Improving Access or Inducing Demand? Analyzing Trade-offs in Healthcare Capacity*

Hiroki Saruya† Masaki Takahashi‡

October 22, 2025

Abstract

This paper studies the trade-offs in healthcare capacity between access to care and supplier-induced demand. We first develop an economic model where admission/discharge decisions of healthcare facilities depend on bed occupancy through capacity constraints and demand inducements. We show that the relative importance of the two mechanisms is testable using admission/discharge responses to occupancy fluctuations at different baseline occupancy levels. Applying the framework to Japanese nursing facilities, with patient deaths as exogenous occupancy shocks, we find that short-run admission decisions are mainly driven by capacity constraints. We discuss the assignment of healthcare capacity to improve access while controlling supplier-induced demand.

Keywords: capacity constraint, congestion, supplier-induced demand, nursing facility, long-term care

JEL Codes: D24, I11, I12, I18

*We thank Jason Abaluck, Steven Berry, Haruko Noguchi, Katja Seim, Yuta Toyama, and Tzu-Ting Yang as well as seminar participants at Econometric Society World Congress 2025, EuHEA Conference 2024, JHEA Conference 2023, Keio University, Sophia University, Tri-Country/Asia Pacific Health Economics Symposium, Waseda University, and Yale University for their helpful comments. Masaki Takahashi gratefully acknowledges the support of JSPS Grant-in-Aid for Scientific Research (20H01514, 23K12490). The findings and conclusions expressed are solely those of the authors and do not represent the views of any agency of the Government of Japan. All errors are ours.

†Economic and Social Research Institute at Cabinet Office, Government of Japan, and Sophia University. Email: hiroki.saruya@aya.yale.edu

‡Sophia University. Email: msk.tkhs@gmail.com

1 Introduction

The efficient delivery of goods and services under capacity constraints is a central issue in economics. In many markets, such as health care, housing, and transportation, governments regulate supply capacity for various reasons, including reducing unnecessary services and public spending, managing externalities, and ensuring comfort and safety. However, such regulations can tighten the supply capacity constraints, hinder valuable market transactions, and lead to shortages of essential goods and services. Striking a balance in supply capacity trade-offs is a common policy challenge across many sectors.

In this paper, we study the trade-offs in healthcare capacity between demand inducements and access to care. The imposition of capacity constraints on healthcare providers has been justified by concerns about supplier-induced demand: providers may use excess capacity for patients with low medical needs to increase revenues. More recently, researchers have begun to focus on the negative side of capacity constraints, such as congestion costs: productivity and accessibility in health care may decrease when providers are congested. The existing literature is limited because the positive and negative aspects of capacity constraints are analyzed separately, and their trade-offs are not examined in a unified manner. Understanding the trade-offs is essential for determining how to allocate healthcare capacity to increase necessary care while controlling unnecessary care. We fill these gaps and explore how to manage supply capacity in the healthcare sector.

Specifically, we focus on nursing facilities and analyze how bed occupancy affects their admission and discharge decisions. In inpatient care, bed occupancy is a key measure of capacity utilization and capacity constraints, and the number of (newly) treated patients measures access to health care (Alexander and Schnell, 2024) and production quantity (Grieco and McDevitt, 2017).[1] The increased demand for nursing facility care has raised concerns among policymakers and academics about the accessibility of nursing facilities. Recent studies find evidence that facilities engage in selective admissions (He and Konetzka, 2015; Gandhi, 2023; Corredor-Waldron, 2022) and discharges (Hackmann et al., 2024) when bed supply is limited. However, increasing capacity may or may not be a desirable policy, depending on the relative importance of capacity constraints and provider incentives to induce demand.

To provide a unified framework for assessing the role of capacity utilization, we build a novel economic model that shows how a facility's admission and discharge decisions depend

[1] Admissions and discharges also measure bed turnover, another important policy target.

on bed occupancy. In the model, the marginal cost of serving an additional patient[2] weakly increases with occupancy, possibly reflecting the altruistic facility's concern about quality deterioration or reduced motivations of workers, as well as pecuniary costs such as diminishing returns and higher input costs. It also approximates physical capacity constraints by assigning prohibitively high costs to services above capacity. The model also allows for (loose) income targeting or occupancy targeting by allowing the marginal benefit of income to be larger at lower occupancy, which incentivizes the facility to induce demand in response to a reduction in occupancy. These shape restrictions need not hold globally; e.g., we may impose them only above certain occupancy level, and focus our analysis on such observations.[3] In addition, the facility incurs admission and discharge costs, which represent frictions in adjusting patient volume in the short run.

The model predicts that the facility will respond to an exogenous decrease in occupancy by increasing admissions and decreasing discharges. The responses are driven by two mechanisms. First, *an income effect* incentivizes the facility to increase admissions to compensate for lost revenue. Second, *a cost effect* allows the facility to increase admissions by reducing the marginal cost of service. The magnitude of the responses depends on admission and discharge frictions. In the frictionless case, the facility adjusts the net admissions to exactly offset the occupancy reduction. In contrast, in the frictional case, it only partially offsets the occupancy shock. This suggests that an occupancy shock has a persistent effect on the facility's admissions, discharges, and occupancy.

The model also guides us on how to empirically assess the relative importance of the mechanisms. In our model, the cost (income) effect is larger (smaller) at higher occupancy. Thus, if the variation in the cost (income) effect mainly explains the variation in the admission/discharge responses at different levels of baseline occupancy, then the responses will be more (less) intense at higher occupancy levels. Moreover, we can place some bounds on the levels of the two effects at any occupancy, using empirically observable quantities. Disentangling these mechanisms is crucial for policy discussions. Previous studies on supplier-induced demand (Gruber and Owings, 1996; Ikegami et al., 2021), which emphasize the income effect as the driver of care provision under loose capacity constraints, would imply that lowering occupancy may induce wasteful care provision. In contrast, if the cost effect is a key driver, then relaxing capacity constraints can increase

[2]We use the term "patients" to refer to the people who use nursing facility services.

[3]For example, the cost function may be non-convex at lower occupancy due to economies of scale.

valuable care provision.[4]

We empirically test the above theoretical predictions in the context of Japanese nursing facilities for rehabilitation and transitional care, similar to the skilled nursing facilities (SNFs) in the U.S. Japan has the highest rate of population aging in the world (United Nations, 2019), and nursing facilities play an important role in the long-term care of the elderly. On the supply side, facilities have difficulty recruiting care workers. The high demand and staffing frictions together suggest a congestion problem: as we show empirically, facilities with a higher occupancy rate tend to have a higher patient-to-staff ratio, which leads to a higher marginal cost of service due to higher overtime pay or lower altruistic utility of care workers.[5] Extremely congested facilities may face binding capacity constraints, making additional service provision impossible. On the other hand, we study non-profit nursing facilities, which are likely concerned with securing non-negative profit but not with excess profit, consistent with our modelling assumption. Moreover, the simple reimbursement system based on a per-diem payment adjusted to care needs allows us to focus on bed management decisions rather than the content of care, without much concern about facilities picking profitable patients. Using administrative claims data, we construct facility-by-date panel data on bed occupancy, the number of admissions/discharges/deaths, and in-facility patient characteristics.

To test the theoretical predictions, we need exogenous shocks to the occupancy rate. A regression of admissions on occupancy does not necessarily yield the causal effect of occupancy, because occupancy may be affected by unobserved (pre-determined) quality[6] or operational efficiencies that also affect admissions. We address this problem by exploiting daily patient deaths as exogenous occupancy shocks. In our main analysis, we instrument daily occupancy rates by the number of patient deaths on the preceding day. Lagged patient deaths are a relevant instrumental variable (IV) for occupancy because discharges, including death-induced ones, generate empty beds on the following day (see

[4]This distinction is not absolute: both income-induced admissions and cost-induced admissions may be seen as a mechanism behind supplier-induced demand and may involve beneficial or wasteful services. Nonetheless, to derive stronger welfare conclusions, we may assume that the latter mechanism generates services that are more beneficial to patients than the former (see footnote 42). For expositional simplicity, we only refer to the former, i.e., demand inducement due to income motives, as demand inducement.

[5]Care workers who face higher occupancy may derive lower altruistic utility from serving an additional patient because they spend shorter time with the patient, which reduces the quality of care that they provide. Alternatively, heavier workload may directly reduce their altruistic motivation (burnout).

[6]Facilities' quality consists of two parts: (i) occupancy-dependent quality (e.g., care intensity) and (ii) quality that is shaped prior to occupancy (e.g., staff skills or equipment). Endogeneity arises from the latter, whereas the former is regarded as a mechanism behind admission responses to occupancy shocks.

Section 4.1). The identification assumptions are that the exact timing of patient deaths (but not necessarily the longer-run volume of deaths, which is captured by facility-year fixed effects and other controls) is exogenous to confounding factors related to the facility's daily admission/discharge decisions and other patients' preferences, and that patient deaths affect admissions and live discharges only via occupancy. We first implement an event study design to examine the effect of exogenous discharges due to patient deaths on admissions and discharges, investigating detailed dynamic effects. We then estimate regressions of weekly/monthly/quarterly admissions and discharges on daily occupancy, using lagged patient deaths as an IV.

We examine the validity of the identification assumptions from various perspectives. First, we examine whether the timing of death events is exogenous. Using an event study design, we demonstrate that the pre-trends in admissions/discharges are similar between facility-dates that face patient deaths and those that do not, consistent with the timing of deaths being exogenous. Second, we examine whether deaths and admissions/discharges are both correlated with facility-level health shocks (e.g., influenza outbreaks). We show that the trend in discharges to hospitals for acute care does not change around the timing of patient deaths, rendering the above concern unlikely. Finally, to satisfy the exclusion restriction, death events must affect admissions and discharges only through occupancy. The demand side is unlikely to respond to daily death events quickly, due to the time lag between application and both admissions and death events, information frictions, and the requirement for advance planning for discharges. On the supply-side, there is a concern that admissions/discharges could be directly affected by death events, for example due to staff burdens from extra paperwork or emotional strain. However, it should be noted that the outcomes in IV regressions are admissions/discharges aggregated over several weeks after death events, which are unlikely to be affected by the burden associated with patient deaths on a specific past day. We also show that patient deaths have little effect on same-day admissions/discharges, when the recorded occupancy rate is not affected. This result also suggests that the death-related burden is negligible in facility decisions.

We find that patient deaths increase admissions on the next day and thereafter, whereas they have much weaker effects on live discharges. Admissions increase as early as the day after patient deaths, and the increase persists for about a month. The IV regressions imply that a 1pp decrease in the daily occupancy rate increases admissions by 0.64pp and decreases discharges by 0.21pp over the next 12 weeks, implying that 84% (=64%+21%, with rounding) of the vacated beds are filled. Based on our model, the

results suggest that both admissions and discharges are frictional, with discharge frictions being much greater. In addition, the admission responses to a 1pp occupancy reduction is greater at higher occupancy levels, suggesting that the cost effect rather than the income effect is the main driver of the responses. Our baseline estimates imply that at least 72.4% of the 1-week admission response at baseline occupancy strictly between 95% and 100% is explained by the cost effect (0.21pp out of 0.29pp). We find similar patterns when we instrument for the baseline occupancy level, in addition to the local occupancy variation around the baseline. Also, the response size increases with baseline occupancy broadly, not just near 100%, suggesting that some fraction of the response is due to increasing marginal costs rather than binding capacity constraints.

Our conceptual framework and empirical results provide insights on the assignment of healthcare capacity. Without our analysis, the policymaker could conclude that assigning additional capacity would lead to wasteful demand inducements and therefore capacity should not be expanded, an idea frequently referenced to justify Certificate-of-Need (CON) laws in the U.S. and other capacity regulations. In contrast, our analysis suggests that expanding capacity can increase relatively beneficial admissions by relaxing supply constraints, especially at high-occupancy facilities, suggesting that such facilities should be prioritized for additional capacity assignment. Although further analysis is needed to inform broader policy issues, such as the optimal level of capacity in a market (not just where capacity assignment should be prioritized), our analysis can provide a stepping stone for addressing such important issues.

This study relates to the growing literature on the effect of occupancy on admission and discharge decisions. Provider incentives for selective admissions have been studied in various settings such as nursing facilities (He and Konetzka, 2015; Gandhi, 2023; Corredor-Waldron, 2022), inpatient wards (Dong et al., 2020), ICUs (Kim et al., 2015), NICUs (Freedman, 2016), and neurology wards (Samiedaluie et al., 2017). Hackmann et al. (2024) find that Medicaid patients (who are less profitable than privately funded patients) are more likely to be discharged from SNFs when occupancy is higher. We contribute to the literature by conceptualizing and examining the mechanisms by which occupancy affects admissions and discharges. In particular, unlike previous studies that emphasize supplier-induced demand and other financial incentives (e.g., Evans, 1974; Gruber and Owings, 1996; Freedman, 2016; Ikegami et al., 2021), we show that capacity constraints can be a key mechanism in our context. Our findings emphasize the importance of considering two competing mechanisms, capacity constraints and demand inducement, when analyzing

the behavior of healthcare providers and healthcare policy in general.

This study also contributes to the literature on the effect of policies on access to health care. The relationship between provider incentives and healthcare access has been studied extensively in the context of the US Medicaid, a public insurance program for low-income population (Baker and Royalty, 2000; Decker, 2007, 2009; Buchmueller et al., 2015; Gandhi, 2023; Alexander and Schnell, 2024; Cabral et al., 2025). Previous studies have emphasized the importance of expanding capacity as a tool to improve access (Gandhi, 2023). Our contribution is to analyze the trade-offs in expanding capacity between access to care and supplier-induced demand.[7] Our analysis is informative of which facilities or markets are likely to benefit from additional healthcare capacity by improving access while controlling supplier-induced demand. In this sense, our study also relates to the discussions of place-based policies (Kline and Moretti, 2014) to improve healthcare access.

Finally, this study contributes to the broad literature on how demand fluctuations combined with capacity constraints and congestion costs affect productivity (Baker et al., 2004; Collard-Wexler, 2013; Butters, 2020; Boehm and Pandalai-Nayar, 2022; Ilzetzki, 2024). Collard-Wexler (2013) simulates that smoothing demand fluctuations for ready-mix concrete expands the market due to congestion costs for delivering concrete. Butters (2020) finds that variation in demand volatility explains a large fraction of variation in hotel occupancy rates, and that eliminating the demand volatility would increase productivity. Boehm and Pandalai-Nayar (2022) study how aggregate supply curves of the manufacturing industries are convex due to micro-level capacity constraints. These results suggest that additional capacity is potentially valuable at high capacity utilization rates. On the other hand, another strand of literature studies excess capacity, in particular in declining industries (e.g., Takahashi, 2015; Nishiwaki, 2016; Okazaki et al., 2022). Our study provides a unified framework to study one type of trade-offs in capacity expansion, namely improving access (productivity) vs. inducing demand (due to extra capacity).

This paper is organized as follows. Section 2 provides institutional background on our empirical analysis. In Section 3, we present a conceptual framework. Section 4 describes our data. Section 5 presents the empirical strategy. Section 6 reports the estimation results. In Section 7, we discuss the policy implications of our results. Section 8 concludes.

[7] We emphasize the supply-side trade-off between eased capacity constraints and supplier-induced demand. This issue is related to but different from the issue of how to distinguish supplier-induced demand from patient-induced demand (e.g., due to shorter distance to physicians under higher physician density), a major issue in the literature on supplier-induced demand. Our research design allows us to minimize the effects of demand shocks and to focus on the supply side.

