state positions, raising the marginal hire's productivity θ_m^H and the scarcity wage $w_H = \gamma_H \theta_m^H$; this increase raises the idle cost of buffering and tilts hiring toward spot. The same activation lowers the slack ratio R_H, degrading match quality $h(R_H)$, dragging w_H back down, and raising the friction $1-\gamma_H$ in D. The resulting congestion tilts hiring back toward buffering. Rationing dominates when f is thin at θ_m^H, since the cutoff must travel a long productivity distance per unit of mass. The idle cost enters N weighted by $\pi_H(1-\delta_H)$, which approaches 0 as $\delta_H \to 1$: the precautionary motive disappears and the mechanical shrinkage of the idle-weight dominates. Keeping δ_H bounded below 1 preserves enough idle risk for the rationing channel to move the threshold. See Appendix E for the formal condition and proof.

[14] COVID-19 likely also raised the outside option b (hospital competition for nurses). An increase in b reinforces all five predictions: higher idle costs shift the equilibrium toward spot, while induced market thickening raises w_H.

Separations rise as more workers cycle through spot employment. Figure 4 shows that the agency premium rose after COVID. In the model this is consistent with the high-demand scarcity wage rising as the spot market became more important. Appendix G gives the condition under which the premium widens rather than being offset by the changing state weights.

4 General Equilibrium Evidence

To test the model's general equilibrium prediction I first ask whether thicker spot markets are associated with less rationing. Next, 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 CBSAs 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. To visualize the data, rationing here is measured somewhat crudely as the share of facility-days in the county-week on which total RN hours fall below 70 percent of the facility's own within-quarter mean daily hours, a facility-specific threshold that captures departures from each facility's normal staffing level. This residualization controls for the overall scale of nursing activity in the CBSA, so differences across groups reflect market characteristics rather than total hours worked. High-agency CBSAs exhibit persistently lower rationing throughout the sample: for the same number of total nursing hours in the market, CBSAs with a higher agency share in the baseline period have fewer rationing events. After 2020, the gap widens. The increase in rationing that accompanied the COVID demand shock was more muted in CBSAs that entered the pandemic with a deeper spot market.

Table 3 estimates the same relationship with a regression. To remain consistent with the subsequent analyses, 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 rationing measure used in the structural analysis.[15] The regressor of interest is the leave-own-facility-out CBSA agency share, which measures the thickness of the local spot market while excluding the facility's own hours to avoid mechanical correlation. The most demanding specification adds CBSA fixed effects (column 4), using within-CBSA variation across facilities over time. The estimated coefficient is -0.365 and significant at the five percent level, indicating that a ten-percentage-point increase in CBSA agency share is associated with a 3.7-percentage-point reduction in a facility's rationing rate even after absorbing all time-invariant differences across labor markets. The cross-sectional estimates are larger. The bivariate coefficient in column (1) is -0.790 ($p < 0.01$), essentially unchanged at -0.757 after absorbing CBSA scale with log RN hours in column (2), and rises to -0.963 after adding year fixed effects in column (3).[16] Overall, the estimates support the prediction 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. Although labor markets are defined as CBSAs throughout this paper, the UQR specification uses county-level agency share because RIF variance grows as the outcome distribution thins in the upper tail, and the finer cross-sectional variation in county agency share helps to offset this precision loss.[17] The specification controls for log patient census, for-profit status, and case mix index. The left panel shows employee (permanent) HPPD only. The coefficient on county agency share is

[15]This measure was not used in the time-series figure because the minimums change every year so the rationing measure does not have a harmonized definition over time.

[16]The attenuation under CBSA fixed effects suggests that CBSAs with thicker agency markets may differ from thin-market CBSAs on unobservables, or that there are time-varying shocks correlated both with rationing rate and agency share that are less of a concern in the cross-sectional comparison.

[17]The asymptotic variance of the RIF-OLS estimator at quantile τ is proportional to $\mathrm{Var}(\mathrm{RIF}(y; q_\tau))/\mathrm{Var}(X)$, where $\mathrm{Var}(\mathrm{RIF}) \propto 1/f_Y(q_\tau)^2$. As τ approaches the upper tail, $f_Y(q_\tau)$ shrinks and RIF variance inflates; switching from CBSA to county agency share increases $\mathrm{Var}(X)$ and partially offsets this precision loss.

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 county agency share 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 have fewer permanent workers show no significant decline in total staffing as they substitute idle permanent buffers with flexible spot coverage. The interpretation is that thicker markets raise total hours over most of the distribution, improving coverage for facilities that otherwise fall short, while the high-staffed facilities at the top achieve the same total coverage more efficiently using less permanent staff and more agency labor when needed.

5 Structural Estimation

This section quantifies the magnitude of spot market frictions, the cost of unmet care, and how these quantities have changed over time. I estimate a newsvendor model structurally to 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. Using these estimates I examine how frictions changed across the 2020 demand shock and ask in Section 6 whether the estimated valuations imply systematic over-buffering.

I recover parameters from observed equilibrium choices. The key feature of the setting is the simultaneous observability of 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 with a quadratic cost specification. The gap distribution entering the buffer first-order condition is taken directly from the empirical distribution of daily shortfalls.[18]

5.1 Newsvendor Setup

Each SNF chooses permanent RN hours H (the buffer) to minimize expected daily staffing cost. Taking a standard inventory newsvendor formulation mapped to labor, permanent nurses cost mc_e per hour and are paid whether or not they are needed; agency nurses cost $w_a > mc_e$ 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:

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

The first term is the cost of the permanent workforce; 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:

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

The textbook solution to the newsvendor problem is to set the optimal buffer at the $(1 - mc_e/w_a)$ quantile of the demand distribution.[19] Because we observe the gap empirically for every facility, the optimal buffer can be calculated.

[18] Olivares et al. (2008) apply this inversion approach of Rust (1987) to a static newsvendor problem, recovering implicit overage and underage costs from observed operating-room reservation decisions at a single hospital.

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

5.2 Structural Parameters and Identification

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. These frictional costs reflect search effort, placement fees, scheduling frictions, and mismatch between skill of the agency nurse and job requirements. 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 and $\theta \geq 0$ is the marginal rationing cost, reflecting care quality degradation, regulatory risk, and the cost of care gaps. The rationing cost is assumed quadratic.[20] A_t denotes agency hiring in period t and S_t is the shortfall from target hours in period t.

The parameter θ governs the slope of the facility's inverse demand for agency hours. On a day with shortfall S_t, the marginal reduction in rationing cost from purchasing one additional agency hour is $\theta(S_t - A_t)$, which is decreasing in A_t. The interior first-order condition $\partial c_t/\partial A_t = 0$ equates this to the marginal wage cost $w_a(1+\alpha)$, yielding the facility's optimal agency demand, with θ determining how steeply the rationing benefit falls as additional hours are purchased.

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

[20] A linear rationing cost would imply that facilities always close the gap completely on agency-using days, leaving no residual shortfall. In the data, 38 percent of agency days have a residual shortfall exceeding half an hour, consistent with the interior solution implied by the quadratic specification. The estimation also allows for corner cases.

shortfall. The facility minimizes expected total cost:

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

The parameter α is the frictional markup on the spot wage; in the theory it maps to match efficiency through $\gamma_H = 1/(1+\alpha)$. When the spot market thickens, density (g) and congestion-relief (h) improvements raise γ_H and α falls.[21]

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.

How α and θ are identified depends on a facility's position relative to two corners of the staffing problem: whether it ever uses agency labor, and whether its permanent buffer ever falls short of need. I work through three exhaustive cases.

Case A: interior solution (agency-using facilities). A facility that calls in agency on a meaningful number of days reveals both margins of its problem. The ex-post FOC for A_t identifies the residual shortfall r tolerated on agency-using days, and the ex-ante FOC for the buffer identifies the permanent staffing level. Together they identify α and θ jointly:

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

[21] The canonical newsvendor sets $\alpha = 0$ and $\theta \to \infty$, forcing the shortfall 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).

where r_i is the mean residual shortfall on agency days at facility i (the average of $S_t - A_t$ when $A_t > 0$). Equation (18) identifies the ratio $(1+\alpha)/\theta$. The ex-ante buffer FOC, obtained by differentiating (17) with the optimal A_t^* substituted in, identifies α given r_i:

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

Given observed r_i, mc_e, w_a, and the empirical gap CDF, (19) determines α and (18) then yields $\theta = w_a(1+\alpha)/r_i$. The two equations identify $\hat{\alpha}_i$ and $\hat{\theta}_i$ separately for each facility.

Case B: corner solution (non-agency facilities with positive shortfalls). A facility that almost never calls in agency labor but still has days where its permanent staffing falls short of need is at a corner of the agency margin: the effective spot price $w_a(1+\alpha)$ is high enough that absorbing every shortfall through rationing dominates paying for agency. Equation (18) is uninformative because there are too few agency days to estimate r_i with precision. With no agency use, the facility's only choice is the level of permanent staffing H; substituting $A_t \equiv 0$ into (17) reduces the program to $\min_H mc_e H + (\theta/2)\,\mathbb{E}[\max(G_t, 0)^2]$. The first-order condition with respect to H is

$$mc_{e,i} = \theta_i \cdot \mathbb{E}[\max(G_{it},\, 0)], \tag{20}$$

evaluated at the facility's chosen H and the empirical gap distribution it implies, which gives $\theta_i = mc_{e,i}/\mathbb{E}[\max(G_{it}, 0)]$. The interpretation is that at its chosen H the facility is indifferent at the margin between hiring one more permanent hour (cost mc_e) and tolerating the expected shortfall that hour would cover (valued at $\theta \times \mathbb{E}[\max(G, 0)]$). For Case-B facilities a facility-specific $\hat{\alpha}_i$ is not separately identified; the CBSA-level estimate $\hat{\alpha}_{ct}$ defined below is used when α for these facilities is needed for welfare and counterfactuals. For the small share of CBSA-years with no Case-A facilities (typically rural CBSAs with no observable agency market), $\hat{\alpha}_{ct}$ is unavailable and the hour-weighted mean of $\hat{\alpha}_{ct}$ across other CBSAs in the same state-year is used instead.[22]

Case C: fully-buffered facilities. A facility whose staff exceeds the target on nearly every observed day is at a second corner: rationing is essentially never operative. Operationally, these facilities behave as if $\theta \to \infty$, which collapses (17) to the canonical newsvendor of equation (14) with the frictional markup applied to the agency wage. The buffer is set at the empirical quantile $H^b = F_G^{-1}\big(1 - mc_e/[w_a(1+\alpha)]\big)$ of the facility's daily gap distribution, where F_G is the empirical CDF of the facility's daily gap over Mon–Fri days in the facility-year. In this scenario, every shortfall above H^b is fully covered by agency at cost $w_a(1+\alpha)$, with α equal to the CBSA-level pooled estimate $\hat{\alpha}_{ct}$ when available, or the state-year hour-weighted mean as a fallback. This theoretically implied buffer is used for counterfactuals in the welfare analysis below.

CBSA-level estimation of α. In the theory, $\alpha = 1/\gamma_H - 1$ measures the equilibrium quality discount prevailing in the CBSA's spot market. Averaging the facility-level $\hat{\alpha}_i$ from Case A within CBSA-years is feasible but introduces substantial sampling error: each $\hat{\alpha}_i$ is identified from only 50–150 facility-days. Applying the two FOCs directly to the pooled CBSA-day gap distribution yields more precise estimates at the relevant unit of analysis. For each CBSA-year, I pool all facility-days from Case-A facilities and construct the CBSA-level version of (18)–(19). Equation (18) yields a CBSA-level $\hat{r}_{ct}$ from the mean of $S_{it} - A_{it}$ across all agency-using days in the CBSA-year. The CBSA-year version of (19) is

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

[22]In the 2018–2024 sample, this state-year fallback covers about 19 percent of imputed Case-B/C facility-years; a residual 0.5 percent in states with no observable agency activity in any CBSA are dropped from the imputed sample.

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

5.3 Defining the Staffing Target

The gap and shortfall measures entering the FOCs require a definition of daily staffing need at each facility. A natural target is the CMS staffing benchmark for RN HPPD implied by each facility's CMS staffing star rating and resident acuity.[23] CMS publishes expected staffing thresholds for each staffing domain star level and scales them for resident acuity using the ratio of the facility's Resource Utilization Group Version IV (RUG-IV) case-mix index to the national average: $h^*_{jt} = \text{RN\_thresh}[\text{star}_{jt}] \times \frac{\text{CM\_RN}_{jt}}{\overline{\text{CM\_RN}}_t}$, 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 $\text{need}_{jt} = h^*_{jt} \times \text{census}_{jt}$, where census_{jt} is patient census for facility j in period t, and the daily gap is $G_{jt} = \text{need}_{jt} - H_{jt}$.[24]

I conduct two analyses to confirm that this target is relevant and that actual staffing tracks h^*_{jt}. The left panel of Figure 7 shows raw employee HPPD distributions by star rating for each facility-year. As expected, 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 0.26 HPPD for 2-star facilities to 0.75 for 5-star facilities.[25] The right panel re-centers the x-axis by each facility's own h^*_{jt} and plots the histogram of HPPD relative to distance

[23]The HPPD thresholds reflect the cut points CMS uses to assign the staffing domain rating in the Five-Star Quality Rating System, which in turn feeds into the facility's composite overall rating (Centers for Medicare & Medicaid Services, 2022). They are not binding regulatory floors on hours worked.

[24]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.

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

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.

As a descriptive validation, I run a stacked difference-in-differences (see Appendix G for implementation details) examining how facility HPPD evolves around years in which the facility's CMS staffing benchmark changes. It asks whether staffing moves discretely with changes in the benchmark in a way that is larger than would be expected from smooth continuation of pre-existing trends, which could mechanically arise because star ratings are a function of staffing. Figure 8 shows that employee HPPD increases by 0.045 at event year 0 following a one-star increase. This jump is substantially larger than what a continuation of the pre-event trend would predict, consistent with facilities adjusting their staffing to the new benchmark. There is no evidence that an increase in minimums results in a higher hourly wage of regular staff RNs, and if anything there is a wage decline, inconsistent with the monopsony prediction following an expansion of employment. This finding tracks with the results in Matsudaira (2014) who found no wage response to employment gains for state-level minimums. The agency wage is unaffected by the star change, as would be expected if it were competitively determined. Agency share is effectively flat.[26]

5.4 Empirical Implementation

This subsection maps the identification cases of Section 5.2 onto the data and walks through the mechanical estimation steps.

Gap construction. The formal model in Section 2 focuses on demand-side uncertainty, but in practice the daily coverage gap also reflects supply-side shocks such

[26] The lack of change 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).

as employee call-outs and scheduling gaps. Patient counts fluctuate daily with admissions, discharges, and acuity; employee call-outs generate supply-side shocks to effective permanent hours. Both sources are observed in the PBJ, and both enter the main empirical objects (such as the staffing shortfall) symmetrically, so the estimation treats the combined daily gap as the relevant object without requiring the two sources to be separated.

Because daily employee hours vary with call-outs and scheduling variability, for the empirical implementation I write $H_{it} = H + \mu_{it}$, where H is the facility-chosen buffer, set at the within-facility-year mean of H_{it}, and μ_{it} is mean-zero scheduling noise. This formulation is consistent with the theory under the assumption that the within-facility-year distribution of μ_{it} is invariant to H. Specifically, with random scheduling noise, the resulting buffer FOC takes the same form as in the model and is identified directly from the empirical CDF of the realized gap G_{it}.[27] 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 $\text{need}_{it} = h^*_{it} \times \text{census}_{it}$ uses the acuity-adjusted CMS minimum from Section 5.3. 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 $A_{it} > 0$.

Sample construction and partition. Estimation uses Mon–Fri days only; facility-years are required to have at least 200 calendar days (including weekends) of valid PBJ records to ensure that the empirical gap distribution is sufficiently characterized. Employee hourly compensation mc_e (salary plus fringe benefits) is the HCRIS-derived total compensation per paid hour. The CBSA-year agency marginal cost w_a is the hours-weighted mean of facility-level contract RN hourly rates from HCRIS, averaged

[27] Expected daily cost as a function of the chosen buffer is $C(H) = mc_e\, H + \mathbb{E}[c_A(G_t)]$, where $G_t = D_t - H_{it} = (D_t - H) - \mu_{it}$ is the realized gap and $c_A(G) = 0$ on $G \le 0$, $(\theta/2)G^2$ on $0 < G \le r$, and $w_a(1+\alpha)(G-r) + (\theta/2)r^2$ on $G > r$, with $r = w_a(1+\alpha)/\theta$. Since $\partial G_t/\partial H = -1$, the FOC for the optimal buffer is $C'(H) = mc_e - \theta\,\mathbb{E}[G \cdot \mathbf{1}_{\{0<G\le r\}}] - w_a(1+\alpha)\,P(G > r) = 0$. Substituting $\theta = w_a(1+\alpha)/r$ yields equation (19), with the expectation taken under the realized gap distribution.

across all facilities in the CBSA-year that report contract RN costs.[28]

The full estimation sample contains 11,752 SNFs and 56,381 facility-years matched to HCRIS wages and the CMS inputs needed to compute the acuity-adjusted minimum. All facility-years enter the analysis; each falls into exactly one of the three cases of Section 5.2. Case A (29.8 percent; 16,829 facility-years) are agency users with at least five agency-using days. Case B (46.5 percent; 26,189) are non-agency facilities with positive shortfalls. These are defined as facility-years with fewer than five agency days but whose expected positive shortfall on weekdays is at least 0.1 hours. Case C are fully-buffered facilities (23.7 percent; 13,363). These are facility-years with $\mathbb{E}[\max(G_{it}, 0)] < 0.1$ hours, where rationing is essentially never operative.

Estimation steps. The mechanical steps that implement the identification of Section 5.2 are:

Case A. Compute r_i and the empirical CDF of G_{it} from the facility-year, apply to (19) to solve for α_i, then recover θ_i from (18). The procedure is closed-form.[29]

Case B. Compute $\mathbb{E}[\max(G_{it}, 0)]$ over weekdays in the facility-year and apply (20); set α to $\hat{\alpha}_{ct}$, or to the state-year hour-weighted mean of $\hat{\alpha}_{ct}$ if the CBSA-year value is unavailable.

Case C. Assign $\hat{\theta}_i \to \infty$ and apply the canonical newsvendor of (14) with $\alpha = \hat{\alpha}_{ct}$ (with the state-year fallback as in Case B).

The pooled CBSA-year estimator $\hat{\alpha}_{ct}$ used in Cases B and C is the solution to the CBSA-level version of (21) described in Section 5.2.

[28]Using the CBSA-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 CBSA. The CBSA-year wage requires at least two reporting facilities per CBSA-year to exclude isolated outlier reports.

[29]With r_i fixed by (18), the bracketed term in (19) is a known constant, so the equation is linear in $w_a(1 + \alpha)$ and has a unique solution.

5.5 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.