2 Institutional Background

2.1 Nursing Facility Industry in Japan

We study nursing facilities in Japan, which are financed by the public long-term care insurance (LTCI). Japan's LTCI is a social insurance program mainly for people over the age of 65 who require long-term care (LTC) services. Eligibility for LTCI benefits is determined by an in-person health examination. The health examination evaluates the applicant's physical and mental disabilities and calculates a health score that indicates the applicant's level of care needs. Applicants are eligible for LTCI benefits if their health score is above a minimum threshold. Eligible LTCI beneficiaries can use various LTC services, including both home and institutional care, at a coinsurance rate. Because of the rapid aging of the population, public spending on LTCI continues to increase. The total annual cost of LTCI was 11.5 trillion JPY in fiscal 2023, 1.94% of Japan's nominal GDP (Ministry of Health, Labor and Welfare, 2023b). Institutional care, including nursing facilities, accounts for about one-third of total costs.

We focus on a type of nursing facilities called Geriatric Health Services Facilities (GHSFs).[8] Their primary goal is to provide high-quality inpatient rehabilitation and transitional care to LTCI beneficiaries and to restore their physical abilities to the point where they can live at home or in the community.[9] Thus, they are similar in their mission to the U.S. Skilled Nursing Facilities (SNFs). Unlike most SNFs, however, GHSFs are non-profit organizations: they may earn a profit to keep their facilities afloat, but they are not allowed to distribute the profit to shareholders or other parties. The establishment of a GHSF and changes to its bed capacity require the approval of the prefectural governor. As of April 2022, there were 4,230 GHSFs nationwide, with approximately 355,900 patients admitted (Ministry of Health, Labor and Welfare, 2023a).

GHSFs provide care for two types of patients. "Long-stay" patients are admitted to the facility for rehabilitative care in order to return to the community. "Short-stay" patients, on the other hand, visit GHSFs to receive temporary assisted living services, typically for temporary unavailability of family caregivers (e.g., respite). Stay types are defined based on claims codes rather than by length of stay. In our empirical analysis, we primarily focus on the admissions and live discharges of the long-stay patients, because short-stay

[8] They are called "Kaigo Roujin Hoken Shisetsu" or "Roken" for short in Japanese.

[9] According to Japan Association of Geriatric Health Services Facilities (2015), GHSFs' motto is to "improve the user's function to enable them to go back home".

admissions and discharges are more likely to be influenced by exogenous factors.[10]

Various healthcare professionals work in GHSFs to provide appropriate care, including physicians, nurses, care staffs, physiotherapists, and social workers. The supply of care workers is not keeping pace with the increasing demand for care due to the rapid aging of the population. As a result, nursing facilities, including GHSFs, are facing shortages of care workers. According to Care Work Foundation (2016), 62.6% of facilities reported being understaffed, and 73.1% of the understaffed facilities reported recruitment difficulties as the main reason for staff shortages.[11] Due to this staffing friction, each worker will have to see a larger number of patients at higher occupancy (as Figure 2 shows), implying that serving an additional patient becomes increasingly costly due to both pecuniary (e.g., higher overtime pay) and non-pecuniary (e.g., reduced altruistic utility) reasons.

2.2 Admission, Treatment, and Discharge

To be admitted to a GHSF, LTCI beneficiaries must apply for admission to the facility, in consultation with physicians and social workers, and must meet several conditions. Upon receipt of the application, the facility interviews the applicant to ascertain their physical condition, living arrangements, and medical needs. Because GHSFs cannot provide acute medical care, patients must be in a stable condition. The facility decides whether to admit the patient based on the interview and documentation, such as a medical certificate.

GHSFs provide rehabilitative care according to each patient's care plan. In the early stages of inpatient care, a care plan is developed based on the patient's goals. The care plan is reviewed periodically as treatment progresses.

When a patient is ready for discharge, the facility plans their discharge in consultation with the patient and their family. They work together to prepare the patient's post-discharge living environment, including the LTC services to be used at home. Patients who wish to live outside the current facility are discharged either to their home or to a nursing home where they can remain for the rest of their lives. If patients require acute care, they may be transferred to a hospital.

[10] Many short stays are due to a planned absence of a family caregiver, the timing of which is likely to be fixed in advance, while many others are due to a family caregiver's emergency, in which case facilities will find it difficult to reject the application (Ministry of Health, Labor and Welfare, 2017). The timing of short-stay discharges is also influenced by the restrictions on the lengths of short stays.

[11] Low wages (57.3%) and demanding jobs (49.6%) were major cited reasons for recruitment difficulties. Because revenue from services is capped by government-set reimbursement rates, facilities do not have the flexibility to raise wages by increasing service prices.

GHSFs also provide end-of-life care for patients who choose to spend their final days in the facility. End-of-life care is provided to relieve pain, suffering, and stress for patients so that they can maintain human dignity until the end of life. End-of-life care at GHSFs includes pain relief through medication, prevention of bedsores, and psychological care to reduce anxiety and fear.

2.3 Reimbursement Policy

Reimbursement for GHSFs depends on the beneficiaries' care needs. Beneficiaries are assigned to one of seven groups based on the health score mentioned in Section 2.1. The groups consist of support levels 1 and 2, and care levels 1–5 in ascending order of care-needs levels (i.e., care level 5 means the highest needs). Appendix Table A1 describes the general health status for each care level. Only beneficiaries classified as care level 1–5 may be admitted to a GHSF.

GHSF reimbursement consists of two components: a per-diem fixed payment and a fee-for-service (FFS) payment. The fixed payment is paid to the facility for a patient's stay for one day, regardless of the content of care. To reflect the burden of care, the amount of the per-diem payment is set higher for higher care levels. The FFS payment is paid for specific medical procedures, such as short-term intensive rehabilitation, dementia care, and end-of-life care. Appendix Table A2 shows per-diem fixed and FFS payments by care levels, using our analysis sample described in Section 4.[12] The fixed payment accounts for roughly 90% of the total reimbursement for GHSFs for serving long-stay patients. Thus, bed occupancy is more important to the facilities' revenue than the content of the care provided to long-stay patients.

Summary. The Japanese GHSF is an attractive setting for empirical analysis, because of its economic importance and its reimbursement system which mitigates concerns about patient selection. The institutional characteristics also guide our modeling: (1) The non-profit facilities may be concerned with securing profits, but not with excess profits. They are also concerned with patient welfare. (2) The numbers of admissions and discharges are main choice variables, not which patients or which services to select. (3) Serving an additional patient becomes increasingly costly at higher occupancy, due to the congestion

[12]Since we can only observe each patient's FFS payment at the monthly level, the daily averages of the FFS payment are calculated by dividing the patient's total FFS payment by the number of days in the facility. See the tablenote of Appendix Table A2 for more details.

problem caused by labor shortages and frictions.

3 Conceptual Framework

To guide our empirical analysis, we present an economic model of a facility's admission and discharge decisions. The model captures two competing mechanisms, demand inducements and capacity constraints (congestion), and generates testable predictions regarding their relative importance.

3.1 Model Setup

A representative facility chooses the number of new patients to admit, a, and the number of in-facility patients to discharge, d, to maximize its objective function given the number of patients currently in the facility, n. We fix capacity[13] and express each variable as a ratio to capacity (e.g., n denotes occupancy rate). The utility of the facility is

$$U(n,a,d) = \underbrace{V(rp)}_{\text{income utility}} + \underbrace{b^P p - C^P(p)}_{\text{service utility}} + \underbrace{b^A a - C^A(a)}_{\text{admission utility}} + \underbrace{b^D d - C^D(d)}_{\text{discharge utility}}, \tag{1}$$

where $p = n + a - d$ is the occupancy rate after admissions and discharges are realized, and r (constant) is the per-patient reimbursement rate net of marginal costs.

The first term represents utility from gross profit rp, converted by V which may capture fixed costs (e.g., $V(R) = R - FC$). V satisfies $V'(\cdot) > 0$, $V''(\cdot) \leq 0$, and $V'''(\cdot) \geq 0$.[14] We allow V to express loose income targeting, e.g., an approximate non-negativity constraint on profit.[15] An income effect, including literal income targeting as a limit case, is a common way to explain supplier-induced demand (McGuire and Pauly, 1991; Gruber and Owings, 1996), and a highly concave utility function is a way to express a large income effect (see also Camerer et al., 1997). In our context, the non-profit facility may strongly desire to avoid operating in the red, while it may care less about excess profit. Alternatively, it may target an occupancy level, and the incentive to induce demand may increase if occupancy falls below the target level.

[13] This is for ease of exposition; our framework is applicable to study capacity changes.

[14] The conditions hold for many common candidates for V, e.g., a CARA utility function, a CRRA utility function, and $V(R) = R^k$ for $k \in (0,1]$. Nonnegativity of the third derivative is also a common assumption to derive the concavity of the consumption function (e.g., Carroll and Kimball, 1996).

[15] E.g., $V(R) = v(R - FC)$, where $v'(\tilde{R})$ is high at $\tilde{R} < 0$ and low at $\tilde{R} > 0$.

Following the literature on non-profit organizations (Lakdawalla and Philipson, 1998; Gaynor and Vogt, 2003), we assume that the facility's utility depends on its output. The second term of Eq.(1) represents the altruistic utility derived from serving p patients. $b^P \geq 0$ denotes the benchmark per-patient utility from service and $C^P(p)$ denotes a weakly convex "congestion cost" that satisfies $C^{P\prime\prime\prime} \geq 0$.[16] Per-patient utility declines steeply as the number of patients increases, possibly for the following reasons. First, congestion may reduce per-patient utility by lowering service quality, e.g., by reducing the amount of time workers spend with each patient (Shurtz et al., 2022) and other inputs. Second, congestion may harm the altruistic motivations of workers who aim to provide good care. Third, C^P may approximate the capacity constraint (i.e., $C^P(n) = +\infty$ for $n \geq 1$), if it increases rapidly as occupancy approaches capacity. Finally, with an appropriate reinterpretation of r, C^P may capture the pecuniary cost of service that increases nonlinearly with volume.[17] Increasing marginal costs can result from diminishing marginal product of inputs or higher labor costs (e.g., higher overtime pay).

The last two terms of Eq.(1) represent the utility derived from achieving the admission and discharge missions. The facility's mission is to provide access to quality care for anyone in need and return them to their home. We capture the facility's desire to achieve this objective by including additional terms to the utility. $b^A a - C^A(a)$ represents the utility derived from quality-adjusted admissions, where $b^A \geq 0$ is the benchmark utility per admission (in addition to b^P) and C^A is weakly convex and captures the reduction in quality due to higher admission volumes (e.g., poor performance in assessing patient needs or coordinating the admission process). Similarly, $b^D d - C^D(d)$ represents the utility from quality-adjusted discharges, where $b^D \geq 0$ and C^D is weakly convex. C^A and C^D may also reflect pecuniary cost.[18]

We impose the global shape restrictions on the income utility and cost functions for simplicity, but they can be relaxed to local assumptions. For example, C^P may be non-convex at lower occupancy rates, due to economies of scale. In this case, we may impose the shape restrictions only when occupancy is relatively high, and we may apply the

[16] The conditions mean that the marginal cost grows (weakly) faster at higher occupancy. They hold for many common candidates for C^P, e.g., the exponential function and $C^P(p) = p^k$ for $k = 1$ or $k \geq 2$.

[17] E.g., if $V(R) = R - FC$ where $R = rp$ is *revenue*, then the first two terms of (1) become the sum of *profit* $rp - C^P(p) - FC$ and altruistic utility $b^P p$. Alternatively, $V(R)$ may be concave in revenue R due to revenue targeting, with $C^P(p)$ capturing all variable costs.

[18] Admission cost may reflect the cost of assessing patients' needs, coordinating the admission process, and moving patients to the facility. Discharge cost may include similar factors. These costs can be convex, for example, due to higher labor costs of workers in charge of discharges.

theoretical framework to empirical analysis of high-occupancy cases.

The facility's decision problem is

$$\max_{a\geq 0, d\in[0,n]} U(n,a,d). \tag{2}$$

We treat n, a and d as continuous variables. We assume that problem (2) has an interior solution $(a^*, d^*) = (a^*(n), d^*(n))$ that satisfies the first-order conditions, and that the resulting occupancy rate is also interior.

3.2 Admission/Discharge Responses to Occupancy Reduction

We examine how the admissions and discharges (a^*, d^*) respond to a decrease in occupancy n, which can also be interpreted as an increase in capacity.[19] Denote the optimal admissions and discharges at $n = \bar{n}$ by $\bar{a} = a^*(\bar{n})$ and $\bar{d} = d^*(\bar{n})$, and let $\bar{p} = \bar{n} + \bar{a} - \bar{d}$. Also, denote the marginal cost by $MC^g(\cdot)$ for $g = P, A, D$, and let $MB^A(p) = rV'(rp) + b^P + b^A$ denote the marginal benefit of admission. We assume that the admission and discharge cost functions are weakly quadratic: $MC^A(a) = \kappa_1^A + \kappa_2^A a$ and $MC^D(d) = \kappa_1^D + \kappa_2^D d$, with $\kappa_1^A, \kappa_2^A, \kappa_1^D, \kappa_2^D \geq 0$. No assumption on MC^P or MB^A is required for the following results, except that $-MB^A$ and MC^P are weakly increasing.

Proposition 1. ***(Frictional Responses)*** *Suppose $\kappa_2^A > 0$ and $\kappa_2^D > 0$. Then, for any $\bar{n} \in (0,1)$, the following statements hold at $n = \bar{n}$, and any weak inequality in the statement holds strictly if and only if $V''(r\bar{p}) < 0$ or $C^{P\prime\prime}(\bar{p}) > 0$.*

(i) (Responses to exogenous discharges)

- *(a) $-\frac{\partial a^*}{\partial n} \geq 0$ and $-\frac{\partial d^*}{\partial n} \leq 0$.*
- *(b) Holding $MC^A(\bar{a})$ and $MC^D(\bar{d})$ constant, we have $\frac{\partial}{\partial \kappa_2^A}\left|\frac{\partial a^*}{\partial n}\right| \leq 0$, $\frac{\partial}{\partial \kappa_2^D}\left|\frac{\partial a^*}{\partial n}\right| \geq 0$, $\frac{\partial}{\partial \kappa_2^A}\left|\frac{\partial d^*}{\partial n}\right| \geq 0$, and $\frac{\partial}{\partial \kappa_2^D}\left|\frac{\partial d^*}{\partial n}\right| \leq 0$.*

(ii) (Imperfect adjustment) $-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) \in [0,1)$.

Proposition 2. ***(Frictionless Responses)*** *Suppose $\kappa_2^A = 0$ or $\kappa_2^D = 0$. Then, for any $\bar{n} \in (0,1)$, $-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) = 1$ at $n = \bar{n}$. Moreover:*

(i) If $\kappa_2^A > 0$ and $\kappa_2^D = 0$, then $-\frac{\partial a^*}{\partial n} = 0$ and $-\frac{\partial d^*}{\partial n} = -1$.

(ii) If $\kappa_2^A = 0$ and $\kappa_2^D > 0$, then $-\frac{\partial a^*}{\partial n} = 1$ and $-\frac{\partial d^*}{\partial n} = 0$.

[19] Here, "capacity" consists of both equipment (e.g., beds) and staffing.

Proofs are in Online Appendix B. Proposition 1-(i-a) states that admissions increase and discharges decrease as the occupancy rate n decreases. We show in Online Appendix B that the admission response can be expressed as

$$-\left.\frac{\partial a^*}{\partial n}\right|_{n=\bar{n}} = \underbrace{-\frac{\kappa_2^D}{D_{J_F}(\bar{a},\bar{d};\bar{n})}MB^{A\prime}(\bar{p})}_{\text{income effect}\geq 0} + \underbrace{\frac{\kappa_2^D}{D_{J_F}(\bar{a},\bar{d};\bar{n})}MC^{P\prime}(\bar{p})}_{\text{cost effect}\geq 0} \tag{3}$$

where $D_{J_F}(\bar{a},\bar{d};\bar{n})$ is a positive term that depends on κ_2^A and other parameters. The first term represents *an income effect*, whereby a decrease in occupancy induces the facility to increase admissions to compensate for the lost income. The second term represents *a cost effect*, whereby a decrease in occupancy reduces the marginal cost of service and allows the facility to admit more patients. Similarly, the discharge response can be expressed as

$$-\left.\frac{\partial d^*}{\partial n}\right|_{n=\bar{n}} = \underbrace{\frac{\kappa_2^A}{D_{J_F}(\bar{a},\bar{d};\bar{n})}MB^{A\prime}(\bar{p})}_{\text{income effect}\leq 0} \underbrace{-\frac{\kappa_2^A}{D_{J_F}(\bar{a},\bar{d};\bar{n})}MC^{P\prime}(\bar{p})}_{\text{cost effect}\leq 0} \tag{4}$$

and interpreted analogously: a decrease in occupancy induces the facility to serve more patients, i.e., reduce discharges, in order to compensate for the lost income (income effect) and because the marginal cost is reduced (cost effect).

Proposition 1-(i-b) shows that, holding marginal costs constant (hence holding a^* and d^* constant), the admission (discharge) response to occupancy shocks is larger when κ_2^A is smaller (larger) and/or κ_2^D is larger (smaller). A larger κ_2^A means larger "frictions" in adjusting admission, in the sense that the marginal cost of admission increases more steeply (and an analogous interpretation holds for discharge).[20] Thus, the result states that the facility attempts to adjust occupancy in the less frictional way.

Proposition 1-(ii) states that empty beds created by an exogenous occupancy reduction are not fully filled if both admission and discharge adjustments are frictional in the sense that $\kappa_2^A, \kappa_2^D > 0$. The difference $\left(-\frac{\partial a^*}{\partial n}\right) - \left(-\frac{\partial d^*}{\partial n}\right)$ represents the extent to which the occupancy reduction is offset by increased admissions and decreased discharges. The difference is less than one, indicating imperfect adjustment. This in turn suggests that the response may be dynamic in a repeated decision setting: the empty beds are only partially filled each period, leaving room for adjustment in future periods.

[20] Precisely speaking, the admission and discharge costs shape *the costs of adjusting occupancy*. We stick to the term "admission/discharge frictions" for convenience.

In contrast, Proposition 2 gives predictions for "frictionless" cases. If the marginal cost of admission or discharge is constant, then the facility perfectly adjusts its occupancy in response to an occupancy reduction ($-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) = 1$). Moreover, if the marginal cost of admission (discharge) increases while that of discharge (admission) does not, the response is driven solely by discharge (admission), the less costly way to adjust occupancy.

3.3 Evaluating the Importance of the Income and Cost Effects

We now discuss our approach to evaluate the relative importance of the income effect and cost effect, as defined in Eq. (3) and (4). Let $E_g^I(n)$ and $E_g^C(n)$ denote the income effect and cost effect for admissions ($g = A$) or discharges ($g = D$), respectively. Note $|E_A^I(n)| = \psi|E_D^I(n)|$ and $|E_A^C(n)| = \psi|E_D^C(n)|$ for some $\psi > 0$ that is independent of n.

Proposition 3. $\frac{\partial|E_A^I|}{\partial n} \leq 0$ *and* $\frac{\partial|E_A^C|}{\partial n} \geq 0$ *at any* $n \in (0,1)$.

Corollary 1. *At any* $n \in (0,1)$, $-\frac{\partial^2 a^*}{\partial n^2} \geq (\leq) 0$ *and* $\frac{\partial^2 d^*}{\partial n^2} \geq (\leq) 0$ *if* $\frac{\partial|E_A^C|}{\partial n} \geq (\leq) -\frac{\partial|E_A^I|}{\partial n}$.

Proofs are straightforward given the exact expressions for E_A^I and E_A^C, hence omitted. In our model, marginal income is especially valuable at lower occupancy levels, leading to the income effects that decrease with baseline occupancy. In contrast, congestion becomes increasingly severer at higher occupancy levels, leading to the cost effects that increase with baseline occupancy. Thus, if the variation in the admission/discharge responses at different levels of baseline occupancy is mainly driven by the variation in the cost effects, then the responses will be more intense at higher occupancy levels. The converse is true if the variation in the responses is mainly explained by the variation in the income effects.

Although Corollary 1 only concerns whether the *variation* in the income effects or the cost effects mainly explains the variation in the admission/discharge responses, the result can be used to bound the *levels* of the responses attributable to each effect, at a given n. Panel A in Table 1 illustrates how to construct the bounds on income and cost effects. We consider two cases, one where occupancy n is low and one where it is high. We denote the admission response to a 1pp decrease in occupancy in each case as R_l and R_h ($> R_l$), respectively. Suppose that the income effect explains $\alpha \geq 0$ of the admission response when n is low. The cost effect is the remaining part of the admission response, $R_l - \alpha$. Now, because the income effect decreases with n, the income effect at high occupancy should be in $[0, \alpha]$. The bound of the cost effect at high occupancy is then $[R_h - \alpha, R_h]$.

Table 1: Bounds on income and cost effects for admission

			Decomposition of the admission response	
	Occupancy	Admission response	Income effect	Cost effect
(A) Formula	n is low	R_l	$\alpha \in [0, R_l]$	$R_l - \alpha$
	n is high	R_h $(> R_l)$	$[0, \alpha]$	$[R_h - \alpha, R_h]$
(B) Example 1	$n = 50$pp	0.4	0.2	0.2
	$n = 90$pp	0.7	$[0, 0.2]$	$[0.5, 0.7]$
(C) Example 2	$n = 50$pp	0.4	0.4	0
	$n = 90$pp	0.7	$[0, 0.4]$	$[0.3, 0.7]$

Notes: Table 1 shows the formula for the bounds on the income and cost effects for admission. The admission response represents the increase in admissions when the occupancy rate is exogenously decreased by 1pp. The unit of the quantities is the percentage point (pp) of occupancy.

Panel B and C in Table 1 illustrate the discussion above. Suppose that the admission response is 0.4pp at $n = 50$pp and 0.7pp at $n = 90$pp (Panel B). If we assume that the income effect explains 0.2pp (50%) of the response at $n = 50$pp, then the income effect at $n = 90$pp is in $[0\text{pp}, 0.2\text{pp}]$, so the cost effect at $n = 90$pp is in $[0.5\text{pp}, 0.7\text{pp}]$. Without such an assumption on the income effect at $n = 50$pp, we still know that its upper bound is the entire response, 0.4pp (Panel C), so the cost effect (income effect) at $n = 90$pp is at least 0.3pp (at most 0.4pp). This method is useful when, e.g., we discuss the value of the admission responses by examining the fraction of induced demand in them.

3.4 Possible Simplifications and Extensions

The above model can be simplified while maintaining its main insights. First, we can omit the discharge utility and replicate the main predictions about admissions alone (or, conversely, we can study discharge decisions alone). Second, we can omit the terms $b^P p$, $b^A a$, and $b^D d$. The last two of these are included to explain the coexistence of positive admissions and discharges, but we may set these terms to zero and focus only on, e.g., admissions. Then, we do not need $b^P p$ to explain positive admissions.

On the other hand, the model abstracts from several features, and can be enriched by introducing them. First, it omits the facility's choice of pre-determined care quality to influence patient health. Instead, the facility chooses admissions and discharges considering their effects on congestion-dependent service quality, e.g., care intensity (recall footnote

6). So, discharge in our model is a short-run tool to manage congestion, rather than a long-run product of care. Second, it omits dynamics. Although incentives to reserve beds for more profitable future patients can be important in some contexts (Gandhi, 2023), our context features relatively homogeneous patient profitability, as discussed in Sections 2, making dynamics relatively unimportant (see also Section 6.3 for empirical investigation). Also, similar results will hold if we extend the model with opportunity costs of admission. Third, we can relax the assumptions on functional forms, as mentioned above. Finally, the model omits patient and provider heterogeneity, to focus on the role of occupancy. In Section 3.5, we discuss its implications for our empirical analysis.

3.5 Connecting Theory to Empirics

The above analysis shows how we can infer provider incentives by examining the effects of occupancy changes on admissions/discharges. However, its empirical application will be difficult due to heterogeneous demand shocks, e.g., pre-determined provider quality (Einav et al., 2025), and supply shocks, e.g., admission costs. In Online Appendix C, we develop a simple two-period model to discuss empirical implications of such heterogeneity. It is summarized below. In period 1, the facility chooses its initial occupancy n to maximize its utility. In period 2, given n, it chooses new admissions a to maximize its utility. We simplify the above model with the following assumptions: (i) No discharge occurs. (ii) The facility is myopic. Now, we introduce utility shocks ξ such that $\frac{\partial MC^A}{\partial \xi}(a, \xi) < 0$ at each a. ξ may be interpreted as supply shocks or a reduced-form expression for demand shocks: e.g., higher demand shocks (due to better pre-determined quality, better transportation, seasonality, etc.) attract more patients, making admission easier. In addition, the facility utility depends on idiosyncratic shocks η, which is implicit here for simplicity.

Propositions 1-3 and Corollary 1 continue to hold if we hold the shocks (ξ_1, ξ_2, η_2) constant, where the subscripts denote time period. Note that the sole role of η_1 in our model is to explicitly indicate the source of exogenous variation in occupancy n^* conditional on ξ_1. Moreover, the same results also hold for the averages of admissions or their derivatives over idiosyncratic shocks η_2 while holding (ξ_1, ξ_2) constant. This is useful for empirically testing the predictions of the propositions.

In contrast, testing the propositions is hindered by the variation in (ξ_1, ξ_2). Proposition

4 in Online Appendix C shows that the following approximation holds:

$$\frac{\mathrm{Cov}\left(a^*, n^*\right)}{\mathrm{Var}\left(n^*\right)} \approx \underbrace{\frac{\partial a^*}{\partial n}}_{\text{admission response}} + \underbrace{\frac{\partial a^*}{\partial \xi_2}\frac{\partial n^*}{\partial \xi_1}\frac{\mathrm{Cov}\left(\xi_1, \xi_2\right)}{\mathrm{Var}\left(n^*\right)}}_{\text{endogeneity bias}}. \tag{5}$$

where n^* and a^* denote the optimal initial occupancy and new admissions, respectively. Eq.(5) shows that the simple regression of admission on occupancy overestimates the true admission response $\frac{\partial a^*}{\partial n}$ if there exist persistent shocks ($\mathrm{Cov}\left(\xi_1, \xi_2\right) > 0$) that positively affect both current admissions ($\frac{\partial a^*}{\partial \xi_2} > 0$) and current occupancy, i.e., past admissions ($\frac{\partial n^*}{\partial \xi_1} > 0$). In our empirical analysis below, therefore, we need to exploit variation in occupancy n^* that is not driven by persistent shocks ξ but by idiosyncratic shocks η.

4 Data

4.1 Data Sources and Sample Selection

The primary data source is the Survey of Long-Term Care Benefit Expenditures, also referred to as administrative LTCI claims data. The sample period is from April 2011 to March 2018.[21] The claims data contain information on each LTCI beneficiary's monthly use of LTC services, including both home-based care and facility care. We also observe individual characteristics such as age, gender, care level, and coinsurance rate. The data also provide admission dates for all episodes, and discharge dates and destinations for all discharges that occurred in our sample period. If the discharge destination is recorded as "death," it is interpreted that the bed became vacant due to patient deaths.[22] We also use the Survey of Institutions and Establishments for Long-Term Care to obtain annual information on the characteristics of each GHSF, such as the number of beds, physicians, nurses, and care staffs. Combining these datasets, we construct a facility-by-date panel data on the number of in-facility patients, deaths, and admissions and discharges for each facility-date. Based on the number of beds and patients, we can calculate the daily bed occupancy rate for each facility. In LTCI claims data, if patients stay at the facility

[21] In Japan, a fiscal year begins on April 1 and ends on March 31.

[22] A potential concern in using deaths as a daily occupancy shifter is measurement error problems, e.g., discrepancies between the date of death vs. the date of discharge. Given our event study results in Figure 3, this concern is unlikely to be important. Moreover, in the main analysis, we use deaths as instruments, so their measurement error does not invalidate the main empirical strategy.

for inpatient care on a given day, they are included in the number of in-facility patients. Therefore, admissions increase the occupancy rate on the same day, while discharges lower the occupancy rate on the following day.[23]

The sample for our analysis is selected as follows. First, we exclude facilities with a specialized dementia unit, because we cannot observe whether patients are admitted to regular or dementia units, making it difficult to identify the relevant congestion measure. We also exclude facilities whose maximum occupancy rate falls in the bottom or top 1 percentile of the distribution of maximum occupancy across providers. We impose the former restriction to eliminate providers that are always empty, while we impose the latter restriction to exclude occupancy outliers that may be mismeasured.[24]

4.2 Summary Statistics

Table 2 shows summary statistics for our main sample at facility-date level. The average number of beds is 83, with most facilities having between 50 and 100 beds. On average, a facility employs 0.72 full-time equivalent physicians and 8.96 full-time equivalent nurses.[25] The average occupancy rate is 77%, and the breakdown by care levels shows that the main patients in GHSFs are those in care levels 3-5 with high care needs. The average number of daily long-stay admissions and discharges is about 0.2, which means that one long-stay patient is newly admitted or discharged every 5 days.

Figure 1 shows the histogram of occupancy rates in our analysis sample. Occupancy is mostly concentrated in the range of 80-100%. A small number of observations have occupancy rates higher than 100%, possibly due to changes in bed capacity not reflected in the surveys (e.g., temporary use of makeshift beds), although the reimbursement rate is reduced if such excess utilization persists for a period of time. The availability of makeshift bed and the reimbursement rule suggest that the number of beds reported by facilities may be a "soft" constraint that can be slightly exceeded at some cost.[26] There are also observations with extremely low occupancy rates. Although these observations are included in our main analysis sample, excluding facilities with low occupancy rates

[23] In our main empirical analysis, we compute and use occupancy before admissions and discharges occur in the day, which corresponds to n in the theoretical framework in Section 3.

[24] Our main results are unchanged when we remove the sample restrictions based on the maximum occupancy rate. See Appendix Figure A5 and Appendix Table A4.

[25] Since GHSFs are required to have a physician, facilities that do not employ a full-time physician employ a part-time physician.

[26] Our theoretical framework can accommodate occupancy exceeding 100%, without any modifications.