A standard newsvendor model predicts that facilities with more volatile permanent staffing per patient day should use more agency.[30] 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 0.252 (column 1), highly significant and essentially unchanged after adding controls, year fixed effects, or CBSA fixed effects. The optimal-share gradient on employee-hour volatility is smaller (0.094–0.133), 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 9 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 structural parameters recovered from the newsvendor moments. The mean CBSA-year frictional markup from the pooled-moment estimator (21) is $\hat{\alpha}_{ct} = 0.480$, meaning the effective cost of an agency nurse exceeds the posted wage by 48 percent once frictions

[30] Because the expected shortfall $\mathbb{E}[\max(G - H, 0)]$ is a convex function of G for any fixed buffer H, a mean-preserving spread in the gap distribution raises expected agency use at every buffer level, and therefore raises both expected agency use and the optimal agency share at the facility's optimum.

are accounted for; the large standard deviation reflects substantial cross-CBSA heterogeneity. The mean marginal rationing cost across the combined Case A and B sample of 43,164 facility-years is $\hat{\theta} \times r = \$108$ per hour. Among agency-using facilities, the mean tolerated residual shortfall is 5.5 hours per day, implying that even when facilities call in agency nurses they leave meaningful unmet need at the margin.

Panel A of Figure 10 shows the time series of $\hat{\alpha}_{ct}$ from 2018 to 2024 and Table 5 reports the annual distribution of the CBSA-year frictional markup, along with significance tests against the 2018 baseline and year-on-year changes. The mean markup is 0.606 in 2018 and 0.587 in 2019, then falls sharply: by 2020 it has declined to 0.455, significant at the five-percent level. The 2021 and 2022 means of 0.451 and 0.448 are also significant at the five-percent level relative to 2018. The markup then reverses, rising to 0.515 in 2023 and to 0.650 in 2024. Neither value is statistically distinguishable from the 2018 baseline. 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.

Panel B of Figure 10 shows the time series of the hour-weighted mean marginal rationing cost $\hat{\theta} \times r$ (evaluated at each facility-year's typical agency-day shortfall) over the same period. Rationing costs rise monotonically throughout the sample, accelerating after 2020 and reaching roughly \$160 per hour by 2024. Panels C and D of the figure are discussed in Section 6.

6 Welfare

This section uses the facility-level parameter estimates from the newsvendor model to estimate the welfare implications of spot market frictions, counterfactuals on the effects of different levels of frictions, and estimate the marginal external benefit of firms using the spot market.

6.1 Aggregate Demand and Surplus

The welfare analysis that follows uses the newsvendor estimates to estimate market-year demand curves for agency which allows for estimation of surplus.

Starting at the facility-level, on a day with shortfall S_t, the ex-post first-order condition equates the marginal benefit of an additional agency hour to the effective spot price: $\theta_i(S_t - A_t) = w_a(1 + \alpha_i)$. This condition implies that facility i's marginal willingness to pay for agency hours, as a function of quantity A_i purchased, is

$$P_i(A_i) = \theta_i\big(\tilde{G}_i - A_i\big), \tag{22}$$

where $\tilde{G}_i$ is the facility's expected shortfall on agency-using days and the slope θ_i is the rationing cost, both recovered from the structural estimation (Section 5). Inverting equation (22), facility i demands $A_i(P) = \tilde{G}_i - P/\theta_i$ agency hours at price P. For each CBSA-year I compute the aggregate demand schedule as the horizontal sum of these facility-level curves:

$$A^{\text{agg}}_{ct}(P) = \sum_{i \in c} \max\left\{\tilde{G}_i - \frac{P}{\theta_i},\ 0\right\}, \tag{23}$$

and $P^{\text{agg}}_{ct}(\cdot)$ denotes its inverse.

The market equilibrium price paid by facilities is $P^*_{ct} = w_a(1 + \hat{\alpha}_{ct})$, where $\hat{\alpha}_{ct}$ is the pooled frictional markup from equation (21). The equilibrium quantity A^*_{ct} is defined by $P^{\text{agg}}_{ct}(A^*_{ct}) = P^*_{ct}$.[31]

Producer surplus is $\text{PS}_{ct} = \int_0^{A^*_{ct}} P^{\text{agg}}_{ct}(A)\, dA\ -\ P^*_{ct} \cdot A^*_{ct}$, which is the area above the

[31] This linear approximation passes through the equilibrium (P^*, A^*) point but understates the convex expected demand at high prices, possibly understating PS in Table 7.

effective price and below the demand curve up to the equilibrium quantity. Friction waste is the rectangle between the two price lines evaluated at the equilibrium quantity: Friction waste$_{ct} = (P^*_{ct} - w_a) \cdot A^*_{ct} = w_a\, \hat{\alpha}_{ct} \cdot A^*_{ct}$. This quantity represents resources consumed by search, placement, scheduling, and match-quality frictions rather than by nursing care. Figure 11, panel (a), illustrates these components graphically for 2019 aggregating to the national level. Panel (b) shows how the national aggregate demand curve shifts outward from 2018 to 2023 as both the scale of agency market participation and $\hat{\theta}$ increased, before partially receding in 2024.

Table 7 reports annual means of producer surplus and frictional waste across CBSA-years. The quantities are computed separately for each CBSA-year (requiring facilities with estimated θ_i parameters, yielding 1,047 CBSA-years across 276 CBSAs), and the table entries are CBSA-year means weighted by CBSA-year total employee RN hours within each calendar year. Producer surplus per agency hour nearly doubled over the sample, rising from \$201 in 2018 to \$399 in 2024, while the frictional cost per hour ranged from \$26 to \$42. Facilities receive an average of \$2.19 to \$3.63 of surplus per dollar of agency expenditure, rising over time as the demand curve shifted outward.

Panels C and D of Figure 10 display the time series of the same two quantities. Panel C plots the ratio of producer surplus to effective agency expenditure, $\mathrm{PS}/[w_a(1+\hat{\alpha}_{ct})\, A^*_{ct}]$, which rises from 2.2 in 2018 to a peak of 3.6 in 2023 before partially reverting to 3.4 in 2024 as the demand curve shifted outward then partially receded. Panel D plots the frictional waste share, $\mathrm{Friction}/(\mathrm{PS} + \mathrm{Friction})$, which falls from roughly 18 percent pre-pandemic to 12 percent in 2020–2023 as the spot market thickened, then rebounds to 14 percent in 2024 as $\hat{\alpha}_{ct}$ climbs back above its 2018 level.

In principle, the aggregate demand estimates can be used to estimate deadweight loss, but the facility-level first-order conditions are better suited to evaluate inefficiency under different counterfactuals on frictional markups, because they can also be applied to facilities that do not use agency. Table 8 asks how resource costs—hiring frictions, rationing, and precautionary hiring—would change if spot market frictions were lower. The counterfactual varies the frictional markup α, holding fixed each facility's parameters ($\hat{\theta}_i$, $mc_{e,i}$, $\bar{w}_a$) and its empirical gap distribution. For each CBSA-year, α is set to $\alpha^{\text{cf}} = \min\{\alpha^{\text{target}},\ \hat{\alpha}_{ct}\}$, where α^{target} takes one of two values from the cross-sectional distribution of $\hat{\alpha}_{ct}$: the 25th percentile (p_{25}: $\alpha^{\text{target}} = 0.066$) or the median ($p_{50}$: $\alpha^{\text{target}} = 0.315$). The reduction applies only where it improves on the observed value. I then re-solve the FOC from equation (17) for each facility at the new α, which reduces the tolerated residual, $r^{\text{cf}} = w_a(1 + \alpha^{\text{cf}})/\hat{\theta}_i < r^{\text{obs}}$, so agency covers more of each shortfall, and reduces the optimal buffer H_i^b because the effective agency price $w_a(1 + \alpha^{\text{cf}})$ is lower. The buffer satisfies:

$$mc_{e,i} = \hat{\theta}_i \cdot \mathbb{E}\big[\min\big(\max(g_{it} - H_i^b,\, 0),\ r\big)\big], \tag{24}$$

where the only unknown is the buffer H_i^b. The other inputs are taken as observed or pre-estimated: $mc_{e,i}$ and w_a from HCRIS, the rationing cost $\hat{\theta}_i$ from the structural estimation in Section 5, and the frictional markup α either at the facility's observed CBSA-level value $\hat{\alpha}_{ct}$ or at the counterfactual value α^{cf}. The implied tolerated residual $r = w_a(1 + \alpha)/\hat{\theta}_i$ is therefore also fixed by these inputs. Equation (24) is the same FOC as the estimation moment (19) written in unscaled form: the agency-margin saving $w_a(1+\alpha)\, P(g_{it} > H_i^b + r)$ is embedded in the r-saturated tail of the truncated expectation through the relation $\hat{\theta}_i\, r = w_a(1 + \alpha)$.[32] The truncated expectation is evaluated as a sample analog on the facility-year's empirical gap distribution $\{g_{it}\}$ (the actual weekday observations from PBJ), and (24) is solved numerically for H_i^b. The same FOC applies to facilities that previously did not use agency, allowing for

[32] The integrand $\min(\max(g - H^b, 0), r)$ equals 0 on $\{g \le H^b\}$, $g - H^b$ on $\{H^b < g \le H^b + r\}$, and the cap r on $\{g > H^b + r\}$ (the "r-saturated tail"). The saturated tail contributes $r \cdot P(g > H^b + r)$ to the expectation; multiplying by $\hat{\theta}_i$ and using $\hat{\theta}_i r = w_a(1 + \alpha)$ turns this into the agency-saving term $w_a(1 + \alpha)\, P(g > H^b + r)$.

an extensive-margin response to the changing friction.[33]

The table decomposes total resource loss into three components, each computed per facility per day at the counterfactual buffer $H_i^{b,\mathrm{cf}}$ and then aggregated:

$$\mathrm{Friction}^{\mathrm{cf}} = \big[w_a(1+\alpha^{\mathrm{cf}}) - w_a\big] \times A_{it}^{\mathrm{cf}}, \tag{25}$$

$$\text{Precautionary expenditures}^{\mathrm{cf}} = mc_{e,i} \times \max\big(H_i^{b,\mathrm{cf}} - g_{it},\ 0\big), \tag{26}$$

$$\text{Rationing cost}^{\mathrm{cf}} = \frac{\hat{\theta}_i}{2}\big[\min\big(\max(g_{it} - H_i^{b,\mathrm{cf}},\ 0),\ r^{\mathrm{cf}}\big)\big]^2, \tag{27}$$

where $A_{it}^{\mathrm{cf}} = \max\big(g_{it} - H_i^{b,\mathrm{cf}} - r^{\mathrm{cf}},\ 0\big)$ is counterfactual agency hours on day t and $r^{\mathrm{cf}} = w_a(1+\alpha^{\mathrm{cf}})/\hat{\theta}_i$. The empirical gap g_{it}, the wage w_a, the marginal labor cost $mc_{e,i}$, and the rationing cost $\hat{\theta}_i$ are held fixed at their observed/estimated values; only α, H_i^b, r, and A_{it} change between the observed and counterfactual scenarios. The same expressions evaluated at $\alpha = \hat{\alpha}_{ct}$ and $H_i^b = H_i^{b,\mathrm{obs}}$ give the observed levels reported in the first column of Panel A. Friction waste is the frictional markup above the spot wage. Precautionary expenditures is the regular nurse expenditure on days they are not needed to satisfy the target. Rationing cost is the welfare loss from unmet patient need, evaluated at the estimated rationing parameter. Note that precautionary expenditure and rationing can both be privately optimal and efficient given demand uncertainty. However, any reduction in either of these under a lower markup is a social resource gain.

Panel A of Table 8 reports the estimates for 2019. Because the estimation sample has 7,546 facilities in 2019, I scale the estimates by a factor of two to approximate the 15,000 SNFs nationally. Total observed weekday waste is $1.2 billion per year, of which precautionary capacity accounts for $539 million (44 percent), rationing

[33]This counterfactual holds the CBSA-level frictional markup α fixed at the counterfactual cap and re-solves only the facility-level FOCs. It therefore abstracts from the general-equilibrium feedback emphasized in Section 2: when α falls and more facilities shift to spot, the equilibrium agency share rises, which would further reduce α. The reported estimates are thus a lower bound on the welfare gains from reducing frictions to the benchmark market.

\$409 million (34 percent), and friction \$269 million (22 percent).[34] Under the 25th percentile frictional markup ceiling, total waste falls 39 percent. Frictional loss drops 62 percent, precautionary labor expenditures fall 46 percent, and rationing costs fall 15 percent. Under the median markup ceiling total waste falls 18 percent.

Panel B shows the staffing adjustment. The reported H^b is the FOC-implied buffer relative to observed employee staffing: a negative value means the model recommends total permanent hours ($emp + H^b$) below the facility's current employee hours. The observed $H^b = -8.5$ leaves $E[\text{idle}] = 3.6$ hours per day positive because the daily gap distribution has a wide left tail (high-slack days). The change columns are signed counterfactual minus observed. Under the 25th percentile cap the buffer falls by an additional 4.5 hours per facility per day; of that, 1.6 hours comes off slack days (idle falls), while agency hours rise by 3.4 (a 117 percent increase from a base of 2.9 hours). The buffer response is dampened by the rationing cost $\hat{\theta}$: because agency leaves a tolerated residual of r hours unfilled, reducing α also reduces r and raises the marginal value of each remaining buffer hour, partially offsetting the incentive to cut buffer.

6.2 Marginal External Benefit of Spot Market Participation

I now use the structural estimates to compute the marginal external benefit when a firm shifts one nurse from its private buffer to the spot market. This is the empirical counterpart to B^* in the theoretical model: the gain in hiring efficiency that accrues to all other facilities sharing the CBSA spot market.

Since α is a real resource cost (search, screening, scheduling, match-quality losses), lower frictions translate directly into resources saved on every deployed agency hour. Expressed as a fraction of the engagement wage, the per-engagement externality is $B^*_{ct} = |\hat{\beta}_\alpha| \cdot s_{ct}$, where s_{ct} is the national HCRIS S-3 Part V annual agency share of RN hours (the series plotted in Figure 2). The regression coefficient $\beta_\alpha \equiv \partial\alpha/\partial s$,

[34] Weekends are excluded since the main estimation sample is weekday staffing.

estimated below, gives the drop in the frictional markup per unit of agency share added. Each unit of α saved frees w_a per deployed agency hour, and the share of total RN hours actually deployed in the spot market is exactly s_{ct}, so per dollar of spot wage the externality is $|\hat{\beta}_\alpha| \cdot s_{ct}$.

The parameter β_α is taken from the regression of $\hat{\alpha}_{ct}$ on CBSA agency share s_{ct}, with log RN hours as a scale control. Table 6 reports this regression, run on non-overlapping halves of each CBSA-year to avoid mechanical correlation between $\hat{\alpha}_{ct}$ and agency share (see Appendix G). The within-CBSA estimate is $\hat{\beta}_\alpha = -2.129$ ($p < 0.00$); a ten-percentage-point increase in agency share is associated with a 0.213 reduction in the frictional markup—about 36 percent of the pre-pandemic mean ($\bar{\alpha}_{2018\text{–}19} \approx 0.60$).[35]

Substituting $\hat{\beta}_\alpha = -2.129$ and aggregating by year, the marginal external benefit as a percentage of the spot wage is 6.5 percent in 2018, rising to 17.3 percent in 2022 as the market thickens, and partially reverting to 15.0 percent in 2024, with a sample mean of 11.8 percent (Table 9). The rise reflects the expansion of the agency share s_{ct} over this period; the regression slope $|\hat{\beta}_\alpha|$ is fixed across years.

7 Conclusion and Discussion

This paper studies the choice between holding labor as private buffer capacity and obtaining it from a spot market after demand is realized. Accounting for spot market liquidity has general equilibrium implications for the standard labor demand problem. The same structure appears in any market where firms or consumers choose between

[35] Taking the theoretical model at face value, the OLS coefficient is a conservative (attenuated) estimate of the structural parameter because positive omitted-variable bias from demand intensity δ_H pushes $\hat{\beta}_\alpha$ toward zero. In the model, $\alpha_{ct} = 1/[g(L_{ct})\, h(R_{H,ct})] - 1$, with spot size L_{ct} and slack $R_{H,ct} = L_{ct}/[\delta_{H,ct} F(\theta^*_{ct})]$. Conditioning on s_{ct} and log RN hours n_{ct} absorbs variation in θ^*_{ct} and N_{ct}, leaving $\delta_{H,ct}$ as the residual driver of α_{ct} through R_H. The omitted-variable bias is $(\partial\alpha/\partial\delta_H)|_{\theta^*,N} \cdot \mathrm{Cov}(\delta_H, s \mid n)/\mathrm{Var}(s \mid n)$, which is positive: higher δ_H reduces slack and raises α ($\partial\alpha/\partial\delta_H > 0$), and shifts sorting toward spot ($\mathrm{Cov}(\delta_H, s \mid n) > 0$ via $\partial\theta^I/\partial\delta_H > 0$). Positive bias attenuates $\hat{\beta}$ toward zero, so $|\hat{\beta}_\alpha|$ is a lower bound on $|\beta_\alpha|$.

private buffer capacity and a shared adjustment margin. Beyond inventory and labor markets, an example is capital markets, like compute, and the lease or buy decision.

The welfare results show that in a competitive market the marginal buffer nurse creates more expected value in the spot market than in private employment. The implied Pigouvian corrective subsidy relies on a tax mechanism that may not exist in these markets. However, in labor markets there are institutions that can be seen as solving the problem of missing spot markets, albeit introducing other distortions. An example is craft union hiring halls, used in the building trades, longshore work, and similar industries. Workers register with the hall and employers request labor through it; dispatch follows seniority, so workers accumulate a positional asset by remaining attached even during slack periods. Entry is controlled through apprenticeship and licensing, preventing seniority dilution. The queue therefore solves the commitment problem as workers stay available because their queue position has value, and they respond when called because refusing dispatch destroys it. In terms of the model, the hall sustains market participation in both demand states, keeping frictional markups low and rationing limited in the tight state. More generally, it is interesting to consider whether rent dissipation serves as a partial remedy for the unpriced value of availability.

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.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of financial economics*, 2022, *144* (2), 370–395.

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.

Caballero, Ricardo J. and Arvind Krishnamurthy, "Collective Risk Management in a Flight to Quality Episode," *The Journal of Finance*, 2008, *63* (5), 2195–2230.

Centers for Medicare & Medicaid Services, "Design for Nursing Home Care Compare Five-Star Quality Rating System: Technical Users' Guide," Technical Report, Centers for Medicare & Medicaid Services July 2022. Updated July 2022.

Chen, M Keith, Peter E Rossi, Judith A Chevalier, and Emily Oehlsen, "The value of flexible work: Evidence from Uber drivers," *Journal of political economy*, 2019, *127* (6), 2735–2794.

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.

Foster, Lucia, John C Haltiwanger, and Cornell John Krizan, "Aggregate productivity growth: Lessons from microeconomic evidence," in "New developments in productivity analysis," University of Chicago Press, 2001, pp. 303–372.

Gaines, Kathleen, "How to Become a Travel Nurse: Requirements and Salary," Nurse.org June 2025. Accessed 2026.

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.

Health Carousel Travel Nursing, "How Much Experience Do You Really Need to Be a Travel Nurse?," Health Carousel Travel Nursing Blog n.d. Accessed 2026.

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.

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," *Econometrica*, 2026, *94* (3), 991–1025.

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.

Olanrewaju, Sherif, Susan Loeb, and Manuel Rosaldo, "INTERNATIONALLY EDUCATED NURSES CARING FOR OLDER ADULTS: A SCOPING REVIEW," *Innovation in Aging*, 2022, *6* (Supplement_1), 678–678.

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.

Pissarides, Christopher A, "Short-run equilibrium dynamics of unemployment, vacancies, and real wages," *The American Economic Review*, 1985, *75* (4), 676–690.

Rust, John, "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 1987, pp. 999–1033.

Scott D. Pollock and Associates, "Understanding Green Cards for Nurses in the U.S.," Scott D. Pollock & Associates, P.C. 2026. Accessed 2026.

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

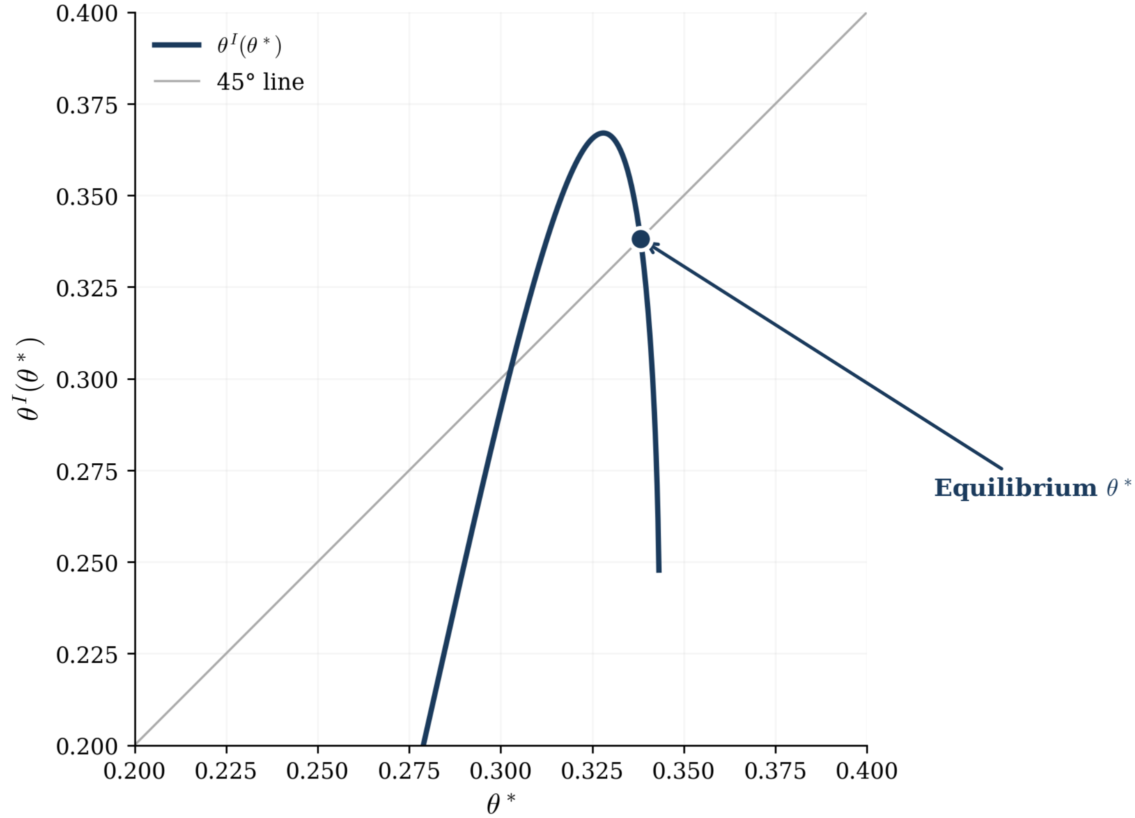


Figure 1: **Equilibrium threshold.** The indifference-threshold curve $\theta^I(\theta^*)$ from (10) and the 45-degree line. The equilibrium cutoff θ^* is labeled at the intersection. Parameters: $F = \text{Beta}(2, 2)$, $N = 0.85$, $\delta_H = 0.45$, $\delta_L = 0.05$, $\pi_H = 0.30$, $b = 0.015$.

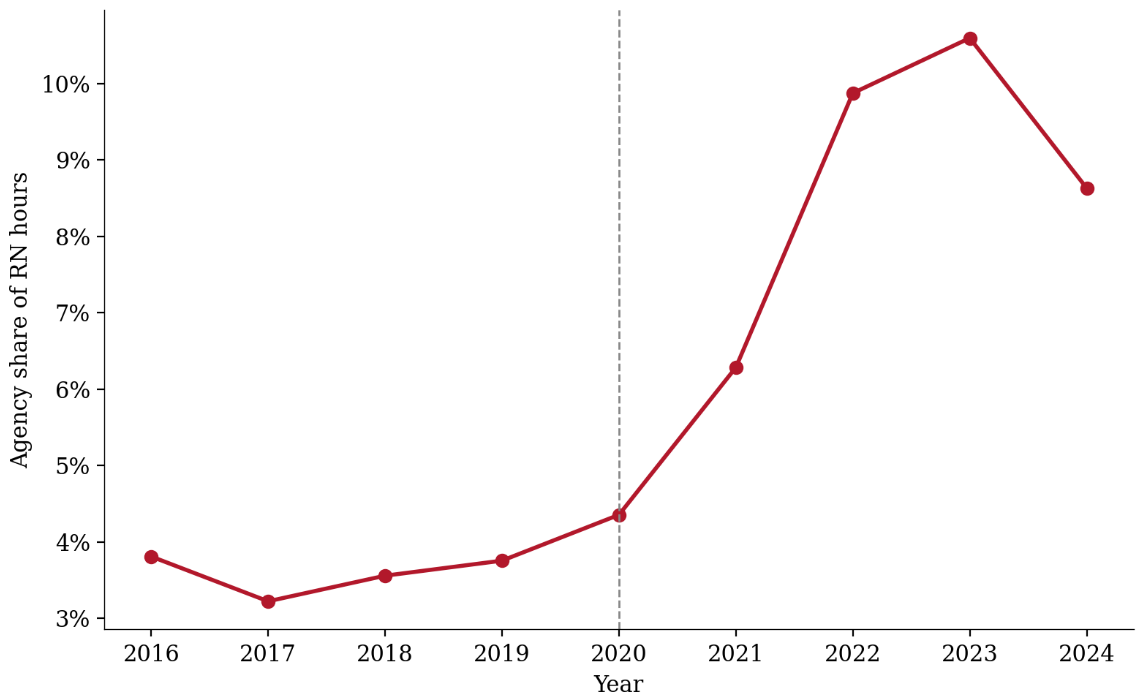


Figure 2: **RN agency share by year, SNFs.** Contract hours as a fraction of total RN hours, HCRIS Worksheet S-3 Part V, unbalanced panel 2016–2024.

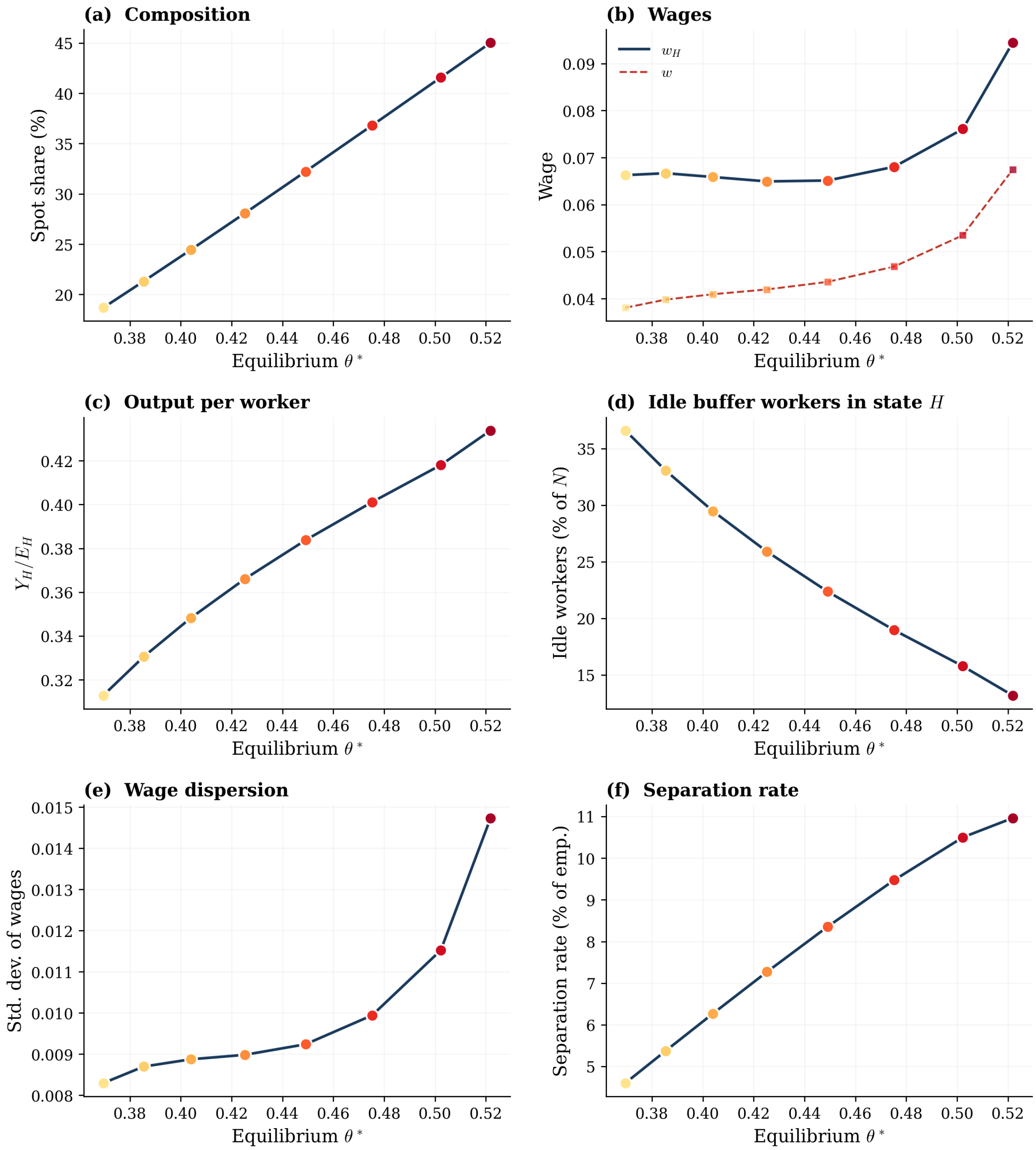


Figure 3: **Equilibrium outcomes under joint demand boom.** Panel (a): spot share (composition). Panel (b): scarcity wage w_H and buffer wage $\bar{w}$. Panel (c): output per worker (Y_H/E_H). Panel (d): idle buffer workers in state H (as % of N). Panel (e): wage dispersion (std. dev. of wages). Panel (f): separation rate. The model gives equilibrium employment as a function of the aggregate state but does not track individual workers across periods, so the separation rate requires an additional step. I assume that the aggregate state $s \in \{H, L\}$ is drawn i.i.d. each period at probabilities π_H, π_L and that workers separate when the state transitions from H to L (employment falls from N to $E_L < N$, where E_L is total employment in state L). A transition occurs with probability $\pi_H \pi_L$ and displaces $N - E_L$ workers; the plotted rate divides this per-period flow by expected employment $\pi_H N + \pi_L E_L$. 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.

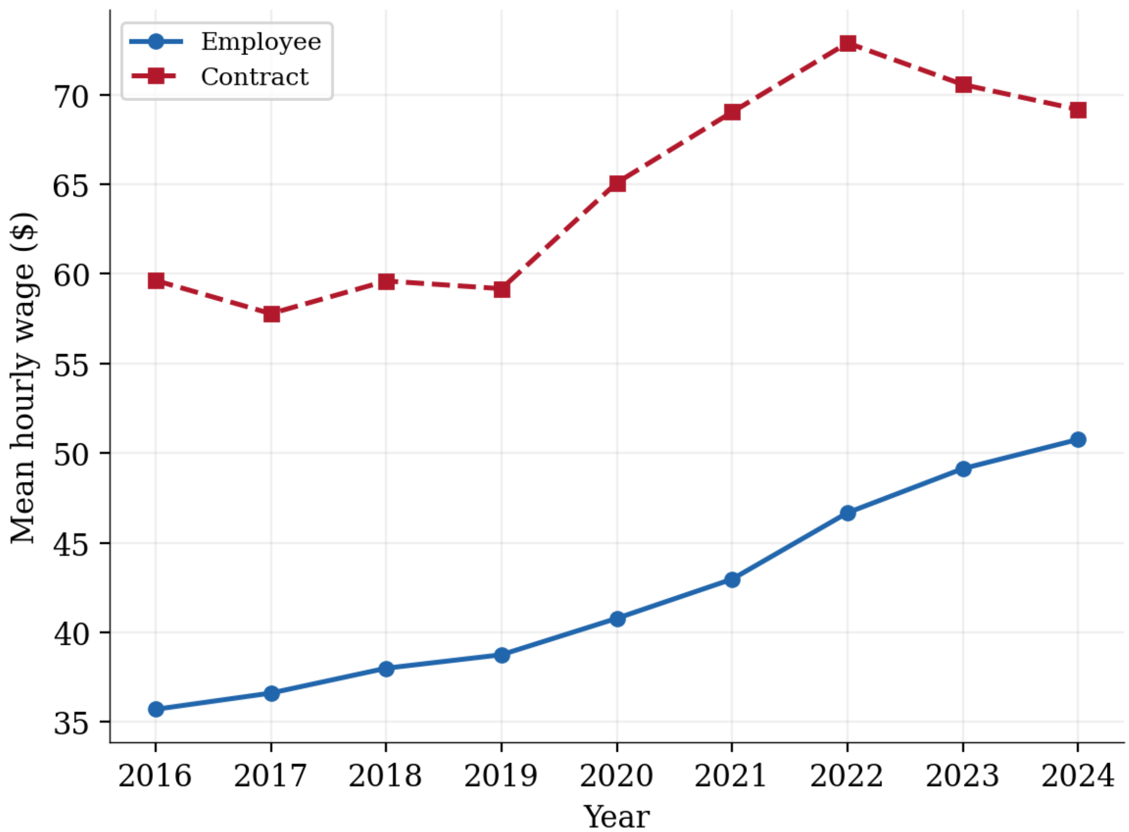


Figure 4: **RN average hourly wage: employee vs. contract, 2016–2024.** Mean hourly wage for employee (permanent, solid blue) and contract (dashed red) registered nurses, conditional on positive reported wage, HCRIS Worksheet S-3 Part V. Sample: SNFs with positive reported wages in both staffing modes, unbalanced panel 2016–2024.

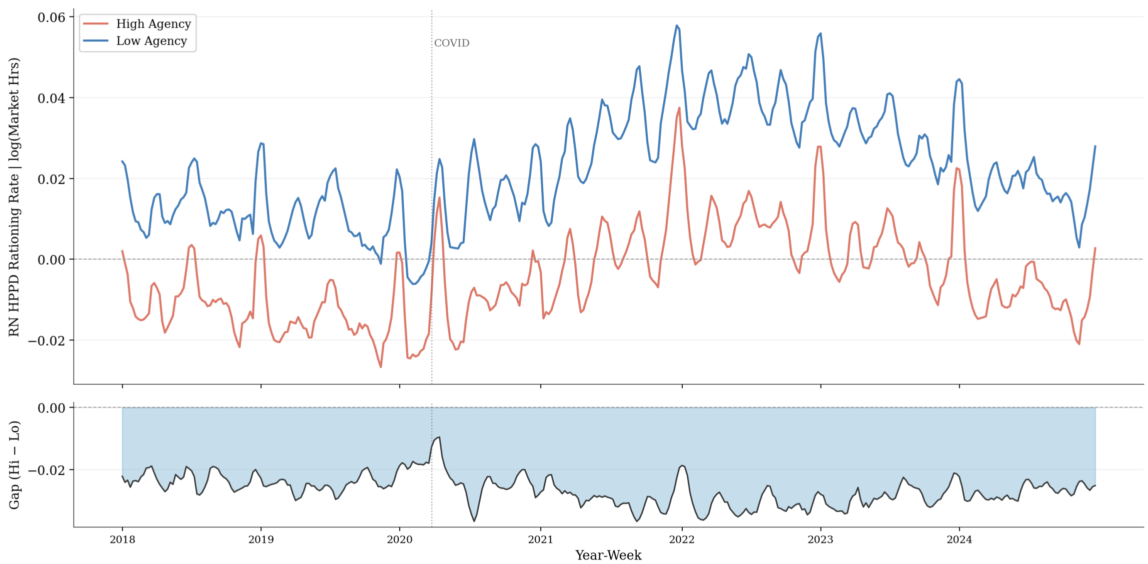


Figure 5: **RN rationing rate: high vs. low baseline agency CBSAs.** Weekly RN HPPD rationing rate residualized on log total nursing market hours in the CBSA-week, plotted as a 4-week moving average, 2018–2024. CBSAs are split at the median pre-2020 RN agency share. Top panel: residualized rationing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

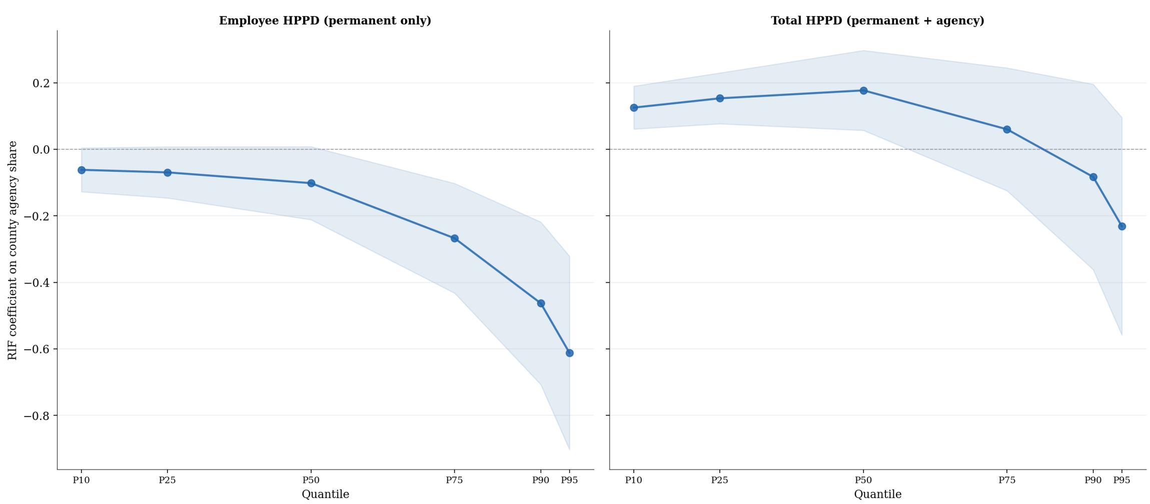


Figure 6: **Effect of county agency share on the HPPD distribution.** Firpo-Fortin-Lemieux unconditional quantile regression (UQR) coefficients on county leave-own-facility-out RN agency share, 2018–2024. Census-weighted; standard errors clustered by county; shaded bands show 95% confidence intervals. Left panel: dependent variable is employee (permanent) RN HPPD. Right panel: dependent variable is total RN HPPD (employee plus agency). Specification controls for log census, for-profit status, and case mix index.