Table 2: Summary Statistics

	Mean (1)	SD (2)	p10 (3)	p90 (4)
Facility-date (Obs. = 6,759,473, #Facilities = 3,073)				
Number of beds	83.44	30.45	48	100
Number of physicians	0.72	0.56	0	1
Number of nurses	8.96	3.76	5	13
Number of care staffs	26.02	10.37	13	38
Occupancy rate (pp)	76.73	30.20	12.86	98.15
Care level 1	8.11	6.77	0	17.14
Care level 2	13.74	8.21	1.25	24.00
Care level 3	17.72	9.11	2.50	28.00
Care level 4	20.27	10.25	3.00	32.00
Care level 5	15.70	11.77	2.00	30.00
Number of admissions	0.47	0.85	0	2
Short stay	0.27	0.65	0	1
Long stay	0.20	0.47	0	1
Number of discharges	0.47	0.85	0	2
Short stay	0.27	0.65	0	1
Long stay	0.19	0.49	0	2
Home	0.04	0.21	0	0
Hospital	0.07	0.27	0	0
Death	0.01	0.12	0	0

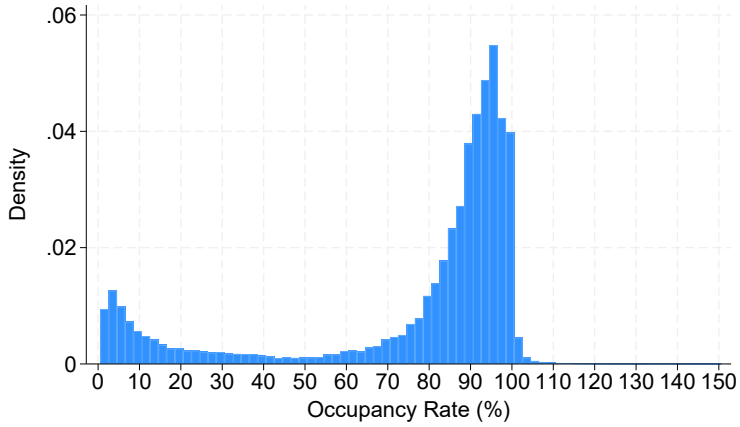
Notes: Table 2 presents summary statistics for the facility-date panel. The last two columns present 10th and 90th percentiles. The number of physicians, nurses, and care staffs are full-time equivalent. The occupancy rate (overall and by care levels) is the number of patients in the category divided by the number of beds, expressed in pp. The other variables are expressed in level.

yields qualitatively similar estimation results (see Section 6.2). Thus, our qualitative results are not affected by, e.g., economies of scale at lower occupancy.

Figures 2a and 2b show the binscatter of the patient-to-nurse ratio and the patient-to-care staff ratio against occupancy rate, using facility-fiscal year observations.[27] The average number of patients per nurse (care staff) is about 8 (2.7) at 80% occupancy, and it rises to about 9.5 (3.4) as occupancy approaches 100%. This suggests that it is difficult for facilities to adjust staffing levels based on occupancy rates, implying less nurse time

[27]For ease of viewing, the plot uses only observations with at least 50% occupancy.

Figure 1: Distribution of Occupancy Rates

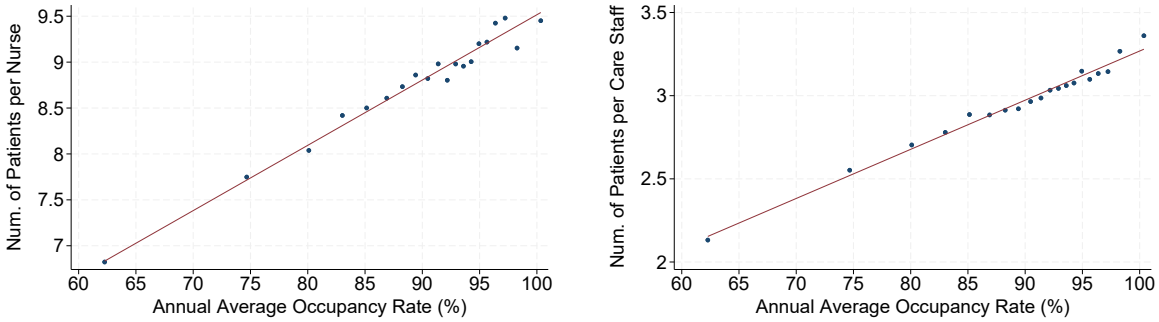


Notes: Figure 1 shows the histogram of occupancy rates in our analysis sample.

Figure 2: Occupancy and Patient-to-Staff Ratio

(a) Nurse

(b) Care Staff



Notes: Figures 2a and 2b show the binscatter of the patient-to-nurse ratio and the patient-to-care staff ratio versus annual average occupancy rate, using facility-fiscal year observations with at least 50% occupancy. The numbers of nurses and care staffs are full-time equivalent.

per patient at higher occupancy rates. Given that staffing levels are not adjusted on an annual basis, short-term staffing adjustments are unlikely to affect facility capacity, which is consistent with the shortages of care workers noted in Section 2.1.

5 Empirical Strategy

5.1 Effect of Occupancy on Admissions and Discharges

A naive way to test the predictions of Propositions 1-3 and Corollary 1 is to regress admissions and discharges on occupancy rate and examine its coefficients. However, this approach suffers from endogeneity because occupancy may be affected by facilities' unobserved pre-determined quality or operational efficiencies that also affect admissions/discharges (see Section 3.5). We address this problem by exploiting patient deaths. When a patient dies in a facility, she is discharged from the facility, which reduces occupancy on the following day.[28] Patient deaths lead to a marginal change in occupancy, which corresponds to the marginal change in n in Propositions 1-3 and Corollary 1. The key assumption is that the timing of patient deaths is unrelated to the preferences of live patients and exogenous to facility decisions.

5.1.1 Event Study

We first use an event study regression to examine the timing of the response, in particular whether the facilities respond immediately to occupancy shocks. It also allows us to test for the presence of a pre-trend in admissions/discharges, which would likely indicate a violation of our identification assumption.

Specifically, we estimate the following regression:

$$Y_{jt} = \sum_{k=-84}^{84} \beta_k Deaths_{jt-k} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{6}$$

where Y_{jt} denotes the number of admissions or live discharges of long-stay patients in facility j on date t. $Deaths_{jt}$ denotes the number of patients who died on date t in facility j. To normalize the scale of admissions and discharges across facilities, we divide these variables by the number of beds in the facility. Thus, they are measured as a percentage point (pp) change in occupancy. The parameter of interest is β_k, which represents the "effect" of patient deaths on the number of admissions or live discharges k days before ($k \leq 0$) or after ($k > 0$) the death event.[29] We estimate β_k from 84 days (12 weeks)

[28] Recall that discharges lower the occupancy rate on the following day (see Section 4.1).

[29] Because multiple patients may die in a day, our specification corresponds to an event study design for multiple events with different intensity (Schmidheiny and Siegloch, 2023).

before and after patient deaths. The regression model includes fiscal year-specific facility fixed effects γ_{jy} (which control pre-determined quality, staffing, equipment, and other unobservable factors at the facility-fiscal year level) and date fixed effects γ_t. Control variables X_{jt} include average age, female ratio, the share of patients whose coinsurance rate is strictly above 10% (indicating relatively high income), average care levels, and the share of patients receiving terminal care in facility j on date t. Standard errors are clustered at the municipality level because each municipality is an insurer of LTCI and error terms may be correlated within insurers.

Regression (6) can be provided with a theoretical foundation using the framework in Section 3, under additional assumptions. Suppose (i) the admission/discharge functions are linear (at least locally): $Y_{jt} = \alpha_j + \alpha_t + \alpha^Y n_{jt}$, where Y denotes a (admission) or d (live discharge), and (ii) occupancy evolves as $n_{jt+1} = n_{jt} + a_{jt} - d_{jt} - \Delta_{jt}$ where Δ denotes deaths. Then, successively substituting n_{jt} yields Y_{jt} as a function of the lags of Δ and a residual. Adding leads of deaths (for pre-trend analysis) and decomposing the residual using controls and fixed effects yields Eq.(6). Proposition 1 yields testable implications for the regression parameters: for $k > 0$, (a) $\beta_k \in (0, 1)$ for admissions, (b) $\beta_k \in (-1, 0)$ for discharges, and (c) $|\beta_k|$ decreases with k.

5.1.2 Instrumental Variables Estimation

In the main analysis, we conduct instrumental variables (IV) regressions of admissions and discharges, using patient deaths as an IV for occupancy rate. The effect of occupancy rate on admissions and discharges is specified as:

$$Y_{jw(t)} = \beta OC_{jt} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{7}$$

where $Y_{jw(t)}$ denotes the number of admissions or live discharges (in pp) of facility j in the week(s) $w(t)$ following date t.[30] We mainly examine the outcomes in the first one or four week(s) following t, though we also show baseline results for up to 12 weeks. OC_{jt} denotes the occupancy rate of facility j at the beginning of date t, before the facility admits or discharges patients. The parameter of interest, β, denotes the effect of a 1 pp increase in occupancy on outcomes. This allows us to investigate how the frictions in adjusting admissions and discharges differ and how they vary over different time horizons

[30] For example, $Y_{jw(t)}$ may represent the number of admissions during day t through $t + 6$ (one week) or admissions during day t through $t + 27$ (four weeks).

(Propositions 1 and 2). We control for the average number of deaths in the four weeks prior to the date (to capture heterogeneous mortality trends), as well as control variables included in the event study.

The first-stage regression is

$$OC_{jt} = \alpha Deaths_{jt-1} + \tilde{\lambda}' X_{jt} + \tilde{\gamma}_{jy} + \tilde{\gamma}_t + \tilde{\varepsilon}_{jt}, \tag{8}$$

where $Deaths_{jt-1}$ denotes the number of patient deaths (in pp) on date $t-1$.

5.1.3 Effects by Baseline Occupancy

Corollary 1 states that if variation in the cost effect (income effect) is the main driver of variation in admission/discharge responses to occupancy shocks, then the responses will be more (less) intense at higher occupancy rates. We test this prediction by estimating Eq.(7) separately at different occupancy levels (below 85%, 85-90%, 90-95%, and above 95%) and comparing responses across groups. This exercise is conceptually similar to regressing $-\frac{\partial a}{\partial n}$ on n to examine the sign of $\frac{\partial}{\partial n}\left(-\frac{\partial a}{\partial n}\right)$.

However, there are two concerns. First, the comparison of admission responses across occupancy levels may be confounded by heterogeneity that causes both higher baseline occupancy and larger admission responses, rather than by differential congestion. For example, higher-quality facilities may have both higher occupancy and larger responses, because many patients have decided or are ready to be admitted. Therefore, we also conduct an alternative regression analysis that exploits an arguably exogenous variation in baseline occupancy levels. Specifically, we estimate the following regression model:

$$Y_{jw(t)} = \beta_1 OC_{jt} + \beta_2 OC_{jt} \times I\left\{OC_{jt} \geq L^o\right\} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{9}$$

where L^o is an occupancy cutoff. As IVs, we use $Deaths_{jt-1}$ and $Deaths_{jt-1} \times TotalDeaths_{jt-1}$, where $TotalDeaths_{jt-1}$ denotes the total number of deaths in the K weeks preceding day $t-1$.[31] β_2 measures how the magnitude of the admission response to occupancy changes when a facility is exogenously assigned to a higher baseline occupancy level. The identification idea is that deaths prior to day $t-1$ affect the baseline occupancy on day $t-1$ without shifting demand on day $t-1$, at least if we focus on a relatively short period prior to $t-1$.

[31] We use $L^o = 95\%, 90\%, 85\%$ and $K = 2$.

Second, facilities may differ in what they consider to be high occupancy, e.g., due to heterogeneous cost functions. In such a case, a comparison across occupancy levels may reflect heterogeneity other than congestion. To address this concern, we implement an alternative classification of occupancy groups. Specifically, we compute facility-specific quartiles of occupancy and then classify observations into the following groups: below 25th percentile, 25-50th percentile, 50-75th percentile, and above 75th percentile. This allows us to examine how admission responses differ when baseline occupancy becomes high relative to each facility's standard.[32]

5.2 Validity of Identification Assumptions

The key identification assumption for Eq.(6) and (7) is that the timing of patient deaths is exogenous to confounding factors that affect admissions or discharges. Because we control for fiscal year-specific facility fixed effects and date fixed effects, our estimates are unaffected by different admission tendencies across facilities or time-specific shocks to demand for in-facility care. Identification can be challenged by unobserved confounders that affect both trends in admissions/discharges and trends in patient deaths. For example, facilities may increase admissions and skimp on necessary care, both motivated by increasing concerns about profits. We investigate this possibility by testing for a common pre-trend in the number of admissions or discharges between facility-dates that face patient deaths and those that do not. We also note that our estimation exploits variation in deaths residualized by covariates, including the share of patients receiving terminal care and the number of deaths in the previous month, as well as fixed effects. The residual variation in deaths is likely to be exogenous and unexpected, at least at the daily level.

We also examine the concern whether deaths and admissions/discharges are both correlated with facility-level health shocks, such as influenza outbreaks. To check this possibility, we estimate Eq.(6) using the number of discharges to hospital for acute care (hospitalizations) as an outcome. Since GHSFs do not provide acute medical care (see Section 2), facility-level health shocks would be expected to result in increased hospitalizations among in-facility patients. Appendix Figure A1 shows that the trend in hospital-

[32] In estimating separate regressions by baseline occupancy levels, fixed effects account for heterogeneity in facilities' "standard" occupancy within each occupancy group, but they do not account for heterogeneity related to facilities' assignment to occupancy groups. Regression by occupancy percentiles will mitigate the latter problem, by exploiting variation in assignment to facility-specific high vs. low occupancy groups.

izations does not change around the timing of patient deaths, which suggests that patient deaths are unlikely to be correlated with facility-level health shocks.

Another concern in estimating Eq.(7) is that death events may directly affect admissions and discharges, violating the exclusion restriction. For example, death events may cause live patients to avoid admission or seek sooner discharge because they signal the poor pre-determined quality of the facility. Such effects are likely negligible for admissions because (i) there is a time lag between applications and both admissions and death events (which can be more than a month), and (ii) it is difficult for applicants to gather information about daily in-facility deaths. Patients are also unlikely to hasten discharge in response to death events, especially in the short run, because discharge requires advance planning. Also, they are unlikely to update their beliefs about facility quality based on daily (not long-run aggregate) deaths residualized by detailed covariates and fixed effects.

Alternatively, patient deaths may create additional burdens for staff, such as extra paperwork and emotional strain. These burdens could induce facilities to reduce their workload by deferring new admissions or expediting discharges, which could violate the exclusion restriction. First, it should be noted that the outcomes in Eq.(7) are admissions/discharges aggregated over several weeks *after* death events, which are unlikely to be affected by the burden associated with patient deaths on a specific past day. Second, we evaluate the concern by examining the effect of death events on same-day admissions/discharges. Death-related burdens are considered significant on the day of death events, whereas the recorded occupancy rate is unaffected at that time because patient deaths affect occupancy rates on subsequent days. In Section 6.1.1, we show that the overall pre-trend in admissions/discharges is statistically indistinguishable from zero, which equals the effect of patient deaths on same-day admissions/discharges (β_0 in Eq.(6), after normalization). This result implies that patient deaths have little effect on same-day admissions/discharges, and that the death-related burden is negligible in facility decisions.[33]

As described in Section 5.1.3, when we estimate Eq.(7) by occupancy levels, another concern is that assignment to different baseline occupancy levels is non-random, due to demand shocks or cost shocks. We address these concerns by two exercises. First, we estimate an alternative regression Eq.(9), which uses patient deaths in the past two weeks as a source of exogenous variation in baseline occupancy. The two-week patient deaths should be exogenous to the pre-determined quality of facility: given that long-term care is labor-

[33]The direct effects via patient preference or additional burden, if any, would make our estimates conservative.

intensive and staffing adjustments are inflexible (see Section 4.2), facilities' pre-determined quality is unlikely to fluctuate in such a short period and affect patient deaths.[34] Second, we estimate Eq.(7) and (9) by facility-specific occupancy percentiles rather than levels, to exploit within-facility variation in assignment to high vs. low occupancy.