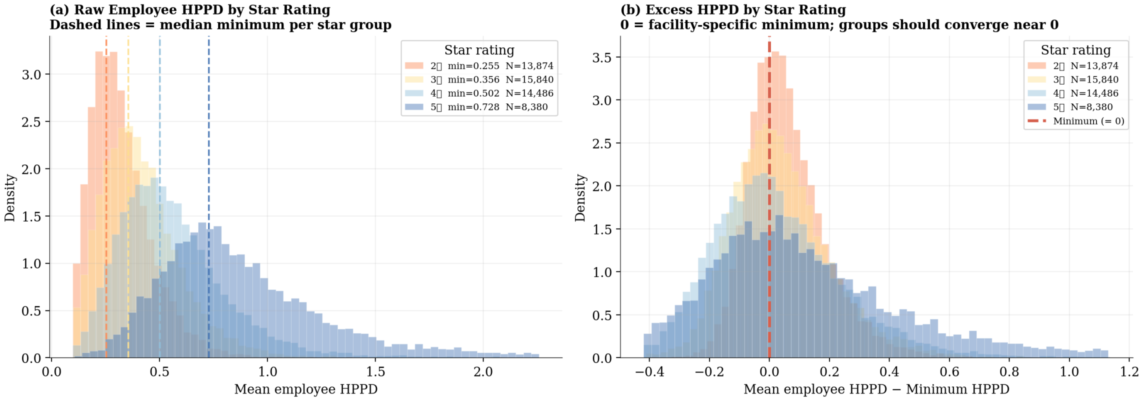


Figure 7: **Employee HPPD distributions by staffing star rating, 2019–2023.** Panel (a): raw mean employee HPPD by star rating (2–5 stars); dashed lines mark the median acuity-adjusted minimum h^*_{jt} for each group (2 star: 0.260, 3 star: 0.362, 4 star: 0.516, 5 star: 0.750). Panel (b): excess HPPD, defined as mean employee HPPD minus the facility-specific minimum h^*_{jt}. Sample sizes: 2 star: $N = 7{,}525$; 3 star: $N = 8{,}553$; 4 star: $N = 7{,}710$; 5 star: $N = 3{,}614$.

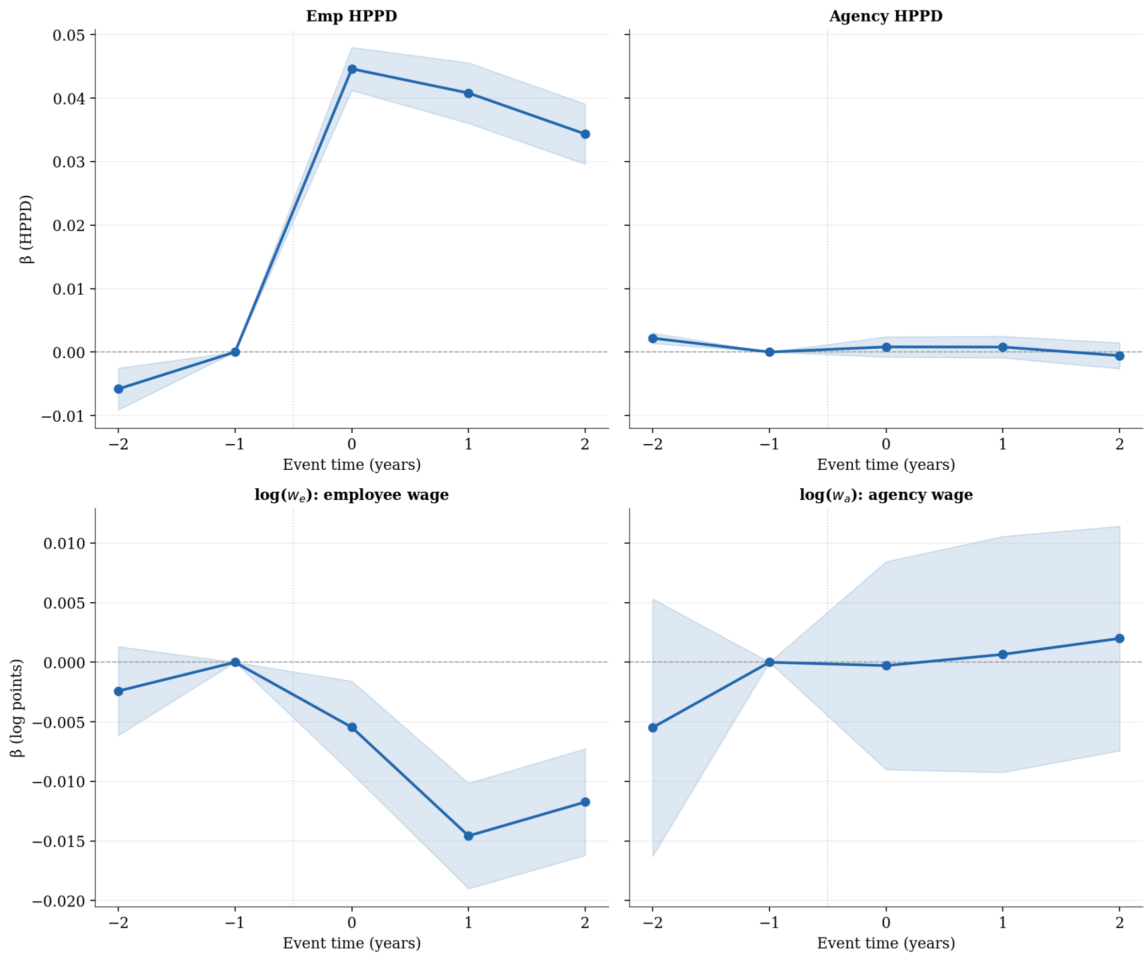


Figure 8: **Stacked difference-in-differences: facility response to shifts in the acuity-adjusted minimum.** Cohorts defined by facilities experiencing a shift in the CMS staffing star rating in 2020, 2021, and 2022; control group is never-treated facilities. The treatment is the signed star-rating change ($d_{ig} \in \{-2, -1, +1, +2\}$), so $\hat{\beta}_\ell$ is the outcome response per one-star shift. Fixed effects: facility-by-cohort and year-by-cohort. Standard errors clustered by facility; shaded bands show 95% confidence intervals. Base period: $\ell = -1$. Panels show employee HPPD in levels (blue), agency HPPD in levels (red), log employee wage w_e (green), and log CBSA agency wage w_a (purple). $N = 40{,}315$ facility-cohort-year observations; $G = 6{,}275$ facilities (5,185 treated, 1,090 never-treated controls).

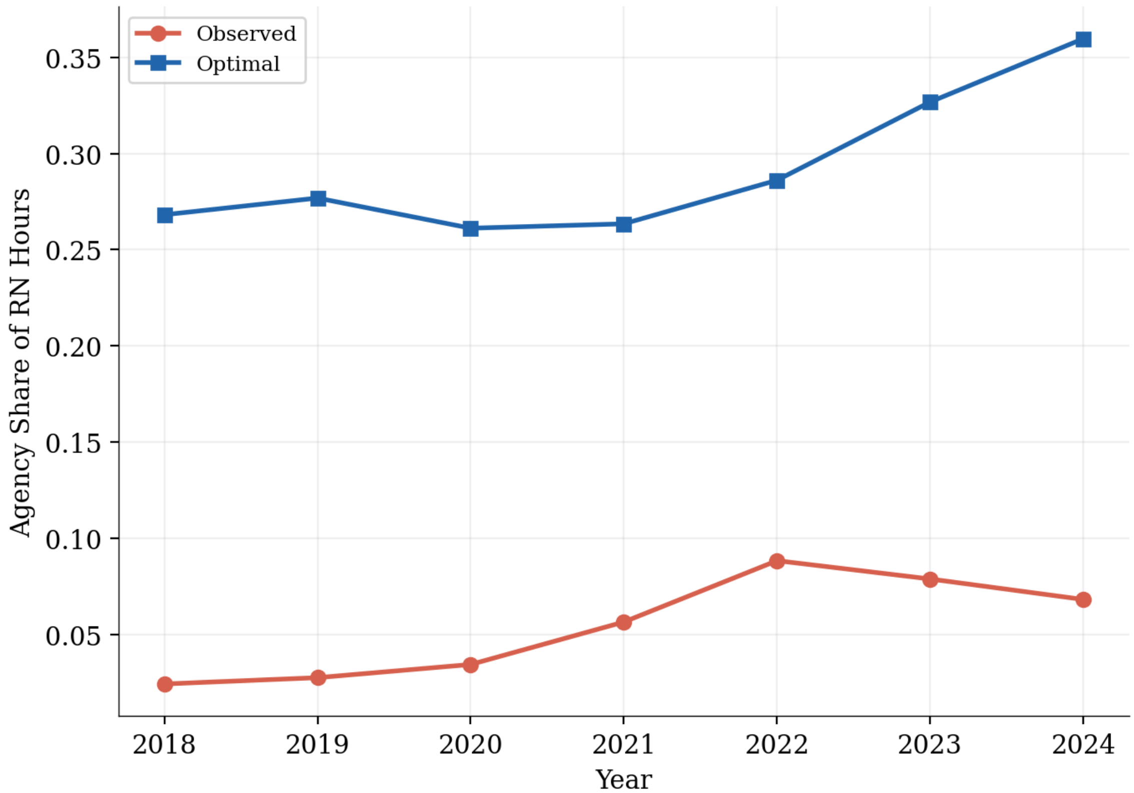


Figure 9: **Observed vs. frictionless-optimal agency use, 2018–2024.** Observed (red) and frictionless newsvendor optimum ($\alpha = 0$, blue) agency share of RN hours by year, averaged across facilities in the estimation sample. The frictionless optimum sets the permanent buffer at the $(1 - mc_e/w_a)$ quantile of the empirical gap distribution.

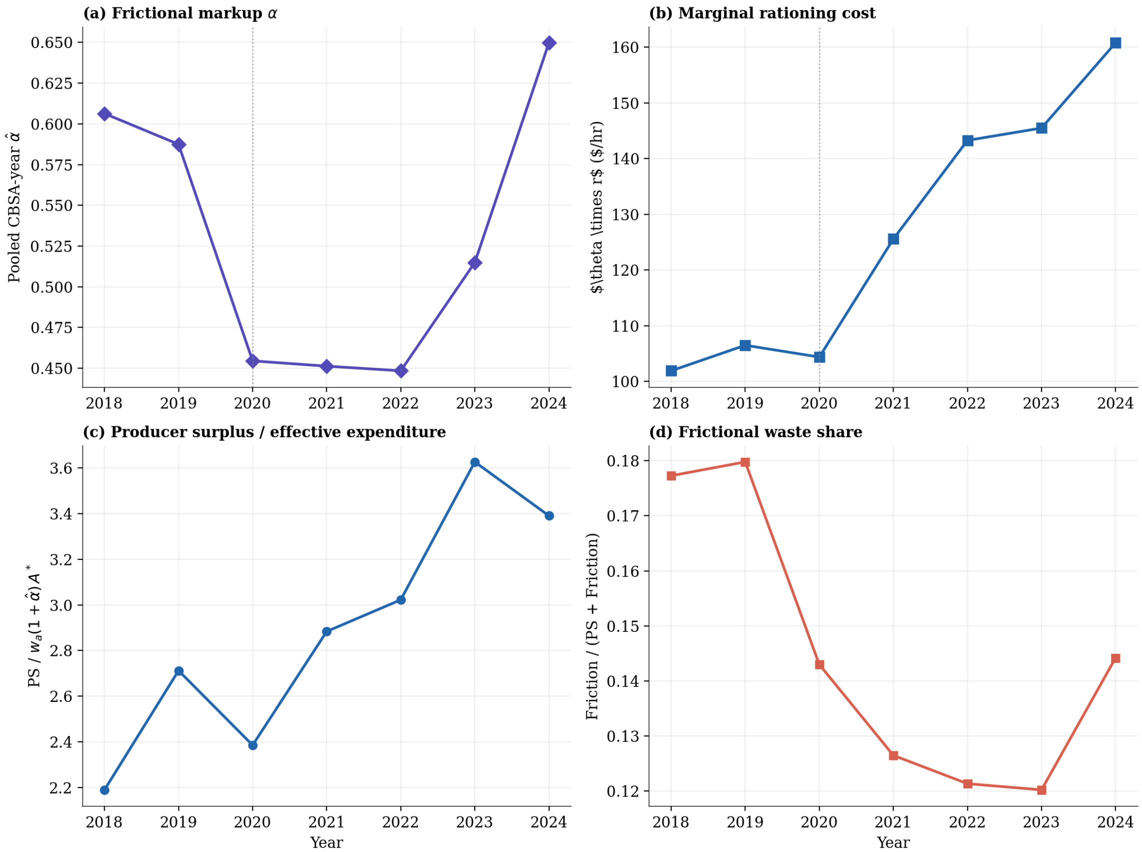


Figure 10: **Pooled CBSA-year structural parameters by year.** Panel (a): hour-weighted mean of the CBSA-year pooled frictional markup $\hat{\alpha}_{ct}$ estimated by applying the two structural FOC moments to the pooled CBSA-day gap distribution. Markup floored at 0 (negative-markup CBSA-years are treated as friction-free) and Windsorized at 1.5. Dotted vertical line marks 2020. Panel (b): hour-weighted facility-year mean of the marginal rationing cost $\hat{\theta} \times r$ ($/hr), pooled across Cases A and B (excluding Case C). Panel (c): hour-weighted CBSA-year mean of producer surplus over effective agency expenditure, $\mathrm{PS}/[w_a(1+\hat{\alpha}_{ct})\,A^*_{ct}]$, from the CBSA-level aggregate demand curve decomposition. Panel (d): hour-weighted CBSA-year mean of the frictional waste share, Friction/(PS + Friction).

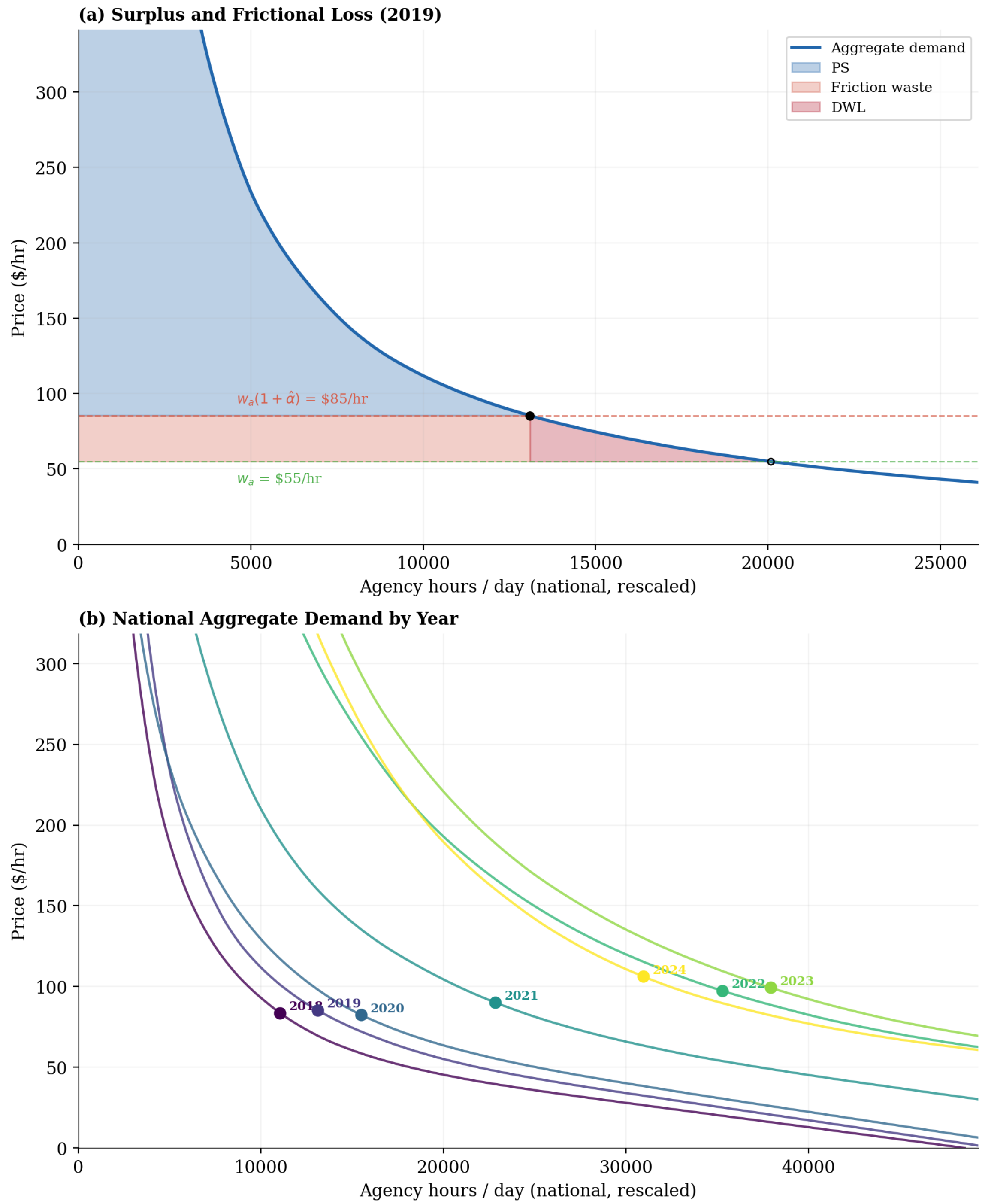


Figure 11: **Aggregate demand curve welfare decomposition.** Panel (a): national aggregate inverse demand for agency RN hours in 2019. The curve is the horizontal sum of facility-level inverse demands $P_i(A_i) = \theta_i(\tilde{G}_i - A_i)$ over all identified facilities ($\hat{r}_i > 1$ hr), with the quantity axis rescaled from the identified sample to the full national SNF universe (assumed at 15,000 facilities). The price axis is unaffected by the rescaling. Producer surplus (PS, blue shading) is the area above the effective spot price $P^* = w_a(1+\hat{\alpha})$; friction waste (light red rectangle) is $(P^* - w_a) \times A^*$. Filled circle marks the observed equilibrium; open circle marks the frictionless counterfactual.

Tables

Table 1: Summary Statistics. Sample: 56,381 facility-years, 2018–2024.

	N	Mean	SD
A. Demand, supply, and coverage			
Daily census (residents)	56,381	84.1	42.8
Daily staffing need (hrs)	56,381	29.3	24.6
Daily employee supply (hrs)	56,381	35.6	27.3
Daily gap: need minus supply (hrs)	56,381	-6.313	18.6
CV of employee hours	56,381	0.368	0.201
CV of daily census	56,381	0.0658	0.0457
Fraction of days total staffing $<$ 90% of need	56,381	0.296	0.307
Observed agency share of RN hours	56,381	0.0558	0.119
B. Wages and cost ratios			
Employee RN compensation w_e (\$/hr)	56,381	43.2	9.521
CBSA agency wage w_a (\$/hr)	56,381	62.9	13.2
C. Structural parameters			
Residual shortfall r (hrs/day on agency days)	16,829	5.511	8.465
Rationing cost $\hat{\theta} \times r$ (\$/hr)	43,018	107.7	119.6
Frictional markup $\hat{\alpha}_{ct}$ (CBSA-year)	1,985	0.480	0.757

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 Worksheet S-3 Part V; CBSA agency wage is the hours-weighted mean of facility-level contract RN hourly rates within the CBSA-year. In Panel C, the residual shortfall r is defined only for Case A facilities (those with ≥ 5 agency-using days, $N = 16,829$), since r is observed from the agency-day gap. The rationing cost $\hat{\theta} \times r$ is reported for the combined Case A+B sample ($N = 43,018$). $\hat{\alpha}_{ct}$ is the pooled-moment CBSA-year frictional markup above the posted agency wage, benefits-corrected.