6 Results

6.1 Effects of Occupancy on Admissions and Discharges

6.1.1 Event Study Results

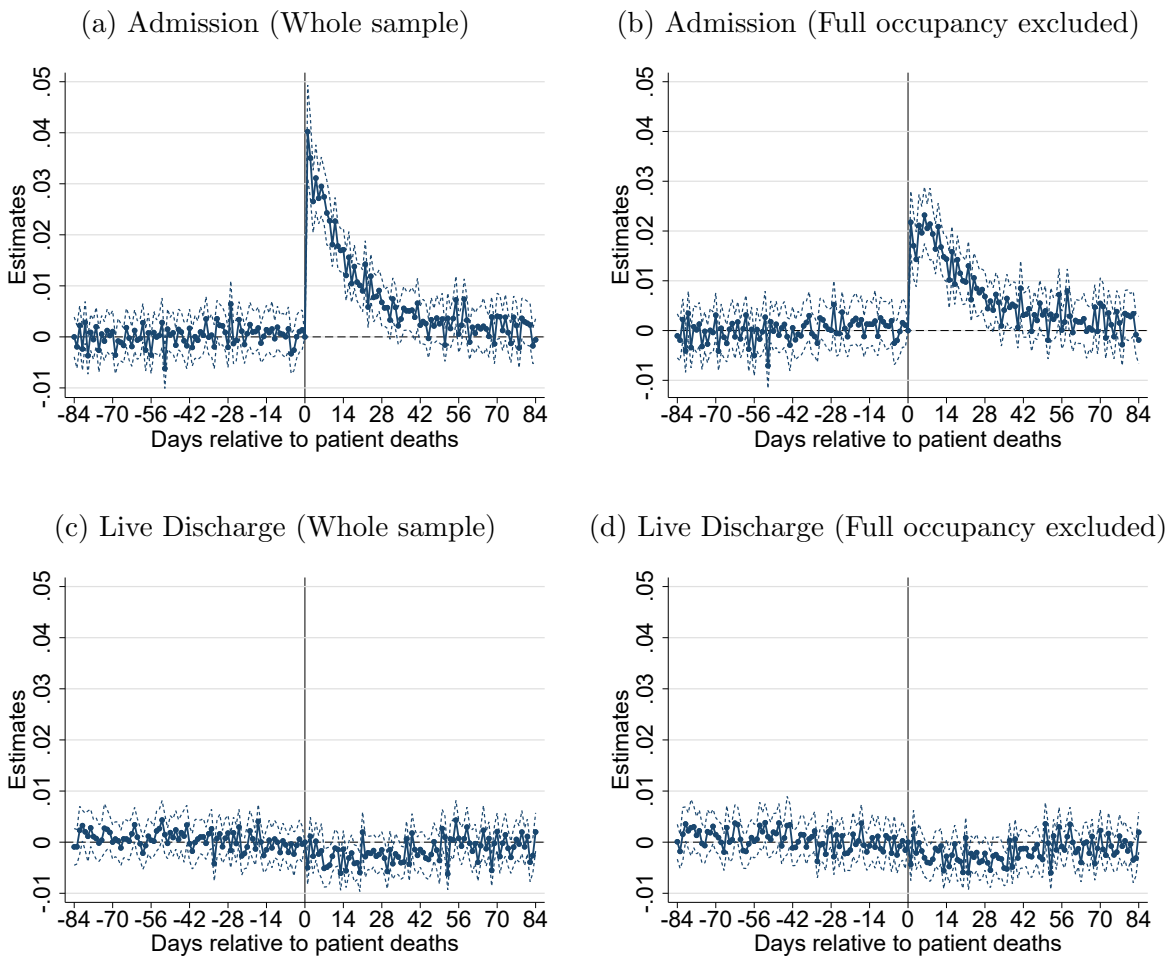
Figure 3 plots the estimates of β_k's in the event study regression (6), which represent the effects of patient deaths on the number of admissions or live discharges measured as a pp change in occupancy. The parameter for the day of patient deaths is normalized to zero ($\beta_0 = 0$). Figure 3a displays the results for admissions using the full sample. The estimates before patient deaths are close to zero and show no pre-trend, supporting the identification assumption for the event study design. This is reasonable given that nursing facilities are unlikely to possess medical technology that could manipulate or predict the timing of patient deaths. Figure 3a also shows that admissions increase significantly immediately after patient deaths. The increase begins on the next day of patient deaths and persists for about a month. Given the long admission process, the immediate responses are likely driven by admissions of patients who are ready and waiting to be admitted. Note that the estimates confirm the implications of the theoretical framework, as described in Section 5.1.1: for $k > 0$, $\beta_k \in (0, 1)$ and its magnitude tends to decrease with k.

To eliminate the mechanical effect of binding capacity constraints, Figure 3b shows the same regression result using observations for which capacity constraints are not binding. Figure 3b shows that even such facilities respond to patient deaths by increasing admissions. However, the magnitude of the response is smaller than that shown in Figure 3a, probably because the latter includes the mechanical effect of binding capacity constraints in addition to the effects of increasing marginal costs and demand inducement for non-binding cases.

Figures 3c and 3d show the results for live discharges using the full sample and the

[34] Nevertheless, patient deaths in the past two weeks are more likely to be endogenous than those in the previous day. We try patient deaths in the past week instead and obtain similar results, although the first stage becomes weaker.

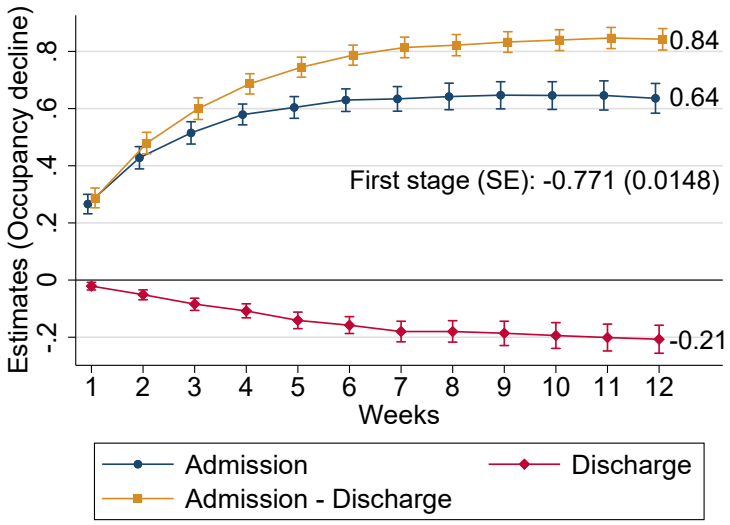
Figure 3: Effect of Patient Deaths on Admissions and Live Discharges



Notes: Figure 3 plots estimates of β_k coefficients from Eq.(6), which is a regression of the number of admissions or live discharges on the number of patient deaths, fiscal year by facility fixed effects, date fixed effects, and other controls. The estimate of β_k on the day of patient deaths is normalized to zero. Standard errors are clustered at the municipality level, and dotted lines show the 95% confidence intervals.

sample with non-binding capacity constraints, respectively. Neither shows a pre-trend. In contrast to admissions, the live discharges decrease after patient deaths, but very slightly. The different responsiveness of admissions and discharges to patient deaths can be explained by different frictions in adjusting admissions and discharges. As described in Section 2.2, discharging a patient requires an in-advance planning, including consultation with the patient and their family. Flexibly changing a patient's planned discharge date may be difficult, and adjusting discharges is likely to be costly. On the other hand, adjusting admissions may be less costly if there are patients who are ready to be admitted.

Figure 4: Effect of Empty Beds on Admissions and Live Discharges Over Time



Notes: Figure 4 plots the estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (blue curve), live discharges (red), and admissions minus live discharges (orange) for the following T week(s), $T = 1, ..., 12$. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors are clustered at the municipality level.

6.1.2 IV and OLS Estimates

Figure 4 plots the IV estimates of β in Eq.(7), with the sign reversed to indicate the effect of a 1pp *decrease* in occupancy. For example, a 1pp decrease in daily occupancy leads to a 0.64pp increase in admissions (blue curve) and a 0.21pp decrease in discharges (red) of long-stay patients over the following 12 weeks, implying that 84%(=64%+21%, with rounding) of the vacated beds are filled during this period (orange). Unlike the event study results in Figure 3, the decrease in discharges is statistically significant, probably because the outcome is aggregated over a longer period. The total response to a reduction in occupancy, $-\frac{\partial a^*}{\partial n} - (-\frac{\partial d^*}{\partial n})$, is positive but less than one, consistent with the theoretical predictions for the frictional case (Proposition 1-(ii)) rather than the frictionless case (Proposition 2). Again, admission responses are larger than discharge responses, suggesting that discharge frictions are larger. Admission and discharge responses tend to increase over time, suggesting that short-run adjustment is more frictional.[35]

Appendix Figure A2 plots the OLS and IV estimates of the coefficients on occupancy in the regression Eq.(7). Although both (reversed) estimates are positive for admissions

[35]The magnitude of responses is not monotonically increasing in weeks, because the admissions and discharges in the "control group" change over time as well.

and negative for discharges, the IV estimates are larger than the OLS estimates, especially for admissions. This suggests that the OLS under-estimates the effect of an occupancy reduction, probably because congested facilities tend to admit more patients.

6.2 Heterogeneous Effects by Occupancy Levels

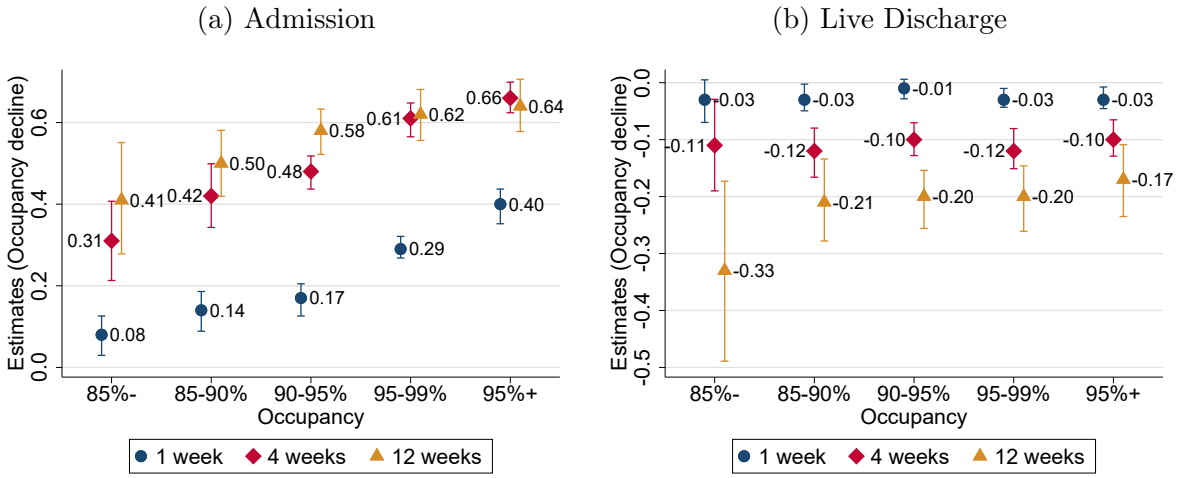
We now estimate Eq.(7) separately by occupancy rates on day $t-1$. Figure 5a plots the coefficients on occupancy from the regressions of 1-week, 4-week, and 12-week admissions, for each occupancy group: below 85%, 85-90%, 90-95%, and above 95%. To show that the result for the last group is not entirely driven by a mechanical admission responses from fully occupied facilities, we also show the results for facilities with occupancy strictly below 1 (labeled "95-99%"). The figure suggests that the admission responses tend to be larger at higher occupancy rates. The pattern holds whether or not we drop observations with binding capacity constraints. The increasing (in occupancy) response is consistent with the cost effect being a more important driver of variation in admission responses than the income effect (Corollary 1). In contrast, Figure 5b shows no systematic relationship between the discharge responses and baseline occupancy. This may be because discharges are less manipulable for capacity utilization management, and thus less reflective of facility incentives, than admissions. The qualitative results are similar when we group observations by occupancy percentiles rather than levels (see Appendix Figure A3).

Note that the differences in admission responses between occupancy levels shrink for longer-run outcomes, especially 12 weeks. This is possibly because admissions approach the steady-state level in the long run. We therefore focus on the effects on 1-week or 4-week admissions and discharges, interpreted as short-run outcomes, in what follows.

Panel A of Table 3 presents the estimates from Eq.(9). Because there is no systematic heterogeneity in the discharge responses across baseline occupancy levels, we only show the results for admissions. The table shows that the positive responses to an occupancy reduction are larger when the baseline occupancy is higher.[36] The same pattern holds

[36]The coefficient on uninteracted occupancy denotes the admission response of the low-occupancy group, which seems to be somewhat larger than the estimates shown in Figure 5a. For example, the regression result of 1-week admissions with an interaction term for occupancy above 85pp (third column in Panel A of Table 3) implies that the admission response of the group with occupancy below 85% is 0.26, whereas the admission response of the same group shown in Figure 5a is 0.08. This is likely because the regression Eq.(9) pools various occupancy groups considered in Figure 5a into a high- or low-occupancy group, making the estimated difference between the groups (captured by the coefficient on the interaction term) smaller and adjusting the estimates of low-occupancy admission responses (captured by the coefficient on the uninteracted occupancy) upward. The purpose of estimating regression Eq.(9)

Figure 5: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Level



Notes: Figure 5 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (Figure 5a) or live discharges (Figure 5b) for the following 1 week (blue), 4 weeks (red), and 12 weeks (orange), separately by baseline occupancy level. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors are clustered at the municipality level.

when we exclude observations with binding capacity constraints, although some interactions are statistically insignificant (see Panel A of Appendix Table A3). Again, the results are qualitatively similar when we use occupancy groups based on percentiles instead of levels (see Panel B of Tables 3/A3).

Our results are qualitatively similar when we lift the sample restriction that excludes observations with extreme values of occupancy, as mentioned in Section 4 (see Appendix Figure A5 and Table A4). The results are also similar when we instead strengthen the sample restriction, by dropping facilities with the maximum occupancy below 50% from our main sample (see Appendix Figure A6 and Table A5).

6.3 Occupancy and Patient Selection

Next, we test for selective admissions (Gandhi, 2023) by examining whether the composition of admission responses differs at different occupancy levels. Figure 6 plots the coefficients for 1-week admissions of each care level, separately by baseline occupancy levels. We divide the care level-specific coefficient by the coefficient for total admissions within each occupancy level, so that the numbers indicate the share of the admission

is to show that different identification assumptions lead to similar qualitative results, and we will mainly refer to the results shown in Figure 5a for the rest of this paper.