Table 2: Between-Within Decomposition of RN Wage Growth

	Avg. wage ($/hr)	Total Δ ($/hr)	Within ($/hr)	Between ($/hr)	% Within	% Between
Panel A: Employee RN only						
2020	36.18	$1.52	$1.94	$0.25	127.7	16.4
2021	37.90	$3.24	$3.82	$0.30	117.9	9.1
2022	41.47	$6.81	$7.49	$0.31	110.1	4.6
2023	43.57	$8.91	$9.99	$0.27	112.2	3.0
Panel B: All RN labor (employee + agency)						
2020	36.94	$1.77	$2.16	$0.24	121.7	13.8
2021	39.27	$4.10	$4.68	$0.27	114.0	6.5
2022	43.96	$8.79	$9.32	$0.30	106.0	3.4
2023	45.86	$10.70	$11.60	$0.31	108.4	2.9

Notes: Between-within decomposition of the change in average RN hourly compensation relative to 2019. "Within" is the contribution from within-facility wage growth, computed by holding each facility's 2019 employment share fixed and averaging the facility-level wage changes weighted by those base-year shares. "Between" is the contribution from reallocation of hours across facilities, computed by holding each facility's 2019 wage fixed and re-weighting facilities by the change in their employment share between 2019 and year t; a positive value means hours shifted toward facilities that already paid above-average wages in 2019. Balanced panel of $N = 13,125$ SNFs observed in all years 2019–2023. Wages and hours from HCRIS Worksheet S-3 Part V (RN-specific). Panel B includes agency RN expenditure and hours.

Table 3: CBSA RN Agency Share and Facility Rationing

	(1)	(2)	(3)	(4)
CBSA agency share (LOO)	-0.7901^{***}	-0.7568^{***}	-0.9629^{***}	-0.3646^{***}
	(0.1396)	(0.1603)	(0.2357)	(0.1241)
log(CBSA RN hours)		-0.0179	-0.0159	0.0331^{***}
		(0.0113)	(0.0114)	(0.0036)
Year FE	No	No	Yes	Yes
CBSA FE	No	No	No	Yes
N	42,958	42,958	42,958	42,958
CBSAs	642	642	642	642
R^2	0.017	0.023	0.030	0.167

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 weighted by facility mean census. Standard errors clustered on CBSA in parentheses. CBSA agency share is the leave-own-facility-out fraction of annual RN hours supplied by agency staff in the CBSA-year, weighted by facility total RN hours. log(CBSA RN hours) is the log of total annual CBSA RN hours. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Agency Share and Employee Supply and Demand Volatility.

	(1) Bivariate	(2) + Controls, Year FE	(3) CBSA FE
Panel A: Observed agency share			
CV employee hours ($\sigma_H/\bar{H}$)	0.252***	0.255***	0.286***
	(0.037)	(0.040)	(0.055)
CV census ($\sigma_N/\bar{N}$)		0.090***	0.098***
		(0.023)	(0.018)
Log(mean census)		0.024***	0.027***
		(0.0044)	(0.0040)
N	44,571	44,571	44,571
R^2	0.192	0.224	0.366
Panel B: Optimal agency share			
CV employee hours ($\sigma_H/\bar{H}$)	0.133***	0.094***	0.111***
	(0.039)	(0.031)	(0.042)
CV census ($\sigma_N/\bar{N}$)		0.384***	0.257***
		(0.045)	(0.037)
Log(mean census)		−0.047***	−0.060***
		(0.0048)	(0.0051)
N	44,571	44,571	44,571
R^2	0.016	0.064	0.223
Year FE	No	Yes	Yes
CBSA 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. *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 CBSA in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table 5: Pooled-Moment CBSA-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	0.606	0.055	0.373	0.085	0.686	204	—	—	—	—
2019	0.587	0.043	0.379	0.115	0.787	222	-0.27	0.786	-0.27	0.786
2020	0.455	0.040	0.247	0.022	0.583	267	-2.23	0.026**	-2.25**	0.024
2021	0.451	0.036	0.221	0.015	0.522	335	-2.36	0.018**	-0.06	0.952
2022	0.448	0.037	0.252	-0.019	0.622	324	-2.38	0.017**	-0.05	0.956
2023	0.515	0.032	0.369	0.121	0.722	327	-1.44	0.149	1.35	0.176
2024	0.650	0.037	0.479	0.203	0.848	306	0.66	0.510	2.78***	0.005

Notes: $\hat{\alpha}_{ct}$ is the pooled-moment CBSA-year frictional markup. For the annual series in this table, values above 1.5 are top-winsorized. *Mean* is hour-weighted across CBSA-years by total employee RN hours in the CBSA-year; *SE* is the bootstrap standard error of that hour-weighted mean (500 resamples with replacement). *Median, P25, P75* are unweighted percentiles across CBSA-years (each CBSA-year counted equally). Tests vs. 2018 and year-on-year use two-sample z-tests on the hour-weighted annual means. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: CBSA Agency Share and Frictional Markup

	(1)	(2)	(3)	(4)
CBSA agency share	-0.8930^{***}	-1.0067^{***}	-1.1009^{***}	-2.1294^{***}
	(0.3225)	(0.2902)	(0.3282)	(0.2648)
log(CBSA RN hours)		0.0427^{*}	0.0477^{**}	0.0312
		(0.0219)	(0.0221)	(0.0206)
Year FE	No	No	Yes	Yes
CBSA FE	No	No	No	Yes
N	3,970	3,970	3,970	3,970
CBSAs	555	555	555	555
R^2	0.012	0.021	0.048	0.565

Notes: Dependent variable is the CBSA-year pooled-moment frictional markup $\hat{\alpha}_{ct}$, estimated by applying the two structural FOC moments directly to the pooled CBSA-day gap distribution. All specifications weighted by CBSA-year total employee RN hours. Standard errors clustered on CBSA in parentheses. CBSA agency share is the fraction of total annual RN hours supplied by agency staff in the CBSA-year. log(CBSA RN hours) is the log of total annual CBSA RN hours. To break the mechanical correlation between $\hat{\alpha}_{ct}$ and agency share, each CBSA-yearś facility-days are randomly split into two DOW-balanced halves and $\hat{\alpha}$ is estimated from one half while agency share is measured from the other, yielding two stacked cross-fit observations per CBSA-year. The regression sample therefore contains two rows per CBSA-year, with the LHS and RHS of each row drawn from non-overlapping days, so no within-split covariance between $\hat{\alpha}$ and share enters the OLS numerator. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7: Spot Market Welfare Decomposition

	2018	2019	2020	2021	2022	2023	2024
Per agency hour ($/hr, CBSA-year mean)							
PS / hr	201.46	253.70	219.05	295.06	319.63	385.25	399.41
Friction / hr	31.15	31.23	25.95	28.85	30.25	33.68	42.04
PS per $ of agency cost	2.19	2.71	2.39	2.88	3.02	3.63	3.39

Notes: Each column is one calendar year; entries are CBSA-year means weighted by CBSA-year total employee RN hours. The aggregate demand curve for each CBSA-year is the horizontal sum of facility-level inverse demands $P_i = \theta_i(\tilde{G}_i - A_i)$ over identified facilities with $\hat{r} > 1$ hr. The market price is $P^* = \bar{w}_a(1 + \hat{\alpha}_{ct})$, with $\hat{\alpha}_{ct}$ floored at 0 (so that negative-markup CBSA-years are treated as friction-free) and winsorized at 1.5. PS is the area above P^* under the demand curve. Friction waste $= (P^* - \bar{w}_a) \cdot A^*$. PS per $ of agency cost = $(\text{PS}/A^*)/P^*$.

Table 8: Resource Waste and Staffing: 2019 National Estimates

Panel A: Waste (annual, \$M, scaled to 15,000 national SNFs)

	Observed	p_{25} red. (%)	p_{50} red. (%)
Friction	\$269	62	20
Idle cost	\$539	46	25
Rationing	\$409	15	7
Total	\$1,216	39	18

Panel B: Staffing (hrs/facility/day)

	Observed	p_{25} change	p_{50} change
Buffer threshold H^b	-8.5	-4.5	-2.0
Expected idle hours	3.6	-1.6	-0.9
Expected agency hours	2.9	+3.4	+1.4
Agency (% change)	—	+117%	+46%

Notes: 2019 estimates. Observed mean $\hat{\alpha} = 0.534$. Identified facilities: 7,546; national scaling to 15,000 SNFs. All quantities computed from the empirical daily gap distribution for each facility-year, using the structural FOC with estimated $\hat{\theta}$. Panel A reports annual waste in \$M: Friction = $[w_a(1{+}\alpha){-}w_a]{\times}A_t$; Idle = $mc_e{\times}\max(H^b{-}\text{gap}_t, 0)$; Rationing = $(\hat{\theta}/2){\times}[\min(\max(\text{gap}_t{-}H^b, 0), r)]^2$. Counterfactual columns show percentage reduction relative to observed. Panel B reports staffing levels (observed) and changes (counterfactual minus observed). H^b is the FOC-implied buffer relative to observed employee staffing; a negative value means the model recommends total permanent hours $(emp + H^b)$ below current employee hours. In Panel B, negative entries in the change columns mean the quantity falls under the counterfactual.

Table 9: Marginal External Benefit Estimates

Year	s_{ct} agency share	$\lvert\hat{\beta}\rvert \cdot s_{ct}$ gross rate	$\frac{n_{\text{all}}-1}{n_{\text{all}}}$ uncaptured share	Marginal external benefit net of private share
2018	0.0356	7.58%	0.964	6.45%
2019	0.0376	8.00%	0.959	6.76%
2020	0.0436	9.28%	0.960	7.84%
2021	0.0629	13.39%	0.957	11.19%
2022	0.0988	21.03%	0.949	17.25%
2023	0.1060	22.57%	0.947	18.43%
2024	0.0863	18.38%	0.948	14.98%
Mean	0.0672	14.32%	0.956	11.84%

Notes: All quantities are expressed as a fraction of the engagement wage $w_a \cdot h_{\text{eng}}$; the engagement length h_{eng} and the spot wage w_a cancel out of the rate. s_{ct} is the national HCRIS S-3 Part V agency share of RN hours (contract over employee plus contract) by year, the same series plotted in Figure 2. $\lvert\hat{\beta}\rvert \cdot s_{ct}$ is the resource-saving externality rate from (**??**) ($\hat{\beta} = -2.129$ from Table 6 column (4)). The marginal external benefit is the uncaptured share $(n_{\text{all}} - 1)/n_{\text{all}}$ of the gross rate; the remaining $1/n_{\text{all}}$ accrues to the focal facility itself.

A Conditions for the Interior Equilibrium

The slack ratio in state s is $R_s(\theta^*) = L(\theta^*)/[\delta_s F(\theta^*)]$, with $L(\theta^*) = N - 1 + F(\theta^*)$. The interior-equilibrium inequalities $R_H < 1$ (high state tight) and $R_L \geq 1$ (low state slack) reduce to

$$\frac{1-N}{1-\delta_L} \ \leq \ F(\theta^*) \ < \ \frac{1-N}{1-\delta_H},$$

a nonempty interval since $\delta_H > \delta_L$, with the corresponding θ^* values lying in $[0,1]$ when $N \geq \delta_L$. Denote the two boundaries by

$$\theta^*_{\text{lo}} \equiv F^{-1}\big((1-N)/(1-\delta_L)\big) \quad (R_L = 1), \qquad \theta^*_{\text{hi}} \equiv F^{-1}\big((1-N)/(1-\delta_H)\big) \quad (R_H = 1).$$

The scarcity-wage formula $w_H = \gamma_H \theta^H_m$ results in an equilibrium wage only when $\gamma_H \theta^H_m > b$; otherwise the wage floor binds at b and only positions with $\theta \geq b/\gamma_H$ hire. Slack-state participation requires $\theta^* > b/\gamma_L(\theta^*)$, so that some low-state spot hiring occurs. The participation-valid region is

$$\mathcal{R}^+ \ \equiv \ \big\{\theta^* : \ R_H(\theta^*) < 1, \ R_L(\theta^*) \geq 1, \ \theta^* > b/\gamma_L(\theta^*), \ \gamma_H(\theta^*)\theta^H_m(\theta^*) > b\big\}\,.$$

The main analysis is conditional on an interior fixed point in $\mathcal{R}^+$. Approaching the upper boundary $\theta^* \uparrow \theta^*_{\text{hi}}$, the rationing condition $\delta_H\big[F(\theta^*) - F(\theta^H_m)\big] = L(\theta^*)$ gives

$$F(\theta^H_m(\theta^*)) \ = \ F(\theta^*) - \frac{L(\theta^*)}{\delta_H} \ \longrightarrow \ 0,$$

so $\theta^H_m \to 0$ and $\gamma_H \theta^H_m \to 0 < b$: $\mathcal{R}^+$ terminates strictly before θ^*_{hi}.

For use throughout the appendix, write $\theta^I = \mathcal{I}/\mathcal{D}$ with

$$\mathcal{I}(\theta^*) \equiv \pi_H(1-\delta_H)w_H(\theta^*) + \pi_L(1-\delta_L)b, \qquad \mathcal{D}(\theta^*) \equiv \pi_H\delta_H(1-\gamma_H(\theta^*)) + \pi_L\delta_L(1-\gamma_L(\theta^*)),$$

the expected idle cost and expected friction cost.

Cutoff structure. Fix the spot-market objects at their equilibrium values. For a non-rationed position of productivity θ, the profit differential between buffering and using the spot market is

$$\Pi^B(\theta) - \Pi^S(\theta) \;=\; \mathcal{D}(\theta^*)\,\theta - \mathcal{I}(\theta^*),$$

linear and strictly increasing in θ since $\mathcal{D}(\theta^*) > 0$. The unique indifference point is $\theta^* = \mathcal{I}/\mathcal{D}$, so positions with $\theta \geq \theta^*$ buffer and positions with $\theta < \theta^*$ use spot.

B Existence of Equilibrium

Lemma 1. *Let $\mathcal{R}^+$ be the participation-valid region of Appendix A, and let $g_{\text{eq}}(\theta^*) \equiv \theta^I(\theta^*) - \theta^*$. Suppose there exists a closed interval $[\underline{\theta}, \bar{\theta}] \subset \mathcal{R}^+$ such that $g_{\text{eq}}(\underline{\theta}) > 0$ and $g_{\text{eq}}(\bar{\theta}) < 0$. Then there exists an equilibrium cutoff $\theta^* \in (\underline{\theta}, \bar{\theta})$ satisfying $\theta^* = \theta^I(\theta^*)$. If additionally $(\theta^I)'(\theta^*) < 1$ at this point, the fixed point is locally stable.*

Proof. On $\mathcal{R}^+$ the spot pool $L = N - 1 + F(\theta^*)$, the slack ratios $R_s = L/(\delta_s F(\theta^*))$, the matching efficiencies $\gamma_s = g(L)h(R_s)$, and the marginal hire $\theta^H_m = F^{-1}((\delta_H - 1)F(\theta^*) + 1 - N)/\delta_H)$ are continuous in θ^* (using continuity of F, F^{-1}, g, h), so $\mathcal{I}$, $\mathcal{D}$, and $\theta^I = \mathcal{I}/\mathcal{D}$ are continuous on $\mathcal{R}^+$. Hence g_{eq} is continuous on $[\underline{\theta}, \bar{\theta}]$, and IVT yields a zero $\theta^* \in (\underline{\theta}, \bar{\theta})$ with $\theta^I(\theta^*) = \theta^*$. Under the continuous adjustment dynamic $\dot{\theta} = \theta^I(\theta) - \theta$, a crossing from above is locally stable when $(\theta^I)'(\theta^*) < 1$. □

The lemma localizes the existence argument to the participation-valid region $\mathcal{R}^+$, which is the only region used in the main text.

C Derivatives of the Spot Market Friction

Each $\gamma_s = g(L) \cdot h(R_s)$ depends on θ^* through $L = N - 1 + F(\theta^*)$ and $R_s = L/(\delta_s F(\theta^*))$. Using $dL/d\theta^* = f(\theta^*)$ and equation (2), 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 is the *density channel* ($g' > 0$); the second is the *congestion-relief channel* ($h' > 0$, stronger in state H because $R_H < R_L$ and h is concave).

Differentiating the tight-state clearing condition $\delta_H(F(\theta^*) - F(\theta_m)) = L$ with respect to θ^* gives $d\theta_m/d\theta^* = (\delta_H - 1)f(\theta^*)/(\delta_H f(\theta_m)) < 0$: each position switching to spot frees one worker but adds only $\delta_H < 1$ units of expected demand, so the rationing threshold falls.

D Proof of Excess Buffering Result

Proof. Differentiate $W(\theta^*) = Y(\theta^*) + u(\theta^*) \cdot b$ with respect to θ^*. Write $z_L(\theta^*) \equiv \theta^L_{\min}(\theta^*) = b/\gamma_L(\theta^*)$ for the slack-state outside-option cutoff. From equation (6),

$$u'(\theta^*) \;=\; \pi_L(1-\delta_L)f(\theta^*) \;+\; \pi_L\delta_L f(z_L)\,z_L'(\theta^*),$$

where the second term reflects that z_L depends on θ^* through γ_L. For $Y'(\theta^*)$, apply the Leibniz rule to each integral in (7), using $d\gamma_s/d\theta^* = f(\theta^*)\Gamma_s$ from (28) and the identity $d\theta_m/d\theta^* = (\delta_H - 1)f(\theta^*)/(\delta_H f(\theta_m))$ (derived by differentiating the tight-state clearing condition). The slack-state integral has a moving lower limit $z_L(\theta^*)$, contributing a Leibniz term $-\pi_L\delta_L\gamma_L\,z_L\,f(z_L) \cdot z_L'(\theta^*)$. Since $\gamma_L z_L = b$, this term equals $-\pi_L\delta_L\,b\,f(z_L) \cdot z_L'(\theta^*)$, which exactly cancels the matching $\pi_L\delta_L b f(z_L)\,z_L'(\theta^*)$

component of $b\,u'(\theta^*)$. Collecting the surviving terms:

$$Y'(\theta^*) = f(\theta^*)\Big\{ -\bar{\delta}\,\theta^* + \pi_H\delta_H\Gamma_H Q_H + \pi_H\delta_H\gamma_H\theta^* - \pi_H\delta_H\gamma_H \cdot \frac{(\delta_H - 1)\theta_m}{\delta_H} + \pi_L\delta_L\Gamma_L Q_L + \pi_L\delta_L\gamma_L\theta^*\Big\}, \tag{29}$$

where $Q_H = \int_{\theta_m}^{\theta^*} \theta\, dF$ and $Q_L = \int_{\theta^L_{\min}}^{\theta^*} \theta\, dF$ is the aggregate productivity of served spot positions in state L, as in (12). 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 the surviving non-employment contribution $\pi_L(1-\delta_L)b\cdot f(\theta^*)$ from $b\,u'(\theta^*)$ (the $\pi_L\delta_L b f(z_L)\,z_L'$ component already cancelled against the slack-state Leibniz term in Y'):

$$W'(\theta^*) = f(\theta^*)\Big\{ -\underbrace{\big[\pi_H\delta_H(1 - \gamma_H) + \pi_L\delta_L(1 - \gamma_L)\big]}_{\mathcal{D}}\,\theta^* + \underbrace{\pi_H(1 - \delta_H)w_H + \pi_L(1 - \delta_L)b}_{\mathcal{I}} + \pi_H\delta_H\,\Gamma_H\,Q_H + \pi_L\delta_L\,\Gamma_L\,Q_L\Big\}. \tag{30}$$

At any equilibrium, $\theta^* = \mathcal{I}/\mathcal{D}$ by (10), so the first two groups cancel:

$$W'(\theta^*) = f(\theta^*)\big[\pi_H\delta_H\,\Gamma_H\,Q_H + \pi_L\delta_L\,\Gamma_L\,Q_L\big]. \tag{31}$$

All factors are positive at any interior equilibrium ($f > 0$ by full support, $\Gamma_s > 0$ since $g', h' > 0$, and $Q_s > 0$), so $W'(\theta^*) > 0$: a marginal reduction in buffering at the competitive equilibrium raises welfare. A global comparison $\theta^{**} > \theta^*$ requires additional structure on W (e.g. single-peakedness) not established here.

If $g' = h' = 0$, both Γ_s vanish and $W'(\theta^*) = 0$: the inefficiency disappears when match quality is independent of the pool. □

E Comparative Static in δ_H

We derive a necessary-and-sufficient condition for θ^I to rise in δ_H, holding θ^*, π_s, δ_L, and b fixed, then reduce it to a sufficient condition with a direct economic reading and state primitive-based assumptions under which it holds. The analysis is conditional on the participation-valid region $\mathcal{R}^+$ (Appendix A), on which $w_H = \gamma_H \theta_m^H > b$ is the equilibrium high-state spot wage. At the equilibrium fixed point $\theta^* = \theta^I(\theta^*)$, the implicit function theorem converts $\partial\theta^I/\partial\delta_H$ into $d\theta^*/d\delta_H$.

Write $\theta^I = \mathcal{I}/\mathcal{D}$ with $\mathcal{I} = \pi_H(1-\delta_H)\, w_H + \pi_L(1-\delta_L)\, b$ and $\mathcal{D} = \pi_H\delta_H(1-\gamma_H) + \pi_L\delta_L(1-\gamma_L)$. The spot pool $L = N - 1 + F(\theta^*)$ does not depend on δ_H.

Since $R_H = L/[\delta_H F(\theta^*)]$ with L fixed, $\partial_{\delta_H} R_H = -R_H/\delta_H$, and $\gamma_H = g(L)\, h(R_H)$ gives

$$\partial_{\delta_H}\gamma_H = -\phi/\delta_H, \qquad \phi \equiv g(L)\, h'(R_H)\, R_H > 0, \tag{32}$$

so higher δ_H activates more positions for a fixed worker pool and congestion worsens. Differentiating the H-clearing condition $\delta_H[F(\theta^*) - F(\theta_m^H)] = L$,

$$\partial_{\delta_H}\theta_m^H = \frac{F(\theta^*) - F(\theta_m^H)}{\delta_H f(\theta_m^H)} = \frac{L}{\delta_H^2 f(\theta_m^H)} > 0, \tag{33}$$

so the marginal served position becomes more productive. With $w_H = \gamma_H\theta_m^H$,

$$\partial_{\delta_H} w_H = \frac{1}{\delta_H}\left[\gamma_H \frac{L}{\delta_H f(\theta_m^H)} - \phi\, \theta_m^H\right], \tag{34}$$

combining rationing tightening (the marginal hire is more productive) with congestion (match quality falls). Defining $\eta_w \equiv (\delta_H/w_H)\,\partial_{\delta_H} w_H$ and using $w_H = \gamma_H\theta_m^H$ together with $L/\delta_H = F(\theta^*) - F(\theta_m^H)$ from clearing,

$$\eta_w = \kappa - \varepsilon_h, \qquad \kappa \equiv \frac{F(\theta^*) - F(\theta_m^H)}{f(\theta_m^H)\,\theta_m^H}, \quad \varepsilon_h \equiv \frac{R_H h'(R_H)}{h(R_H)}. \tag{35}$$

κ is the served-mass-to-local-density ratio: mass served in the H state divided by the density of F at the cutoff times the cutoff productivity. ε_h is the elasticity of h at R_H.

For the denominator, $\partial_{\delta_H}\mathcal{D} = \pi_H(1-\gamma_H) - \pi_H\delta_H\partial_{\delta_H}\gamma_H = \pi_H\big[(1-\gamma_H)+\phi\big] > 0$: both the direct weight on the H-state friction and the friction itself rise. Defining the H-state shares $\sigma_I \equiv \pi_H(1-\delta_H)w_H/\mathcal{I}$ and $\sigma_D \equiv \pi_H\delta_H(1-\gamma_H)/\mathcal{D}$, both in $(0,1)$, differentiation of $\ln\mathcal{I}$ and $\ln\mathcal{D}$ yields

$$\partial_{\delta_H}\ln\mathcal{I} = \frac{\sigma_I}{\delta_H}\left[\eta_w - \frac{\delta_H}{1-\delta_H}\right], \qquad \partial_{\delta_H}\ln\mathcal{D} = \frac{\sigma_D}{\delta_H}\left[1+\frac{\phi}{1-\gamma_H}\right].$$

Since $\partial_{\delta_H}\ln\theta^I = \partial_{\delta_H}\ln\mathcal{I} - \partial_{\delta_H}\ln\mathcal{D}$, the comparative static is positive if and only if $\eta_w > \delta_H/(1-\delta_H) + (\sigma_D/\sigma_I)\big[1+\phi/(1-\gamma_H)\big]$. Substituting (35),

$$\frac{\partial\theta^I}{\partial\delta_H} > 0 \iff \kappa > \varepsilon_h + \frac{\delta_H}{1-\delta_H} + \frac{\sigma_D}{\sigma_I}\left(1+\frac{\phi}{1-\gamma_H}\right). \tag{36}$$

At the equilibrium fixed point, $\mathcal{I} = \mathcal{D}\theta^*$, so $\sigma_D/\sigma_I = \frac{\delta_H(1-\gamma_H)}{(1-\delta_H)\gamma_H}\cdot\frac{\theta^*}{\theta_m^H}$. Using $\phi = \gamma_H\varepsilon_h$, condition (36) simplifies to

$$\kappa > \varepsilon_h + \frac{\delta_H}{1-\delta_H}\left[1+\frac{\theta^*}{\theta_m^H}\left(\frac{1-\gamma_H}{\gamma_H}+\varepsilon_h\right)\right], \tag{37}$$

which makes the buffer-tilt force on the right-hand side scale with the rationing severity θ^*/θ_m^H and the friction-to-efficiency ratio $(1-\gamma_H)/\gamma_H$.

Concavity of h with $h(0) \geq 0$ gives $\varepsilon_h \leq 1$, so $(1-\gamma_H)/\gamma_H + \varepsilon_h \leq 1/\gamma_H$, and writing $\rho \equiv \theta_m^H/\theta^* \in (0,1]$ for rationing severity, the right-hand side of (37) is bounded above by $1 + \frac{\delta_H}{1-\delta_H}(1+1/(\rho\gamma_H))$. A sufficient condition for $\partial\theta^I/\partial\delta_H > 0$ is therefore

$$\kappa > 1 + \frac{\delta_H}{1-\delta_H}\left(1+\frac{1}{\rho\,\gamma_H}\right). \tag{38}$$

The rationing-relief elasticity κ must exceed unity plus a buffer-tilt term composed of three forces: $\delta_H/(1-\delta_H)$ is the mechanical shrinkage of the idle-weight $\pi_H(1-\delta_H)$ in $\mathcal{I}$ (less precautionary value remains when H-state activation is more likely conditional on H); $1/\rho$ amplifies the denominator's δ_H-loading when rationing is severe (smaller ρ means the rationing-relief channel for w_H operates over a larger productivity gap); and $1/\gamma_H$ raises the denominator's response when matching efficiency is low.

The two equilibrium objects in (38), κ and $\rho\gamma_H$, can be bounded using primitives plus an interiority assumption. With $L \geq L_{\min} \equiv (1-N)\delta_L/(1-\delta_L)$ from L-slack and a density bound $f(\theta)\,\theta \leq \bar{f}$ on $[0, \theta^*_{\text{hi}}]$ (with θ^*_{hi} from Appendix A),

$$\kappa = \frac{L/\delta_H}{f(\theta^H_m)\,\theta^H_m} \geq \frac{L_{\min}}{\delta_H \bar{f}}.$$

Assuming $R_H \leq \underline{R} < 1$ at the equilibrium gives $F(\theta^H_m) = F(\theta^*)(1-R_H) \geq F_{\text{lo}}(1-\underline{R})$ with $F_{\text{lo}} = (1-N)/(1-\delta_L)$, so $\rho \geq \rho_{\min} \equiv F^{-1}\big(F_{\text{lo}}(1-\underline{R})\big)/\theta^*_{\text{hi}}$. The lower bound $R_H \geq \delta_L(1-\delta_H)/[\delta_H(1-\delta_L)]$ and monotonicity of g, h give $\gamma_H \geq \gamma_{\min} \equiv g(L_{\min})\, h\big(\delta_L(1-\delta_H)/[\delta_H(1-\delta_L)]\big)$. Substituting these into (38),

$$\frac{L_{\min}}{\delta_H \bar{f}} > 1 + \frac{\delta_H}{1-\delta_H}\left(1 + \frac{1}{\rho_{\min}\,\gamma_{\min}}\right) \tag{39}$$

implies $\partial\theta^I/\partial\delta_H > 0$, with every quantity in (39) a function of primitives $(F, g, h, \pi_s, \delta_s, b, N)$ and the bound $\underline{R}$.

The condition compares two elasticities. The rationing-relief elasticity κ is large when the density f is thin at the marginal hire's productivity θ^H_m, because the cutoff must travel a long productivity distance per unit of additional served mass; raising δ_H then produces a large jump in $w_H = \gamma_H\theta^H_m$, raises the idle cost of buffering, and pushes the equilibrium toward spot. The congestion elasticity ε_h works against this: more activation spreads the worker pool more thinly, lowers R_H, degrades match quality, drags w_H back down, and raises the friction $1-\gamma_H$ in $\mathcal{D}$. Even when rationing dominates congestion ($\kappa > \varepsilon_h$), two further forces oppose the spot shift. The first, $\delta_H/(1-\delta_H)$, is the mechanical shrinkage of the idle-weight $\pi_H(1-\delta_H)$ on w_H in $\mathcal{I}$: buffering is precautionary, and there is less to be precautionary about when H-state activation is near-certain. The second, $(\sigma_D/\sigma_I)\big[1+\phi/(1-\gamma_H)\big]$, captures the rise in $\mathcal{D}$ from the direct δ_H-loading on H-state frictions and the further worsening of $1-\gamma_H$. Both push toward more buffering.

Condition (37) is most likely to hold when the marginal H-state hire sits in a thin region of F (giving large κ), congestion is mildly elastic at the prevailing R_H (giving small ε_h), δ_H is bounded away from 1 so a buffer worker still faces a meaningful chance of being idle in the H state (giving small $\delta_H/(1-\delta_H)$), and the H-state's contribution to the buffer-spot trade-off is large on the numerator side relative to the denominator side (giving large σ_I/σ_D). The first two describe a thin-market equilibrium where the spot pool is small and the cutoff sits in a sparse region of the productivity distribution; the third rules out near-certain H-state activation; the fourth holds when the scarcity wage premium w_H/b is large.

F Proof of Demand Frequency Result

Proof. Write $\theta^I = \mathcal{I}/\mathcal{D}$ and differentiate in π_H at fixed $\theta^*, w_H, b, \delta_s, \gamma_s$, using $\pi_L = 1-\pi_H$:

$$A \;\equiv\; \partial_{\pi_H}\mathcal{I} \;=\; (1-\delta_H)w_H - (1-\delta_L)b, \qquad B \;\equiv\; \partial_{\pi_H}\mathcal{D} \;=\; \delta_H(1-\gamma_H) - \delta_L(1-\gamma_L).$$

Since $\mathcal{D} > 0$, $\operatorname{sign}(\partial\theta^I/\partial\pi_H) = \operatorname{sign}(A\mathcal{D} - B\mathcal{I})$. Set $u_s \equiv \delta_s(1-\gamma_s)$ and $v_s \equiv 1-\delta_s$, so $\mathcal{D} = \pi_H u_H + \pi_L u_L$, $\mathcal{I} = \pi_H v_H w_H + \pi_L v_L b$, $A = v_H w_H - v_L b$, $B = u_H - u_L$. Expanding both products,

$$A\mathcal{D} = \pi_H u_H v_H w_H + \pi_L u_L v_H w_H - \pi_H u_H v_L b - \pi_L u_L v_L b,$$

$$B\mathcal{I} = \pi_H u_H v_H w_H + \pi_L u_H v_L b - \pi_H u_L v_H w_H - \pi_L u_L v_L b.$$

The terms $\pi_H u_H v_H w_H$ and $\pi_L u_L v_L b$ appear in both with the same sign and cancel. Collecting the rest and using $\pi_H + \pi_L = 1$,

$$A\mathcal{D} - B\mathcal{I} \;=\; u_L v_H w_H - u_H v_L b \;=\; \delta_L(1-\delta_H)(1-\gamma_L)\, w_H \;-\; \delta_H(1-\delta_L)(1-\gamma_H)\, b.$$

All factors are strictly positive, so for $b > 0$,

$$\frac{\partial \theta^I}{\partial \pi_H} \;>\; 0 \;\;\iff\;\; \frac{w_H}{b} \;>\; \frac{\delta_H(1-\delta_L)(1-\gamma_H)}{\delta_L(1-\delta_H)(1-\gamma_L)}. \qquad \square$$

G Proof of Thin-Market Comparative Statics

Proof. The scarcity wage is $w_H(\theta^*) = \gamma_H(\theta^*)\, \theta_m^H(\theta^*)$. Differentiating by the product rule,

$$\frac{dw_H}{d\theta^*} \;=\; \frac{d\gamma_H}{d\theta^*}\, \theta_m^H \;+\; \gamma_H\, \frac{d\theta_m^H}{d\theta^*}. \tag{40}$$

By (28), $d\gamma_H/d\theta^* \;=\; f(\theta^*)\, \Gamma_H$. To compute $d\theta_m^H/d\theta^*$, differentiate the tight-state clearing condition $\delta_H[F(\theta^*) - F(\theta_m^H)] = L(\theta^*) = N - 1 + F(\theta^*)$ in θ^*:

$$\delta_H\, f(\theta^*) - \delta_H\, f(\theta_m^H)\, \frac{d\theta_m^H}{d\theta^*} \;=\; f(\theta^*) \;\Longrightarrow\; \frac{d\theta_m^H}{d\theta^*} \;=\; -\frac{(1-\delta_H)\, f(\theta^*)}{\delta_H\, f(\theta_m^H)} \;<\; 0.$$

Substituting both expressions into (40),

$$\frac{dw_H}{d\theta^*} \;=\; f(\theta^*) \left[\; \underbrace{\Gamma_H(\theta^*)\, \theta_m^H(\theta^*)}_{\text{quality-improvement channel}} \;\;-\;\; \underbrace{\frac{\gamma_H(\theta^*)(1-\delta_H)}{\delta_H\, f(\theta_m^H(\theta^*))}}_{\text{rationing-relief channel}} \right]. \tag{41}$$

Since $f(\theta^*) > 0$ on the interior of $[0,1]$ by the full-support assumption, the sign of $dw_H/d\theta^*$ equals the sign of the bracket, and $dw_H/d\theta^* > 0$ if and only if

$$\Gamma_H(\theta^*)\,\theta_m^H(\theta^*) \;>\; \frac{\gamma_H(\theta^*)\,(1-\delta_H)}{\delta_H\, f(\theta_m^H(\theta^*))}. \tag{42}$$

Boundary behavior. Plugging $R_L = 1$ into the H-clearing condition gives $F(\theta_m^H(\theta_{\text{lo}}^*)) = F(\theta_{\text{lo}}^*)(\delta_H - \delta_L)/\delta_H > 0$, so θ_m^H is bounded away from 0 at the lower boundary. As $\theta^* \uparrow \theta_{\text{hi}}^*$, $\theta_m^H \to 0$: in (42) the LHS $\Gamma_H\theta_m^H \to 0$ (Γ_H stays bounded as $F(\theta_{\text{hi}}^*) > 0$) while the RHS converges to $\gamma_H(\theta_{\text{hi}}^*)(1-\delta_H)/(\delta_H f(0)) > 0$, so (42) fails. Under the maintained participation condition $\gamma_H\theta_m^H > b$ (Appendix A), this limit is not approached: $\mathcal{R}^+$ terminates strictly before θ_{hi}^*, and the wage-sign condition need only be checked on $\mathcal{R}^+$.

Thin-market region. Both sides of (42) are continuous in θ^* on $(\theta_{\text{lo}}^*, \theta_{\text{hi}}^*)$. Let $\underline{\theta}$ be the largest $\theta^{*\circ}$ such that (42) holds throughout $(\theta_{\text{lo}}^*, \theta^{*\circ})$; the thin-market region $(\theta_{\text{lo}}^*, \underline{\theta})$ is non-empty iff (42) holds at θ_{lo}^*. On this region, $dw_H/d\theta^* > 0$ by (41).

Statement (iv). By Appendix F, (13) gives $\partial\theta^I/\partial\pi_H > 0$. The implicit function theorem at a stable equilibrium θ_1^* (where $\partial\theta^I/\partial\theta^* < 1$) yields

$$\left.\frac{d\theta^*}{d\pi_H}\right|_{\theta_1^*} = \frac{\partial\theta^I/\partial\pi_H}{1 - \partial\theta^I/\partial\theta^*} > 0.$$

On the thin-market region $dw_H/d\theta^* > 0$, so $dw_H/d\pi_H > 0$ by the chain rule. For the buffer wage $w = \pi_H w_H + (1-\pi_H)b$,

$$\frac{dw}{d\pi_H} = (w_H - b) + \pi_H\,\frac{dw_H}{d\pi_H} > 0,$$

since $w_H = \gamma_H\theta_m^H > b$ in the interior region (else the marginal hire would be unprofitable).

Statement (v). The premium is $w_H - w = \pi_L(w_H - b)$. Differentiating in π_H,

$$\frac{d(w_H - w)}{d\pi_H} = -(w_H - b) + \pi_L \frac{dw_H}{d\pi_H}, \tag{43}$$

a composition effect (the π_L weight shrinks) against a level effect (the scarcity wage rises). The premium widens iff

$$\pi_L \frac{dw_H}{d\pi_H} > w_H - b. \tag{44}$$

$\square$

Appendix: Estimation Details

I. Between-Within Decomposition of RN Wage Growth

Following Foster et al. (2001), the change in the hours-weighted average RN wage relative to 2019 is decomposed as:

$$\underbrace{\bar{w}_t - \bar{w}_0}_{\text{Total}} = \underbrace{\sum_i s_{i0}(w_{it} - w_{i0})}_{\text{Within}} + \underbrace{\sum_i (s_{it} - s_{i0})\, w_{i0}}_{\text{Between}} + \underbrace{\sum_i (s_{it} - s_{i0})(w_{it} - w_{i0})}_{\text{Cross}}, \quad (45)$$

where s_{it} is facility i's share of total national RN paid hours in year t and subscript 0 denotes 2019. Table 2 reports within and between estimates. When within exceeds 100 percent of total, the cross term is negative—facilities that gained hours tended to have below-average wage growth.