Table 3: Effect of Empty Beds on Admissions, with Nonlinear Terms

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.307***	0.291***	0.257***	0.614***	0.600***	0.573***
	(0.0266)	(0.0218)	(0.0161)	(0.0248)	(0.0218)	(0.0186)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.0517***			0.0424***		
	(0.0129)			(0.0140)		
I(Occupancy≥ 90pp)		0.0594***			0.0487***	
		(0.0137)			(0.0159)	
I(Occupancy≥ 85pp)			0.110***			0.0904***
			(0.0270)			(0.0296)
Cragg-Donald F-stats	164.0	141.4	78.05	164.0	141.4	78.05
N	6,673,744	6,673,744	6,673,744	6,673,744	6,673,744	6,673,744

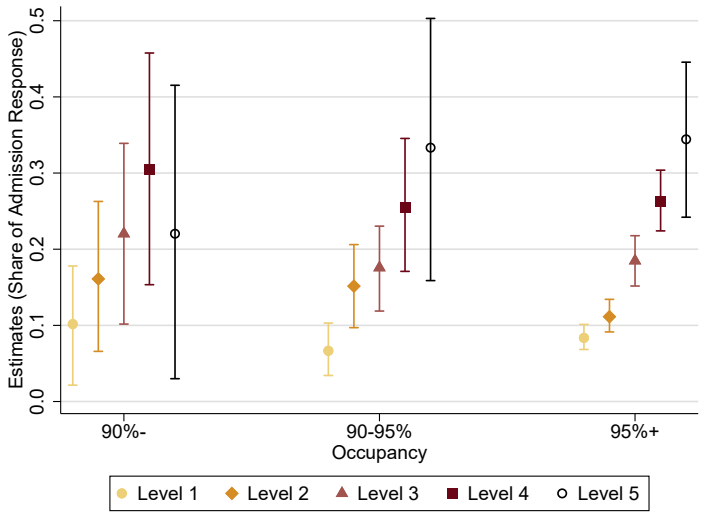
Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.311***	0.301***	0.282***	0.617***	0.609***	0.594***
	(0.0263)	(0.0232)	(0.0202)	(0.0246)	(0.0225)	(0.0204)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0317***			0.0260***		
	(0.00759)			(0.00832)		
I(Occupancy≥ 50pctile)		0.0291***			0.0238***	
		(0.00676)			(0.00759)	
I(Occupancy≥ 25pctile)			0.0445***			0.0365***
			(0.0103)			(0.0117)
Cragg-Donald F-stats	252.2	300.3	215.1	252.2	300.3	215.1
N	6,673,744	6,673,744	6,673,744	6,673,744	6,673,744	6,673,744

Notes: Table 3 shows the estimates of the coefficients in the regression (9), using 1-week or 4-week admission as the outcome and deaths in the previous 2 weeks to construct an instrument for the interaction. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors clustered at the municipality level are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

response accounted for by the particular care level. The figure suggests that, although admissions of higher care levels tend to account for a larger share, the composition does not change systematically with baseline occupancy levels. This suggests that variation in occupancy does not induce facilities to select patients with different care levels. This pattern also holds when we use 4-week admissions (see Appendix Figure A4).

Figure 6: Effect of Empty Beds on Admissions, By Occupancy and Care Levels



Notes: Figure 6 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using long-stay admissions for the following 1 week as the outcome, separately by occupancy and care levels. The coefficients are divided by the coefficient in the baseline regression that pools all care levels, so the numbers represent the share of each care level in the admission response. Standard errors are clustered at the municipality level.

6.4 Interpretations of Admission Responses

The results of the admission responses by occupancy levels are nontrivial, because they allow us to infer the mechanisms underlying the admission responses. Specifically, we interpret the positive and increasing (in baseline occupancy) admission responses to occupancy reductions, shown in Section 6.2, as driven by variation in cost effects. The response pattern is unlikely to be explained by variation in income effects, because responses would then decrease with occupancy under plausible utility functions.

Cost effects are also likely to explain a large part of the overall level of admission responses. The discussion in Section 3.3 suggests that the income effect for 1-week admissions is at most 0.08pp (i.e., the admission response at occupancy below 85%), so the cost effect at occupancy strictly between 95% and 100% is at least 0.21pp (= 0.29pp − 0.08pp), or 72.4% (= $\frac{0.21}{0.29}$) of the total response.[37] Moreover, the income effect is likely less than 0.08pp, because demand inducement is difficult in the short run. Thus, the

[37]The lower bound for the cost effect is even higher if we include observations with binding capacity constraints. Conversely, if we assume that capacity constraints are binding at the occupancy rate of 95% (due, e.g., to reserved beds or labor shortages) and focus on the cost effect at occupancy between 90% and 95%, then the lower bound is lower (0.17pp − 0.08pp=0.09pp) but still a substantial fraction of the total admission response ($\frac{0.09}{0.17} = 52.9\%$).

cost effect will account for a non-negligible part of admission responses at all occupancy levels. Even for 4-week admissions, for which demand inducement is likely easier, the cost effect accounts for at least 0.30pp (= 0.61pp − 0.31pp) or 49.2% ($=\frac{0.30}{0.61}$) of the admission response at occupancy strictly between 95% and 100%.

The results by occupancy levels are also informative about the mechanisms behind the cost effect. Some portion of the cost effect is likely due to increasing marginal costs rather than binding capacity constraints, because the latter would imply flat admission responses except for facilities with binding capacity constraints (in contrast to the increasing responses shown in Figure 5a). On the other hand, the 1-week admission response for the 95-99% occupancy group is significantly smaller than the response for the 95%+ group (Figure 5a), suggesting that binding capacity constraints (Boehm and Pandalai-Nayar, 2022) also play some role in the response, at least in the short run.

We infer that increasing marginal costs will explain some part of admission responses even after we account for facilities' dynamic considerations on capacity constraints (Gandhi, 2023). Unlike the U.S. SNFs studied by Gandhi (2023), which accept patients with different profitability, reimbursement in our setting is based on a universal, care needs-adjusted system. Indeed, we find no evidence of occupancy-induced selection on care levels (Figure 6).[38]

Several reasons may explain why the marginal cost of service increases with occupancy. First, congestion reduces the quality of service, thereby reducing the altruistic utility of the facility. Lower service quality can also result from worker burnout or management slowdowns that prevent efficient care delivery. Second, medical resources may be allocated in order of decreasing productivity, making it costly to serve additional patients at high occupancy rates. Third, the marginal cost of hiring and retaining staff is likely to increase with occupancy. As occupancy increases, facilities will need to pay employees overtime and at a higher rate. In addition, higher occupancy may worsen working conditions, making additional hiring more difficult. Figure 2 is consistent with some of the explanations: the patient-to-staff ratio (a measure of inverse quality and that of the harshness of the work environment) tends to be higher at higher occupancy.

[38]Some beds may be reserved for emergency admissions, effectively imposing capacity constraints strictly below physical bed capacity. Note, however, that our analysis exploits within-facility variation in occupancy levels/percentiles. Thus, as long as the number of reserved beds is constant within each facility, the admission responses at lower occupancy cannot be explained by binding capacity constraints.

7 Policy Implications for Healthcare Capacity

Our conceptual framework and empirical results yield a nontrivial and important insight that a large fraction of admission increases in response to available capacity are driven by cost effects rather than income effects, at least for congested facilities. Many studies have emphasized that available capacity itself leads to costly care provision due to providers' financial incentives, e.g., to induce demand (e.g., Freedman, 2016).[39] On the other hand, a small number of other studies suggest that capacity constraints deter service provision (e.g., Ching et al., 2015).[40] The relative importance of these effects has not been investigated in a unified framework. Understanding the mechanisms behind the effects of capacity on care provision has important implications for capacity policies. If capacity affects care delivery mainly through income effects, an idea emphasized in the vast literature on supplier-induced demand, then capacity expansion would increase wasteful care provision.[41] In an extreme case where the entire admission responses are due to wasteful income effects, we would conclude that additional capacity is the *least* valuable for the most congested facilities because they exhibit the largest income effects, which is the opposite of our conclusion. In contrast, our framework and empirical results suggest that cost effects are the main driver of the admission responses to capacity expansion at congested facilities. This suggests that the incremental admissions are relatively beneficial at congested facilities, and that such facilities should be prioritized when the government considers policies to expand healthcare capacity.

Under some assumptions, the above insight can extend to policy discussions on the assignment of market-level healthcare capacity. Healthcare capacity is typically not planned at the level of individual facilities, but at a broader market level, such as municipalities, considering the population characteristics in the market. Due to limited resources, it is

[39] There can also be dynamic financial incentives to forgo admitting a patient in order to admit a more profitable patient in the future (Freedman, 2016; Gandhi, 2023). As discussed in Section 6.4, such incentives seem relatively weak in our context.

[40] Although Freedman (2016) emphasizes the importance of financial incentives, e.g., to induce demand (also expressed as "income effects" in p.159), he also discusses other mechanisms, including congestion externalities (similar to congestion costs considered in this paper). However, he does not formally assess the relative importance of the mechanisms. Also related are studies by Gandhi (2023) and Hackmann et al. (2024), who document how capacity constraints affect access to care, though their findings are focused on whose access is compromised rather than the overall level of access reduction.

[41] Capacity includes beds, other medical equipment, and staffing. Additional capacity will increase fixed costs (e.g., wages), so it will incentivize facilities to induce demand, just as lower patient volume will. Our theoretical framework unifies both shocks as negative occupancy shocks, so it is applicable to examining policies to expand capacity.

crucial to determine which markets should be prioritized for additional capacity assignment. For this purpose, we can compute the market-level cost effects and income effects (in response to a 1pp decrease in occupancy at each facility in the market) and compare markets based on these measures. To illustrate, suppose that the net social benefit of admissions due to cost effects is positive and weakly higher than the (positive or negative) net social benefit of admissions due to income effects.[42] Now, consider two markets, Market A and Market B, such that Market A has a larger admission response and a smaller income effect than Market B. Then, Market A should be prioritized over Market B for expanding capacity, because the former can achieve a larger increase in total admissions but a smaller increase in low-benefit admissions in response to a capacity expansion, implying larger total benefits under our assumption.[43] Thus, markets such that there is no other market with a larger admission response and a smaller income effect will be prioritized targets of capacity expansion policies. If we further assume the relative social values of cost and income effects, i.e., the "marginal rate of substitution" between the two effects, then we can compare markets in terms of the priority for capacity expansion, using an "indifference curve" of the policymaker. We further illustrate these two ideas in Online Appendix D.

Further analysis is required to inform broader policy decisions. To determine the overall level of capacity expansion (not just where capacity expansion is prioritized), we need various information on the costs and benefits of capacity expansion. They may include, for example, patient preferences for and health effects of admissions initiated by each of cost and income effects, costs of admissions and subsequent care, and public expenditures for increasing capacity (in beds, staffing, and equipment).[44] The preferences and

[42]In Section 6.3, we argue that admissions in response to occupancy reductions are similar across different baseline occupancy levels, in terms of a key profit-relevant characteristic. This suggests that cost and income effects lead to admissions of patients with similar such characteristics. Still, admissions by the two effects may differ in characteristics relevant to patient benefits. The assumption that admissions due to income effects (supplier-induced demand) are of relatively low value is based on numerous studies (Arrow, 1963; McGuire, 2000) that suggest that information asymmetry may cause medical providers to sacrifice patients' interests to prioritize their financial interests. On the other hand, Freedman et al. (2025) find that greater supply of neonatal intensive care units in low-access areas leads to more admissions and lower mortality of newborns with very low birth weight, which suggests that cost effects may improve health outcomes by serving high-needs patients with previously limited access to care. Although health services provided out of an imperfect agency relationship between a physician and a patient may still be effective (Labelle et al., 1994), cost effects will be more valuable, at least to patients, than income effects to the extent that the former are more closely aligned with patient preferences.

[43]This conclusion ignores the costs of capacity expansion. In Online Appendix D, we incorporate such costs under additional assumptions.

[44]There may also be externalities, such as the benefits to the family caregivers of the patients and the

health effects are difficult to study with the research design of this paper, which exploits daily occupancy fluctuations in a facility-level sample. Instead, studying preferences will require demand estimation with an episode-level sample and a method to account for choice set restrictions (caused by providers rejecting some applications) using only data on realized admissions. Also, health outcomes are influenced by patient characteristics and the courses of care during an episode (e.g., care intensity affected by congestion), so it is difficult to determine whether admissions initiated by income effects have lower health benefits than those initiated by cost effects without an episode-level sample. In a companion paper (Saruya and Takahashi, 2025), we study patient preferences and outcomes using an episode-level sample. That said, for more thorough analysis of the trade-offs in healthcare capacity, understanding the conflicting supply-side incentives — capacity constraints and supplier-induced demand — is essential. Our analytical framework provides a stepping stone for further investigation of this important issue.

8 Conclusion

Researchers and policymakers have expressed concern that excess capacity leads to wasteful care provision, while additional capacity may be valuable if valuable care is deterred by congestion costs and capacity constraints. We develop a framework to evaluate the relative importance of these factors in explaining nursing facility admission and discharge decisions. Combining the framework with Japanese long-term care claims data, we find that a marginal reduction in occupancy from baseline leads facilities to increase admissions, and that the admission responses are larger at higher baseline occupancy, consistent with congestion costs and capacity constraints driving the admission responses to occupancy variation. Our results have policy implications for determining which facilities (or markets) should be prioritized for healthcare capacity assignment.

There are some interesting extensions of our study. First, the framework is based on a static decision by a single-agent nursing facility. Dynamics and strategic interactions are unlikely to be highly important in our setting, due to the care needs-adjusted reimbursement system and excess demand. Still, those factors may play some role in shaping access to care in our setting and beyond, requiring an extension of our framework. Second, we do not analyze care outcomes, because they are difficult to study within the framework

congestion costs to patients already in facility.

of this paper, as discussed in Section 7. We study the effects of congestion on patient outcomes in a companion paper (Saruya and Takahashi, 2025).

Our conceptual framework and empirical results can serve as a stepping stone for discussing important policy issues. First, it provides insights on the optimal capacity regulations in health care. Despite the widespread use of policies to constrain supply capacity with the aim of preventing wasteful service provision (e.g., CON laws in health care) and other purposes, such policies can have an adverse effect of worsening access to services.[45] CON laws can even increase healthcare spending if capacity constraints prevent patients from accessing necessary care, worsening their health conditions to the point where they eventually need more intensive treatment.[46] Our findings are useful for discussing which facilities (or markets) should or should not be the target of strict capacity regulations. As discussed in Section 7, however, determining the optimal level of capacity regulation will require various information on the costs and benefits of capacity expansion, which is beyond the scope of this paper. Second, if inflexible staffing is the main source of the cost effects, then labor market reforms to enable more flexible adjustment of staffing can mitigate the negative effect of congestion. We leave analysis of staffing to future research.

References

Alexander, Diane, and Molly Schnell. 2024. "The Impacts of Physician Payments on Patient Access, Use, and Health." *American Economic Journal: Applied Economics* 16 (3): 142–77.

Arrow, Kenneth. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53 (5): 941–973.

Baker, Laurence C., Ciaran S. Phibbs, Cassandra Guarino, Dylan Supina, and James L. Reynolds. 2004. "Within-year variation in hospital utilization and its implications for hospital costs." *Journal of Health Economics* 23 (1): 191–211.

Baker, Laurence C., and Anne Beeson Royalty. 2000. "Medicaid Policy, Physician Behavior, and Health Care for the Low-Income Population." *Journal of Human Resources* 35 (3): 480–502.

[45] Mitchell (2024), based on a review of the literature on CON laws, concludes that "the data suggest CON limits access to care" (p.6).

[46] CON laws have another, more oft-cited, disadvantage that they can limit provider competition (Mitchell, 2024), which can also raise healthcare spending and worsen care outcomes.

Boehm, Christoph E., and Nitya Pandalai-Nayar. 2022. "Convex Supply Curves." *American Economic Review* 112 (12): 3941–69.

Buchmueller, Thomas C., Sean Orzol, and Lara D. Shore-Sheppard. 2015. "The Effect of Medicaid Payment Rates on Access to Dental Care among Children." *American Journal of Health Economics* 1 (2): 194–223.

Butters, R. Andrew. 2020. "Demand Volatility, Adjustment Costs, and Productivity: An Examination of Capacity Utilization in Hotels and Airlines." *American Economic Journal: Microeconomics* 12 (4): 1–44.

Cabral, Marika, Colleen Carey, and Sarah Miller. 2025. "The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid." *American Economic Journal: Economic Policy* 17 (1): 106–43.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *The Quarterly Journal of Economics* 112 (2): 407–441.

Care Work Foundation. 2016. "Nursing Care Labor Survey." https://www.kaigo-center.or.jp/content/files/report/h28_chousa_kekka.pdf (Japanese), Accessed July 31, 2025.

Carroll, Christopher D., and Miles S. Kimball. 1996. "On the Concavity of the Consumption Function." *Econometrica* 64 (4): 981–992.

Ching, Andrew T., Fumiko Hayashi, and Hui Wang. 2015. "Quantifying the Impacts of Limited Supply: The Case of Nursing Homes." *International Economic Review* 56 (4): 1291–1322.

Collard-Wexler, Allan. 2013. "Demand Fluctuations in the Ready-Mix Concrete Industry." *Econometrica* 81 (3): 1003–1037.

Corredor-Waldron, Adriana. 2022. "Spillover Effects of Medicare Policy on Medicaid: Evidence From the Nursing Home Industry." *Working Paper*.

Decker, Sandra L. 2007. "Medicaid Physician Fees and the Quality of Medical Care of Medicaid Patients in the USA." *Review of Economics of the Household* 5 95–112.

Decker, Sandra L. 2009. "Changes in Medicaid Physician Fees and Patterns of Ambulatory Care." *Inquiry* 46 291–304.

Dong, Jing, Pengyi Shi, Fanyin Zheng, and Xin Jin. 2020. "Structural Estimation of Load Balancing Behavior in Inpatient Ward Network." *Working Paper*.

Einav, Liran, Amy Finkelstein, and Neale Mahoney. 2025. "Producing health: measuring value added of nursing homes." *Econometrica* 93 (4): 1225–1264.

Evans, Robert G. 1974. "Supplier-Induced Demand: Some Empirical Evidence and Implications." In *The Economics of Health and Medical Care: Proceedings of a Con-*

ference held by the International Economic Association at Tokyo, edited by Perlman, Mark 162–173, London: Palgrave Macmillan UK.

Freedman, Seth. 2016. "Capacity and Utilization in Health Care: The Effect of Empty Beds on Neonatal Intensive Care Admission." *American Economic Journal: Economic Policy* 8 (2): 154–185.

Freedman, Seth, Lauren Hoehn-Velasco, and Diana R. Jolles. 2025. "Intensive care supply and admission decisions." *Journal of Health Economics* 100 102967.

Gandhi, Ashvin. 2023. "Picking your patients: Selective admissions in the nursing home industry." *Working Paper*.

Gaynor, Martin, and William B. Vogt. 2003. "Competition among Hospitals." *The RAND Journal of Economics* 34 (4): 764–785.

Grieco, Paul L. E., and Ryan C. McDevitt. 2017. "Productivity and Quality in Health Care: Evidence from the Dialysis Industry." *The Review of Economic Studies* 84 (3): 1071–1105.

Gruber, Jonathan, and Maria Owings. 1996. "Physician Financial Incentives and Cesarean Section Delivery." *The RAND Journal of Economics* 27 (1): 99–123.

Hackmann, Martin B., R. Vincent Pohl, and Nicolas R. Ziebarth. 2024. "Patient versus Provider Incentives in Long-Term Care." *American Economic Journal: Applied Economics* 16 (3): 178–218.

He, Daifen, and R. Tamara Konetzka. 2015. "Public Reporting and Demand Rationing: Evidence from the Nursing Home Industry." *Health Economics* 24 (11): 1437–1451.

Ikegami, Kei, Ken Onishi, and Naoki Wakamori. 2021. "Competition-driven physician-induced demand." *Journal of Health Economics* 79 102488.

Ilzetzki, Ethan. 2024. "Learning by Necessity: Government Demand, Capacity Constraints, and Productivity Growth." *American Economic Review* 114 (8): 2436–71.

Japan Association of Geriatric Health Services Facilities. 2015. "Geriatric Health Services Facility in Japan." https://www.roken.or.jp/wp/wp-content/uploads/2013/03/english_2015feb_A4.pdf, Accessed August 4, 2025.

Kim, Song-Hee, Carri W. Chan, Marcelo Olivares, and Gabriel Escobar. 2015. "ICU Admission Control: An Empirical Study of Capacity Allocation and Its Implication for Patient Outcomes." *Management Science* 61 (1): 19–38.

Kline, Patrick, and Enrico Moretti. 2014. "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs." *Annual Review of Economics* 6 (Volume 6, 2014): 629–662.

Labelle, Roberta, Greg Stoddart, and Thomas Rice. 1994. "A re-examination of

the meaning and importance of supplier-induced demand." *Journal of health economics* 13 (3): 347–368.

Lakdawalla, Darius, and Tomas Philipson. 1998. "Nonprofit Production and Competition." Working Paper 6377, National Bureau of Economic Research.

McGuire, Thomas G. 2000. "Physician Agency." *Chapter 9 in the Handbook of Health Economics* 461–536.

McGuire, Thomas, and Mark Pauly. 1991. "Physician response to fee changes with multiple payers." *Journal of Health Economics* 10 (4): 385–410.

Ministry of Health, Labor and Welfare. 2017. "Short-Stay Residential and Medical Care." https://www.mhlw.go.jp/file/05-Shingikai-12601000-Seisakutoukatsukan-Sanjikanshitsu_Shakaihoshoutantou/0000168704.pdf (Japanese), Accessed August 4, 2025.

Ministry of Health, Labor and Welfare. 2023a. "Geriatric Health Services Facilities." https://www.mhlw.go.jp/content/12300000/001131788.pdf (Japanese), Accessed June 2, 2025.

Ministry of Health, Labor and Welfare. 2023b. "Overview of Long-Term Care Cost Statistics." https://www.mhlw.go.jp/toukei/saikin/hw/kaigo/kyufu/23/dl/03.pdf (Japanese), Accessed June 2, 2025.

Mitchell, Matthew D. 2024. "Certificate of Need Laws in Health Care: Past, Present, and Future." *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* 61 1–11.

Nishiwaki, Masato. 2016. "Horizontal mergers and divestment dynamics in a sunset industry." *The RAND Journal of Economics* 47 (4): 961–997.

Okazaki, Tetsuji, Ken Onishi, and Naoki Wakamori. 2022. "Excess capacity and effectiveness of policy interventions: Evidence from the cement industry." *International Economic Review* 63 (2): 883–915.

Samiedaluie, Saied, Beste Kucukyazici, Vedat Verter, and Dan Zhang. 2017. "Managing Patient Admissions in a Neurology Ward." *Operations Research* 65 (3): 635–656.

Saruya, Hiroki, and Masaki Takahashi. 2025. "Congestion-Quality Trade-off: Evidence from Nursing Facilities." Working paper.

Schmidheiny, Kurt, and Sebastian Siegloch. 2023. "On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization." *Journal of Applied Econometrics* 38 (5): 695–713.

Shurtz, Ity, Alon Eizenberg, Adi Alkalay, and Amnon Lahad. 2022. "Physician workload and treatment choice: the case of primary care." *The RAND Journal of Economics* 53 (4): 763–791.

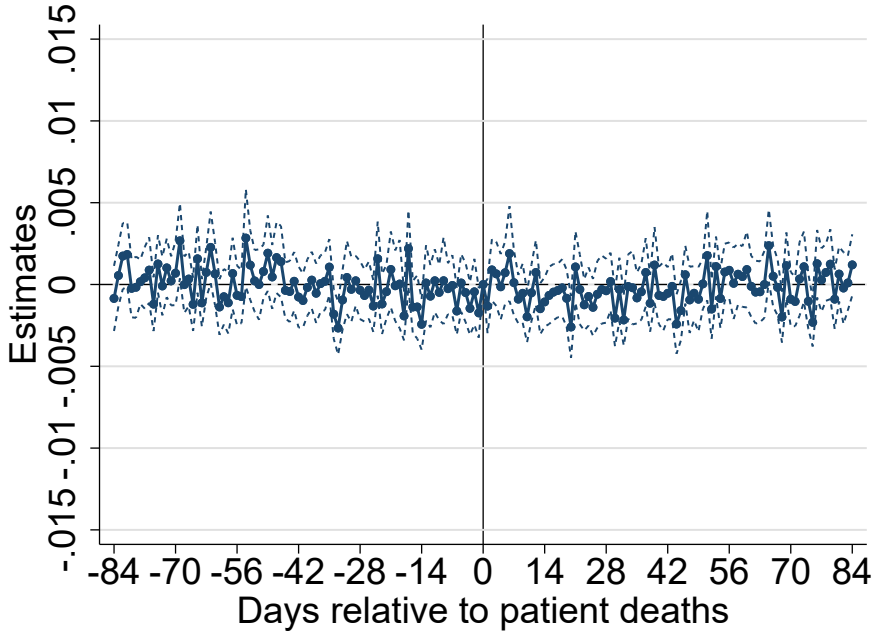
Takahashi, Yuya. 2015. "Estimating a War of Attrition: The Case of the US Movie Theater Industry." *American Economic Review* 105 (7): 2204–41.

United Nations. 2019. "World Population Aging 2019: Highlights." https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf, Accessed June 2, 2025.

Online Appendix

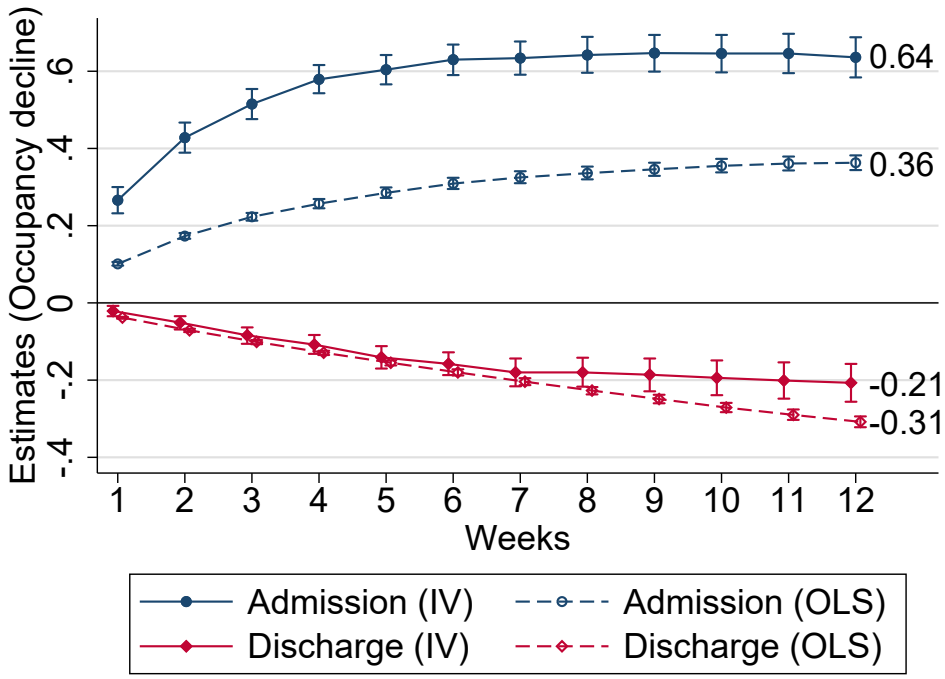
A Additional Figures and Tables

Figure A1: Effect of Patient Deaths on Hospitalization



Notes: Figure A1 plots estimates of β_k coefficients from Eq.(6), which is a regression of the number of hospitalizations on the number of patient deaths, fiscal year by facility fixed effects, date fixed effects, and other controls. The estimate of β_k on the day of patient deaths is normalized to zero. Standard errors are clustered at the municipality level, and dotted lines show the 95% confidence intervals.

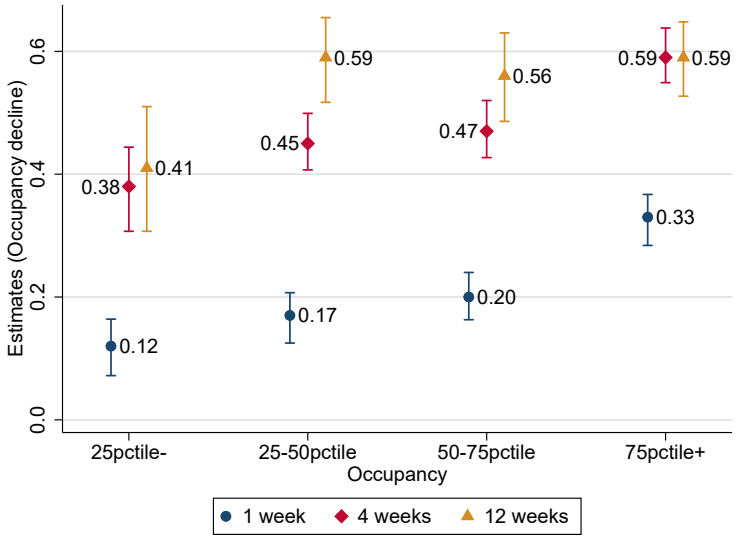
Figure A2: OLS vs IV Estimates of the Coefficient on Occupancy



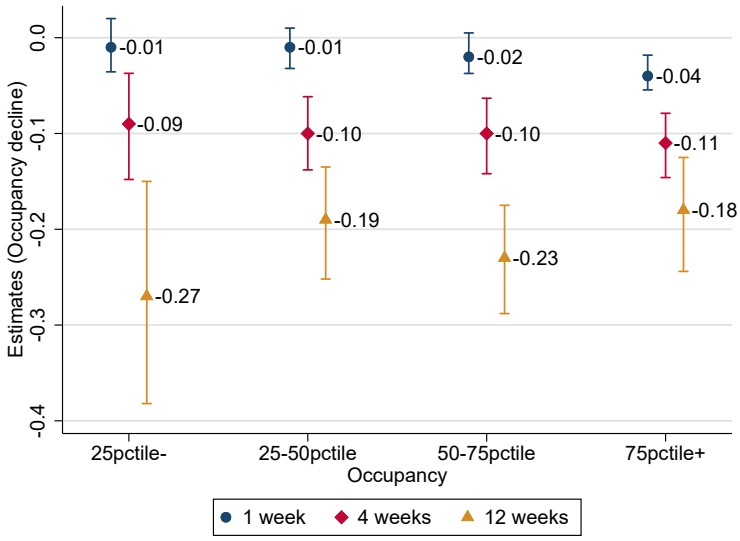
Notes: Figure A2 plots OLS (dotted curves) and IV (bold) estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (blue) and live discharges (red) for the following T week(s), $T = 1, ..., 12$. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors are clustered at the municipality level.

Figure A3: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Percentile

(a) Admission

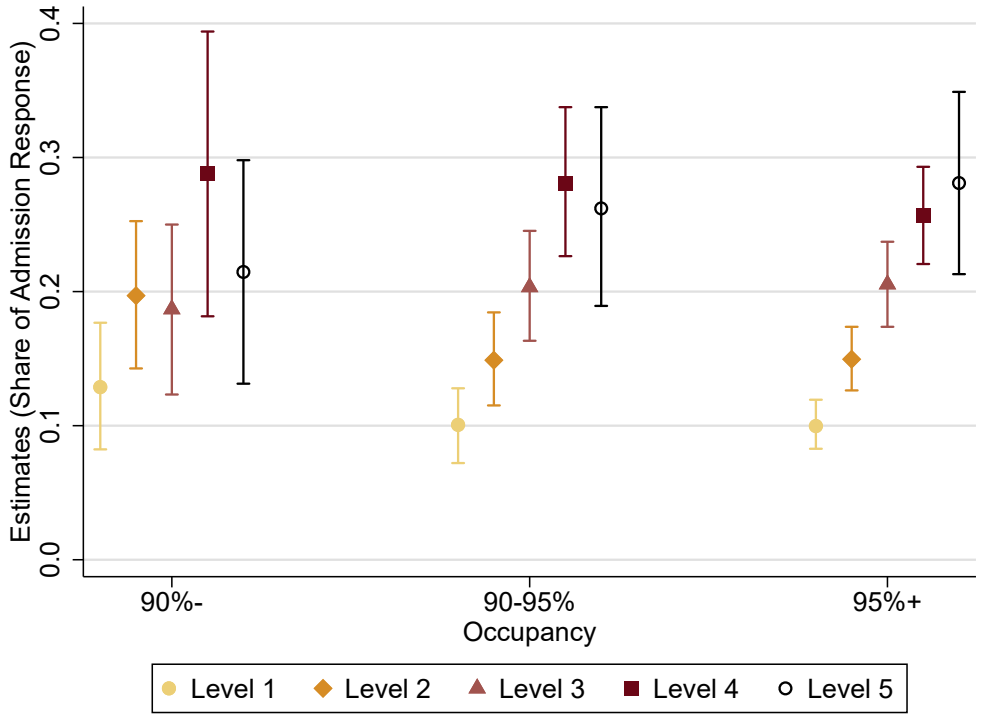


(b) Live Discharge



Notes: Figure A3 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (panel (a)) or live discharges (panel (b)) for the following 1 week (blue), 4 weeks (red), and 12 weeks (orange), separately by baseline occupancy percentile of each facility. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy.