Wages and hours are from HCRIS Form CMS-2540, Worksheet S-3 Part V (RN-specific salary and paid hours, employee and contract separately). Panel A uses employee salary and hours only; Panel B adds contract salary and hours. Observations outside \$5–\$200/hr are dropped. The sample is a balanced panel of $N = 13{,}125$ SNFs observed in all years 2019–2023.

II. Stacked Difference-in-Differences

The event study in Figure 8 avoids the negative-weight problem arising in two-way fixed-effects estimators with heterogeneous and staggered treatment timing (Baker et al., 2022).

Sample construction. Treated cohorts are defined by the year in which a facility first experiences a discrete shift in its CMS staffing star rating: cohorts $g \in \{2020, 2021, 2022\}$ are retained. For each cohort g, I assemble a balanced sub-panel covering a ±2-year window $[g-2,\ g+2]$, consisting of (i) all treated facilities whose star rating shifted in year g, and (ii) all never-treated facilities—those whose star rating never changes across the full panel—as controls. Within each cohort-stack, only facilities observed in every year of the window are retained. The three sub-panels are then stacked into a single dataset, so each never-treated facility contributes observations to every cohort stack in which it is present.

Treatment Definition. The treatment for facility i in cohort g is the signed star-rating change:

$$D_{ig} = d_{ig} \in \{\ldots,\ -2,\ -1,\ +1,\ +2,\ \ldots\}, \tag{46}$$

where $d_{ig} > 0$ indicates an upgrade and $d_{ig} < 0$ indicates a downgrade. Never-treated facilities receive $D_{ig} = 0$. Each β_ℓ is therefore interpretable as the outcome response to a one-star change in the CMS staffing rating.

Estimation equation. Let Y_{igt} denote the outcome for facility i in cohort-stack g at calendar year t, and let $\ell = t - g$ denote relative event time. The estimating equation is:

$$Y_{igt} = \sum_{\ell=-2,\,\ell\neq-1}^{2} \beta_\ell \cdot D_{ig} \cdot \mathbf{1}[t-g=\ell] + \mu_{ig} + \lambda_{tg} + \varepsilon_{igt}, \tag{47}$$

where μ_{ig} are facility×cohort fixed effects and λ_{tg} are year×cohort fixed effects. The omitted period is $\ell = -1$ (the year immediately before the star-rating shift), so all β_ℓ are relative to the pre-event baseline. Fixed effects are removed by iterative demeaning. Standard errors are clustered at the facility level, the unit of treatment assignment. The sample spans $N = 40{,}315$ facility-cohort-year observations across $G = 6{,}275$ facilities (5,185 treated, 1,090 never-treated controls).

III. Split-Sample Cross-Fit Estimator for the Agency Share Regression

Both $\hat{\alpha}_{ct}$ and the CBSA agency share are built from the same facility-day records and both depend on contract hours: $\hat{\alpha}_{ct}$ uses contract hours to compute the residual shortfall r in (18) and the effective agency cost $w_a(1+\alpha)$ in (19), while agency share is simply aggregate contract hours divided by total hours. Using the same data for both variables could therefore produce a spurious relationship even absent any true market-thickening relationship.

To address this, I split each CBSA-year into two non-overlapping halves using a randomized, day-of-week-balanced assignment. Within each day-of-week cell, facility-days are randomly permuted and then assigned to halves alternately. Balancing on day of week prevents weekly scheduling cycles from creating systematic differences in the gap distribution across halves. Both halves must independently satisfy the identification conditions—at least the minimum number of agency-using days and an interior solution with $1+\hat{\alpha}>0$—or the CBSA-year is dropped entirely.

$\hat{\alpha}$ and agency share are then estimated from opposite halves. Specifically, $\hat{\alpha}_A$ is estimated from half A and the corresponding agency share s_B is measured from half B; $\hat{\alpha}_B$ is estimated from half B and s_A from half A. Since each direction pairs an $\hat{\alpha}$ estimate with an agency share computed from non-overlapping days, the mechanical channel through contract hours does not operate.

The agency share regression stacks both directions as two separate observations per CBSA-year: direction 1 uses $(\hat{\alpha}_A,\, s_B)$ and direction 2 uses $(\hat{\alpha}_B,\, s_A)$. By stacking, the OLS numerator contains only cross-split covariances $\text{Cov}(\hat{\alpha}_A, s_B)$ and $\text{Cov}(\hat{\alpha}_B, s_A)$, with no within-split terms. Each stacked observation receives half the CBSA-year employment weight so that the total weight mass per CBSA-year is preserved, and standard errors are clustered on CBSA to account for the within-CBSA-year pairing of the two direction rows.

Appendix: Data

I. Raw Data Sources

Payroll-Based Journal (PBJ). The primary data source is the CMS Payroll-Based Journal system, through which all Medicare- and Medicaid-certified skilled nursing facilities are required to submit daily staffing records. I use the PBJ Daily Nurse Staffing files, released as quarterly CSV extracts covering 2017Q4 through 2024Q4. Each record identifies a facility-date-occupation cell and reports total hours worked, hours attributable to employee (permanent) staff, hours attributable to contract (agency) staff, and the MDS patient census. Each quarterly extract contains approximately four million facility-day-occupation rows; across all quarters the raw data comprise roughly 112 million rows and cover between 15,000 and 21,000 distinct facilities per year.

HCRIS Medicare Cost Reports. Annual facility-level compensation data are drawn from the Healthcare Cost Report Information System (HCRIS), Form CMS-2540 (the freestanding SNF cost report), Worksheet S-3 Part V.[36] The key variables extracted are the employee RN hourly base wage (annual salary divided by paid hours), employee RN total hourly compensation (salary plus fringe benefits divided by paid hours), and the contract RN hourly rate. Values below \$5 or above \$200 per hour for employee wages, and below \$5 or above \$500 per hour for the contract RN hourly rate, are coded as missing on the grounds that they most plausibly reflect reporting errors rather than genuine wage observations.

CMS Star Ratings and Case-Mix Indices. The staffing-domain star rating (1–5 stars) and the RN case-mix index (CM_RN) for each facility-year are obtained

[36] Hospital-based SNFs, which file under Form CMS-2552 with the host hospital and account for roughly 3 percent of U.S. SNFs, are not included.

from the Nursing Home Compare (NHC) November snapshots, which are available for 2019 through 2023. For 2017 and 2018, I assign each facility its 2019 values; for 2024, I assign 2023 values to avoid the confound introduced by the PDPM methodology revision of July 2024. The national average CM_RN for each year is computed from the same NHC source. These inputs are used to construct the facility-year-specific regulatory minimum RN hours per patient-day:

$$h^*_{jt} = \text{RN\_thresh}[\text{star}_{jt}] \times \frac{\text{CM\_RN}_{jt}}{\overline{\text{CM\_RN}}_t}.$$

OMB CBSA Delineation. Facilities are assigned to labor markets using the 2020-vintage county-to-CBSA crosswalk distributed by the NBER which maps five-digit county FIPS codes to Core-Based Statistical Area codes. Counties that fall outside any defined CBSA—predominantly rural areas and territories—receive synthetic market codes so that they remain in the sample as their own markets rather than being dropped.

II. Sample Construction

The PBJ program releases staffing data as quarterly files, and records near quarter boundaries appear in both adjacent files. After concatenating all quarterly files, I identify duplicate records as rows sharing the same facility, county, date, and occupation and retain only one copy. This ensures that facility-days near quarter boundaries are counted once in all hours aggregations. The resulting panel covers 2018–2024 and contains approximately 112 million facility-day-occupation rows.

Structural estimation sample. The facility-year panel used for newsvendor estimation is assembled in several steps. Daily PBJ records are merged with the star ratings panel to construct the daily staffing need ($h^*_{jt} \times \text{census}_{jdt}$) and the daily gap ($G_{jdt} = \text{need}_{jdt} - \text{employee hours}_{jdt}$). The facility-day data are then collapsed to facility-year summaries of the empirical gap distribution, retaining only Monday-through-Friday observations and requiring at least 143 such weekdays per facility-year to ensure that the empirical distribution is well-characterized. The resulting facility-year panel is merged to HCRIS wages on a facility×year inner join, so that only observations with valid compensation records on both sides are retained.

Two further restrictions enforce the conditions needed for point identification of the structural parameters. First, I require each CBSA-year cell to contain at least two facilities reporting agency wages, so that the market-level agency wage $w_{a,ct}$ reflects more than a single reporter and is not mechanically determined by the facility's own cost report. Second, I impose an interior newsvendor solution by requiring that the CBSA agency wage strictly exceeds the facility's employee total compensation ($w_{a,ct} > \text{mc}_{e,jt}$); observations that fail this condition have a corner solution in which the facility would never use agency labor, and the first-order conditions used for identification do not apply. Over-buffered facilities (those whose employee staffing essentially never falls short of need) are retained in the sample and handled via θ imputation in the structural estimation. The final sample contains 56,381 facility-years drawn from 11,752 distinct facilities.

Reduced-form rationing analyses. For the reduced-form analysis of market thickness and rationing, the facility-day RN panel is collapsed to facility-week observations. Market-level regressors—leave-one-out agency share, log total RN hours, and log census—are computed at either the county-week or CBSA-week level depending on the specification. The leave-one-out construction removes the focal facility's own contract and total hours from the market aggregate before computing the agency share; this prevents the mechanical correlation that would otherwise arise between a facility's own staffing choices and the market-level regressor. The baseline agency share entering the market-thickness measure is computed from the first thirteen weeks of each market's observed data. Facilities with fewer than twenty weekly observations in total are excluded.

Appendix Figures

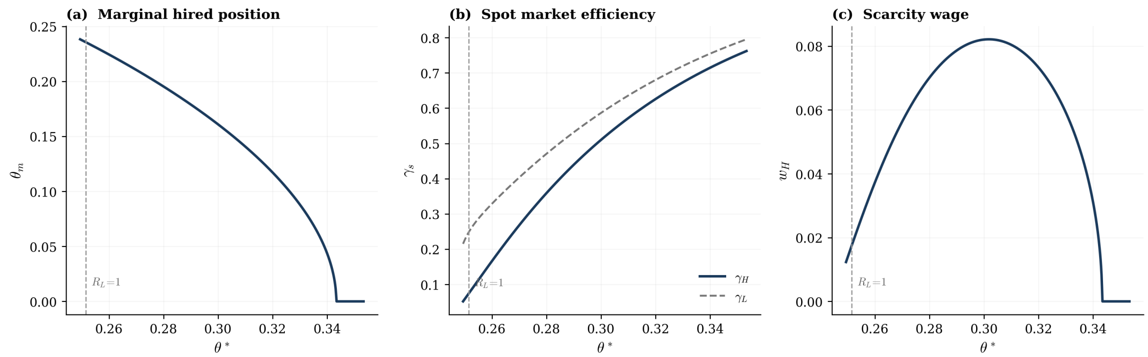


Figure A1: **Scarcity wage components.** Panel (a): marginal hired position θ_m as a function of θ^*. Panel (b): spot market efficiency $\gamma_s = g(L)h(R_s)$ for states H and L. Panel (c): scarcity wage w_H as a function of θ^*. Dashed vertical line marks the tight/slack boundary at $R_L = 1$. Parameters as in Figure 1.

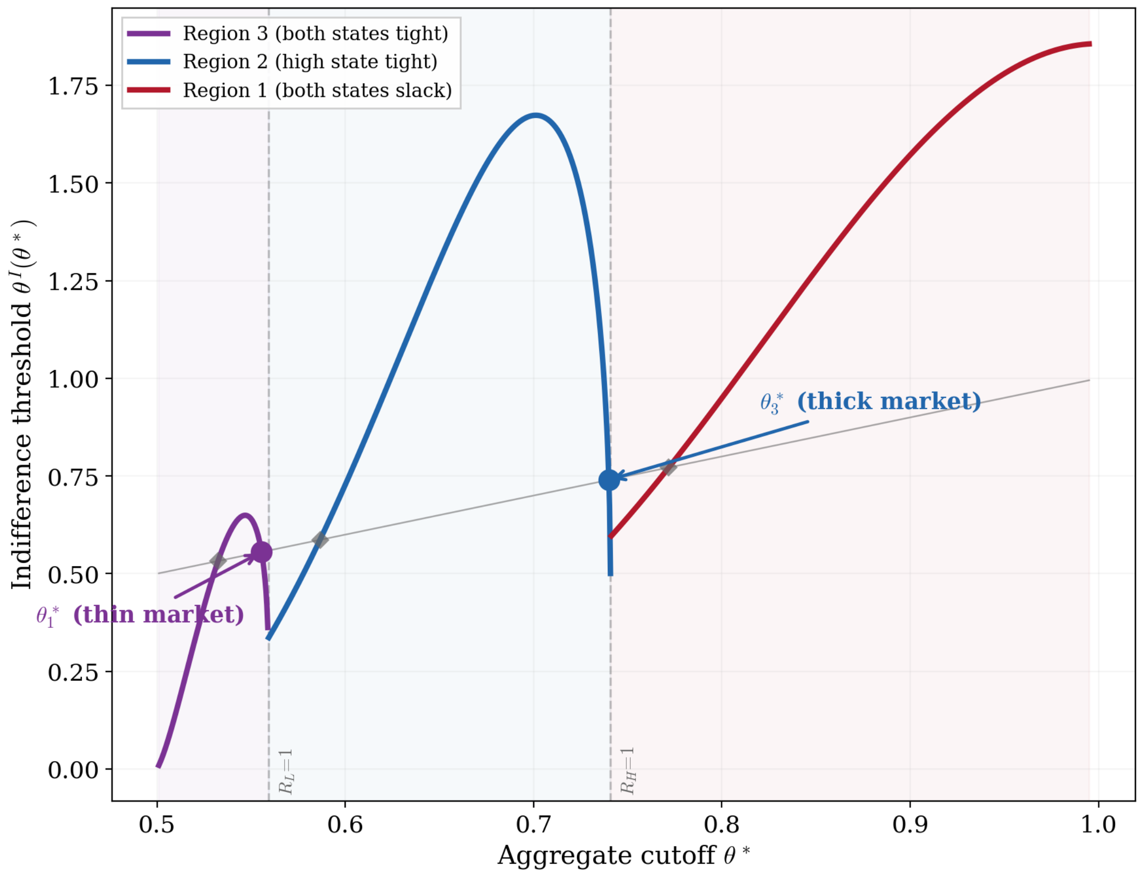


Figure A2: **Multiple equilibria.** Best-response curve $\theta^I(\theta^*)$ plotted across three configurations: both states tight (purple), high state tight only (blue), and both states slack (red). Shaded backgrounds indicate the three configurations; dashed vertical lines mark the boundaries $R_L = 1$ and $R_H = 1$. Filled circles: stable equilibria (θ_1^*, θ_3^*); open circle: unstable equilibrium. Parameters: $F = \text{Beta}(2, 2)$, $N = 0.50$, $\delta_H = 0.40$, $\delta_L = 0.15$, $\pi_H = 0.30$, $b = 0.01$.

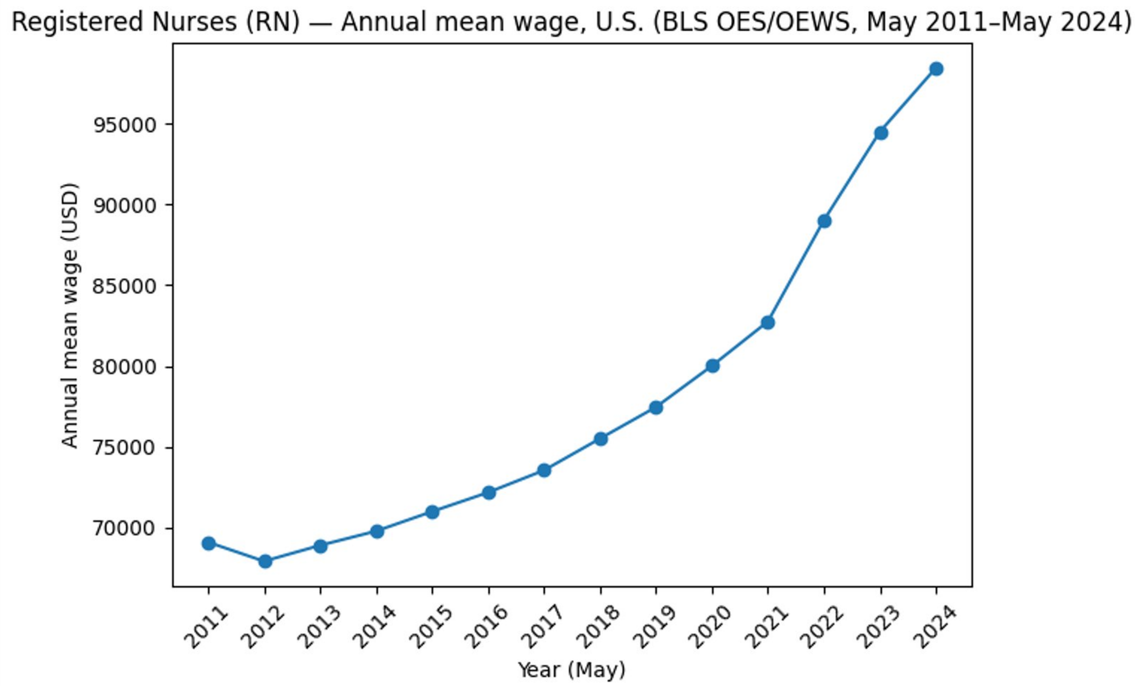


Figure A3: **RN wages, 2011–2024.** Annual mean wage for registered nurses (RN), U.S., BLS Occupational Employment and Wage Statistics (OES/OEWS), May 2011–May 2024.

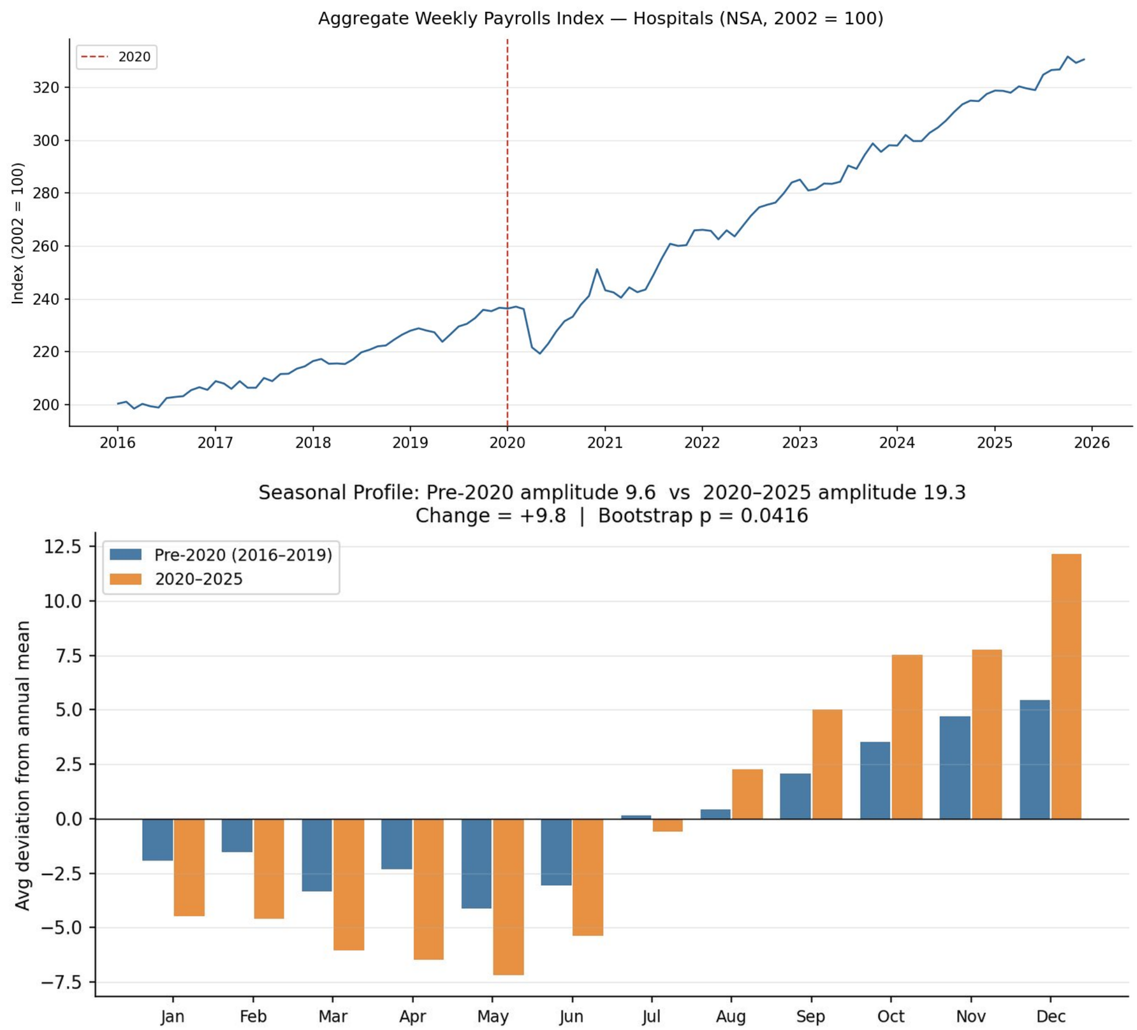


Figure A4: **Hospital employment after 2020.** Top panel: aggregate weekly payroll index, hospitals, BLS CES (not seasonally adjusted, $2002 = 100$), 2016–2026. Red dashed line marks 2020. Bottom panel: seasonal profile of the payroll index, pre-2020 (2016–2019) vs. 2020–2025. Pre-2020 amplitude is 9.6; post-2020 amplitude is 19.3 (change = +9.8, bootstrap $p = 0.0416$).

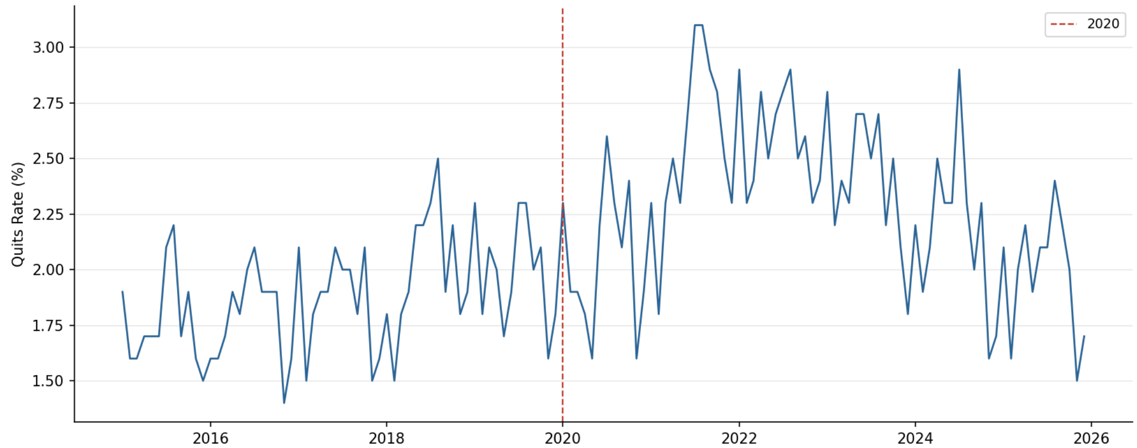


Figure A5: **Quits after 2020.** Monthly quit rate for health care and social assistance, JOLTS (not seasonally adjusted), 2014–2026. Red dashed line marks 2020.

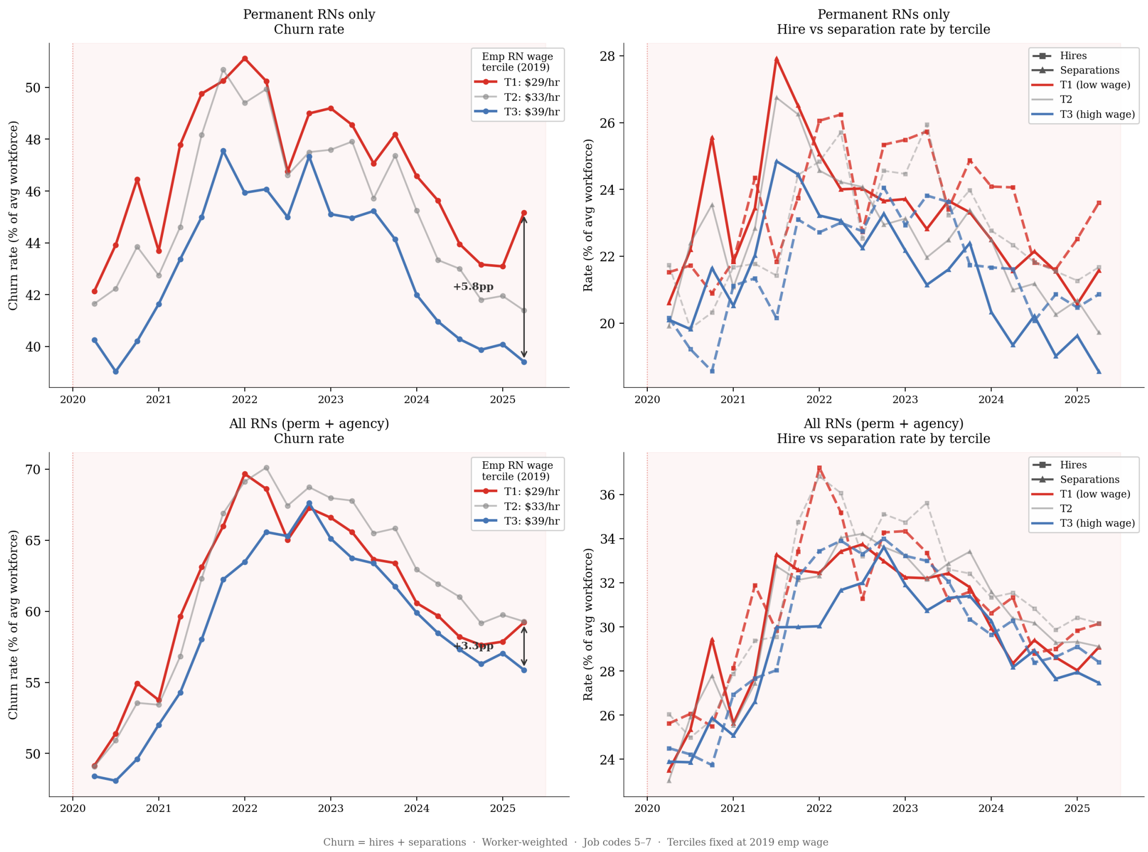


Figure A6: **RN workforce churn by 2019 employee wage tercile, 2020–2025.** Churn rate (hires plus separations as a share of average quarterly workforce) and hire/separation rates by pre-pandemic employee RN wage tercile (T1: \$29/hr, T2: \$33/hr, T3: \$39/hr), PBJ job codes 5–7. Top row: permanent (employee) 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.

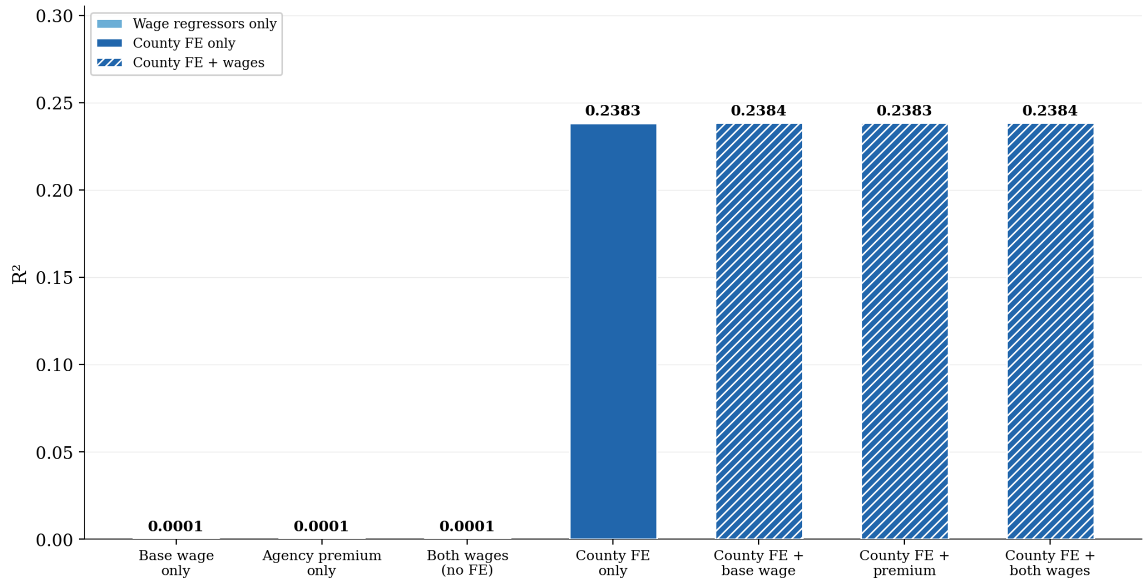
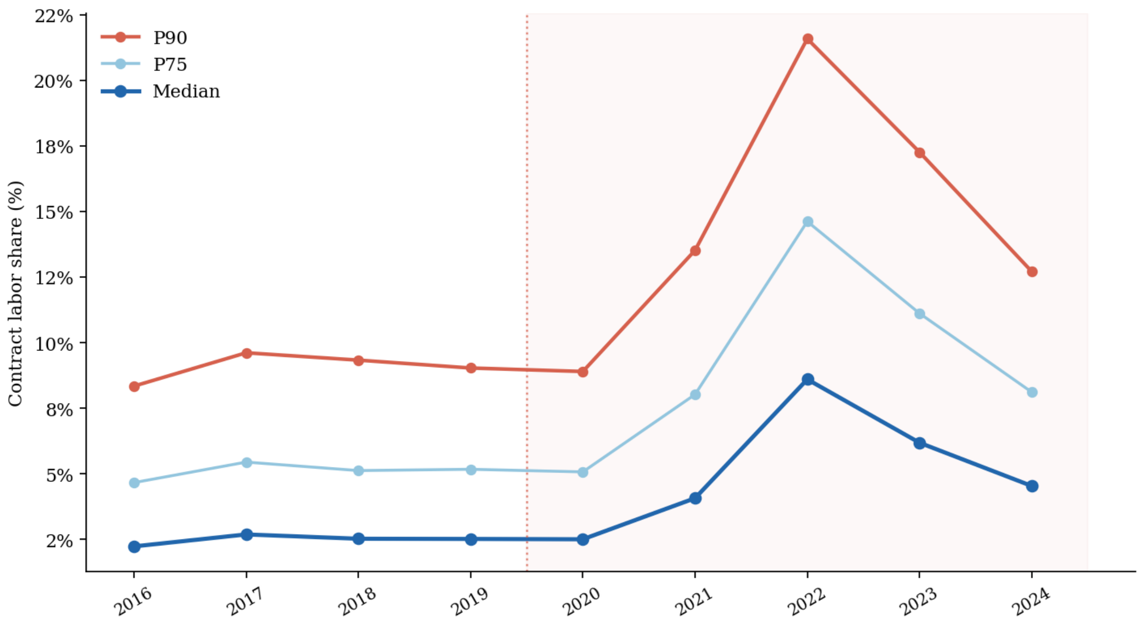


Figure A7: **What explains the change in RN separation rates?** R^2 from OLS regressions of the 2020–2022 facility-level change in RN separation rate on alternative regressor sets: base wage only, agency premium only, both wages without fixed effects, county fixed effects only, county fixed effects plus base wage, county fixed effects plus agency premium, and county fixed effects plus both wages. Sample: $N = 13{,}271$ SNFs across 2,636 counties.



Note: Contract labor share measured in dollars (not hours). Hospital HCRIS (S-3 Part II) does not report contract hours by occupation; RN-specific agency hours are not available in this dataset.

Figure A8: **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 expenditures; RN-specific agency hours are not available in this dataset.

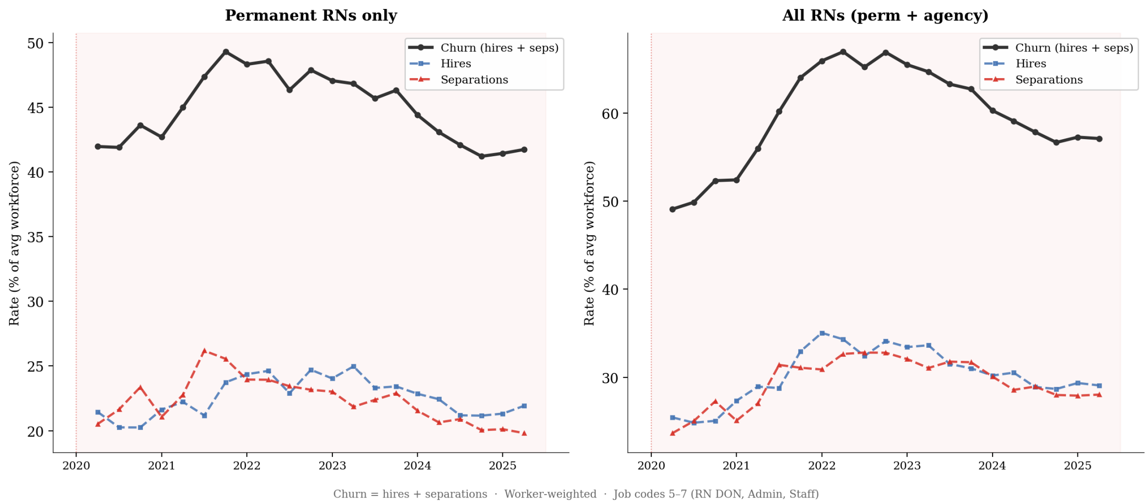


Figure A9: **Quarterly RN workforce churn, 2020–2025.** Hires, separations, and total churn (hires plus separations) as a share of average quarterly workforce, computed from individual-level PBJ employment records using worker identifiers across quarters, PBJ job codes 5–7 (RN DON, Admin, Staff). Worker-weighted. Left panel: permanent (employee) RNs only. Right panel: all RNs including agency workers. Shaded region indicates post-2020 period.

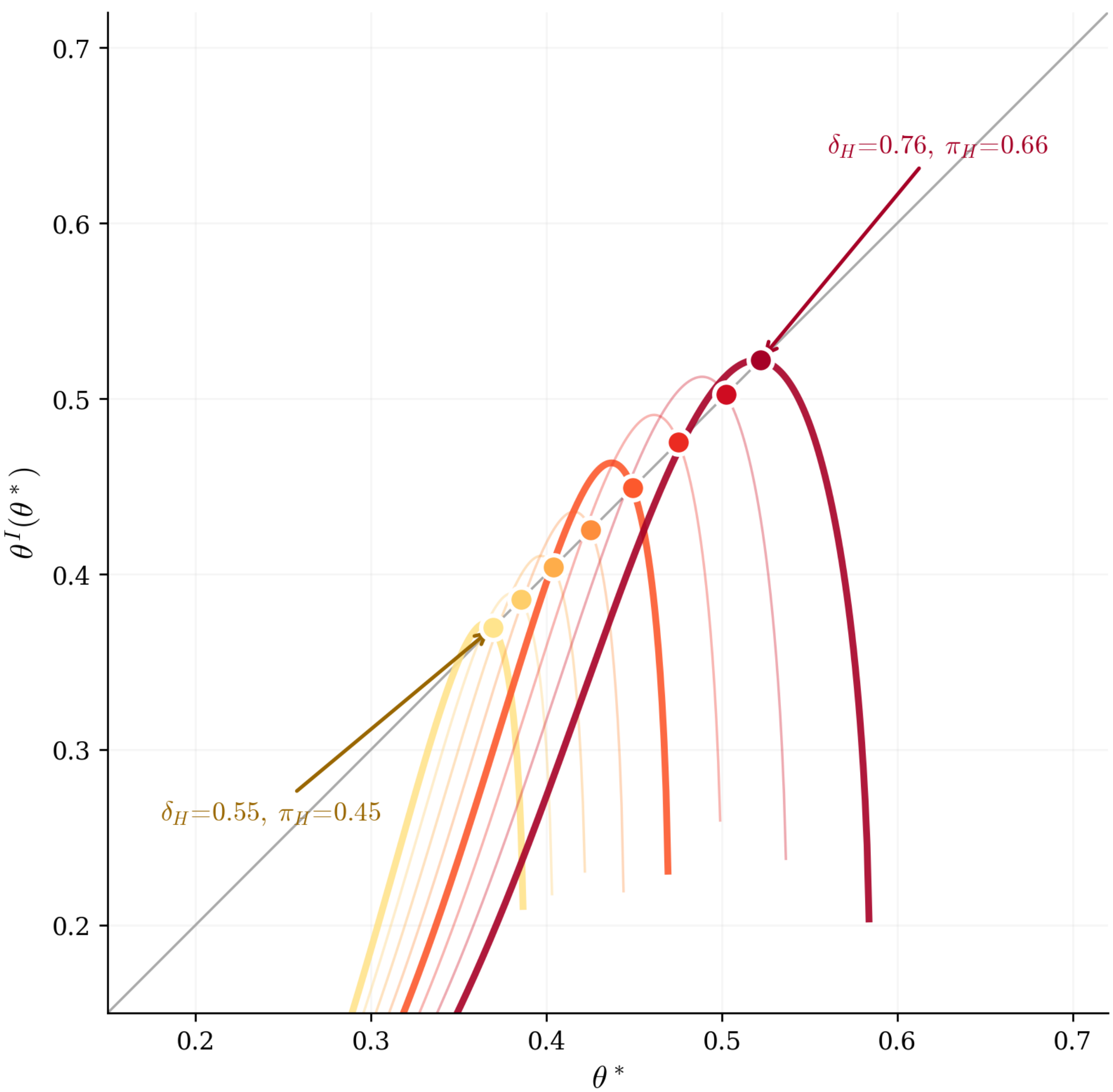


Figure A10: **Equilibrium threshold under joint demand boom.** Each curve plots $\theta^I(\theta^*)$ for (δ_H, π_H) increasing jointly from $(0.55, 0.45)$ (yellow) to $(0.76, 0.66)$ (dark red), holding all other parameters at their baseline values. Filled dots mark stable equilibria.