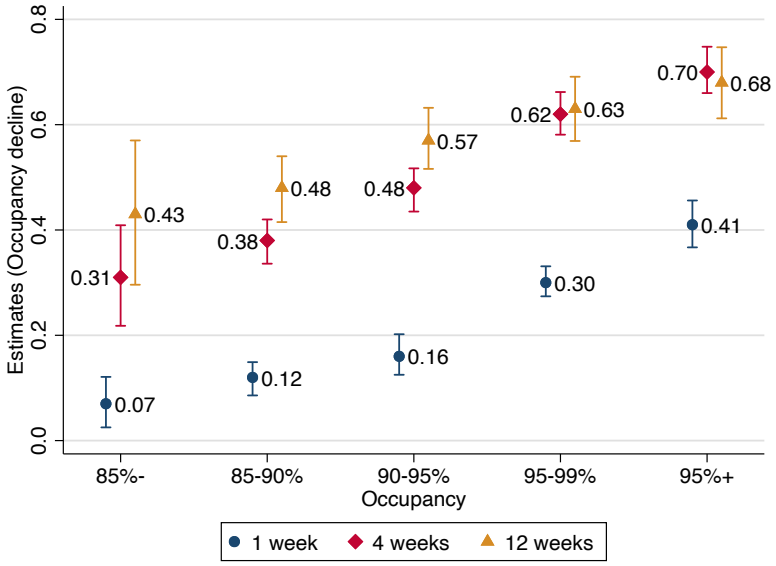
Figure A4: Effect of Empty Beds on Admissions, By Occupancy and Care Levels (4 Weeks)



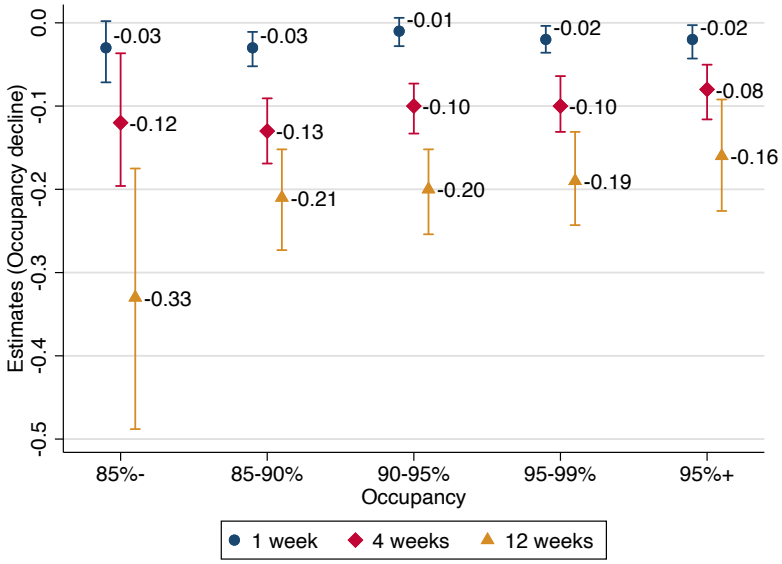
Notes: Figure A4 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using long-stay admissions in the following 4 weeks as the outcome, separately by occupancy and care levels. The coefficients are divided by the coefficient in the baseline regression that pools all care levels, so the numbers represent the share of each care level in the admission response.

Figure A5: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Level (Un-winsorized Sample)

(a) Admission



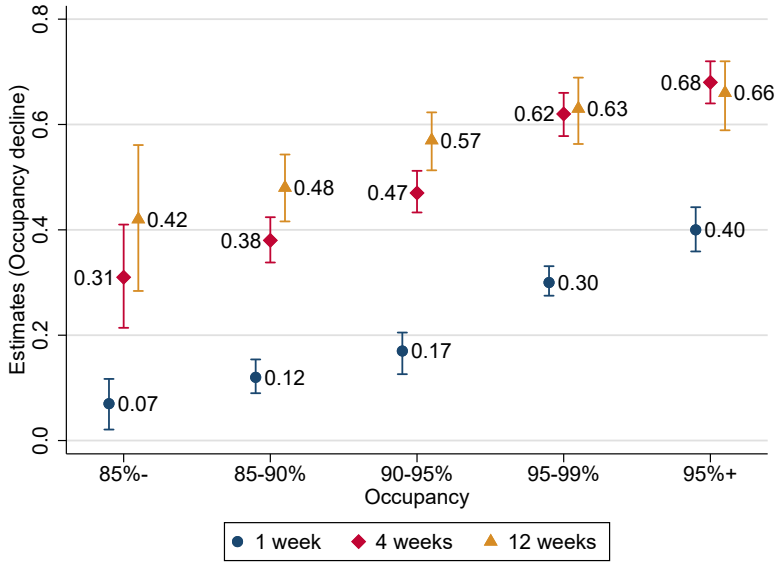
(b) Live Discharge



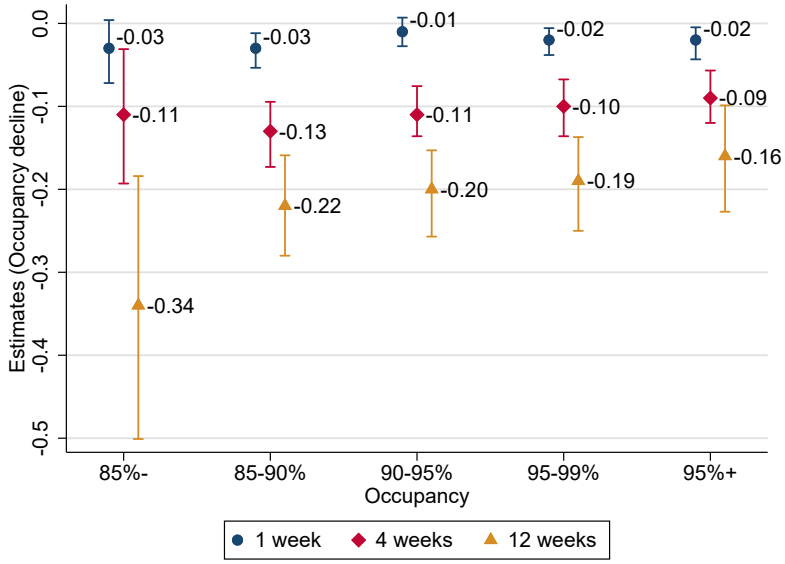
Notes: Figure A5 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (panel (a)) or live discharges (panel (b)) for the following 1 week (blue), 4 weeks (red), and 12 weeks (orange), separately by baseline occupancy level. The estimation sample does not exclude facilities whose maximum occupancy rate falls in the bottom or top 1 percentile of the distribution of maximum occupancy across providers. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy.

Figure A6: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Level (Maximum Occupancy $\geq 50\%$)

(a) Admission



(b) Live Discharge



Notes: Figure A6 plots the IV estimates of the coefficient on occupancy in Eq.(7) and their 95% confidence intervals, using as outcomes long-stay admissions (panel (a)) or live discharges (panel (b)) for the following 1 week (blue), 4 weeks (red), and 12 weeks (orange), separately by baseline occupancy level. The estimation sample exclude facilities whose maximum occupancy rate is lower than 50%. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy.

Table A1: Health Status of Each Care Level

Support level 1-2	The patient is able to perform most of the basic activities of daily living independently, but requires some assistance with complex activities of daily living.
Care level 1	The patient's ability to perform complex activities of daily living has declined further from the state of Support level.
Care level 2	In addition to the condition of care level 1, the patient requires assistance with basic activities of daily living.
Care level 3	Compared to the state of care level 2, there is a significant decline in terms of both basic and complex activities of daily living, and almost full nursing care is required.
Care level 4	In addition to the condition of care level 3, the patient's ability to move is further reduced and it becomes difficult for her to perform daily activities without assistance.
Care level 5	The patient's ability to perform activities of daily living is even worse than the state of care level 4, and it is almost impossible for the patient to perform daily activities without nursing care.

Notes: The table describes typical conditions for patients in each care level.

Table A2: Per-diem Reimbursement

	Fixed (USD) (1)	FFS (USD) (2)	% of fixed pay (3)
Care level 1	78.0	10.3	88.3%
Care level 2	83.1	10.4	88.9%
Care level 3	89.4	10.5	89.5%
Care level 4	95.2	10.6	90.0%
Care level 5	101.3	10.4	90.4%

Notes: The table presents daily averages of fixed and fee-for-service (FFS) payments for long-stay patients in our analysis sample, separately by care levels. The averages are computed through the following steps. (1) Compute the daily averages of fixed and FFS payments within each patient-year-month bin, by computing the total fixed and FFS payments within each bin and then dividing them by the total days of stay within the bin. (2) Aggregate the averages to the care level.

Table A3: Effect of Empty Beds on Admissions, with Nonlinear Terms (Exclude full occupancy)

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.731	0.282***	0.231***	1.089**	0.631***	0.579***
	(0.534)	(0.0407)	(0.0225)	(0.504)	(0.0473)	(0.0285)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.151			0.154		
	(0.140)			(0.133)		
I(Occupancy≥ 90pp)		0.0557***			0.0568**	
		(0.0187)			(0.0243)	
I(Occupancy≥ 85pp)			0.0850***			0.0868**
			(0.0277)			(0.0362)
Cragg-Donald F-stats	5.574	55.68	48.19	5.574	55.68	48.19
N	6,339,068	6,339,068	6,339,068	6,339,068	6,339,068	6,339,068

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.377***	0.321***	0.283***	0.617***	-0.609***	-0.594***
	(0.0803)	(0.0535)	(0.0411)	(0.0246)	(0.0225)	(0.0204)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0258***			0.0260***		
	(0.00975)			(0.00832)		
I(Occupancy≥ 50pctile)		0.0227***			0.0238***	
		(0.00787)			(0.00759)	
I(Occupancy≥ 25pctile)			0.0354***			0.0365***
			(0.0120)			(0.0117)
Cragg-Donald F-stats	87.50	123.3	101.6	87.50	123.3	101.6
N	6,339,068	6,339,068	6,339,068	6,339,068	6,339,068	6,339,068

Notes: Table A3 shows the estimates of the coefficients in the regression (9), excluding full occupancy. We use 1-week or 4-week admission as the outcome and deaths in the previous 2 weeks to construct an instrument for the interaction. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors clustered at the municipality level are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table A4: Effect of Empty Beds on Admissions, with Nonlinear Terms (Un-winsorized Sample)

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.327***	0.309***	0.274***	0.658***	0.642***	0.612***
	(0.0312)	(0.0256)	(0.0198)	(0.0334)	(0.0298)	(0.0260)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.0549***			0.0487***		
	(0.0137)			(0.0153)		
I(Occupancy≥ 90pp)		0.0627***			0.0555***	
		(0.0144)			(0.0170)	
I(Occupancy≥ 85pp)			0.116***			0.103***
			(0.0284)			(0.0318)
Cragg-Donald F-stats	152.8	134.4	75.03	152.8	134.4	75.03
N	6,786,599	6,786,599	6,786,599	6,786,599	6,786,599	6,786,599

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.330***	0.318***	0.299***	0.661***	0.650***	0.633***
	(0.0303)	(0.0264)	(0.0232)	(0.0326)	(0.0298)	(0.0273)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0333***			0.0295***		
	(0.00792)			(0.00885)		
I(Occupancy≥ 50pctile)		0.0303***			0.0269***	
		(0.00696)			(0.00795)	
I(Occupancy≥ 25pctile)			0.0463***			0.0411***
			(0.0106)			(0.0122)
Cragg-Donald F-stats	239.4	290.0	210.3	239.4	290.0	210.3
N	6,786,599	6,786,599	6,786,599	6,786,599	6,786,599	6,786,599

Notes: Table A4 shows the estimates of the coefficients in the regression (9). The estimation sample does not exclude facilities whose maximum occupancy rate falls in the bottom or top 1 percentile of the distribution of maximum occupancy across providers. We use 1-week or 4-week admission as the outcome and deaths in the previous 2 weeks to construct an instrument for the interaction. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors clustered at the municipality level are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Table A5: Effect of Empty Beds on Admissions, with Nonlinear Terms (Maximum Occupancy $\geq 50\%$)

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.306***	0.291***	0.260***	0.614***	0.602***	0.577***
	(0.0266)	(0.0218)	(0.0164)	(0.0246)	(0.0216)	(0.0187)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.0506***			0.0404***		
	(0.0126)			(0.0136)		
I(Occupancy≥ 90pp)		0.0593***			0.0474***	
		(0.0136)			(0.0157)	
I(Occupancy≥ 85pp)			0.107***			0.0856***
			(0.0260)			(0.0284)
Cragg-Donald F-stats	143.5	119.6	69.76	143.5	119.6	69.76
N	5,657,660	5,657,660	5,657,660	5,657,660	5,657,660	5,657,660

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.312***	0.301***	0.283***	0.619***	0.610***	0.595***
	(0.0266)	(0.0234)	(0.0203)	(0.0250)	(0.0229)	(0.0206)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0314***			0.0251***		
	(0.00759)			(0.00839)		
I(Occupancy≥ 50pctile)		0.0289***			0.0231***	
		(0.00678)			(0.00771)	
I(Occupancy≥ 25pctile)			0.0447***			0.0358***
			(0.0105)			(0.0121)
Cragg-Donald F-stats	216.9	256.5	180.2	216.9	256.5	180.2
N	5,657,660	5,657,660	5,657,660	5,657,660	5,657,660	5,657,660

Notes: Table A5 shows the estimates of the coefficients in the regression (9). The estimation sample exclude facilities whose maximum occupancy rate is lower than 50%. We use 1-week or 4-week admission as the outcome and deaths in the previous 2 weeks to construct an instrument for the interaction. The sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors clustered at the municipality level are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1.

B Proof

Proof of Proposition 1

Define a function

$$F(a,d;n)=\left[\begin{array}{c} MB^A(n+a-d)-MC^P(n+a-d)-MC^A(a) \\ MB^D(n+a-d)+MC^P(n+a-d)-MC^D(d) \end{array}\right] \tag{10}$$

where $MB^A(p)=rV'(rp)+b^P+b^A$ and $MB^D(p)=-rV'(rp)-b^P+b^D$. Note $MB^{A\prime}(p)+MB^{D\prime}(p)=0$, which will be used in the algebra below without a mention. The optimal admission and discharge decisions at interior satisfy $F(a^*(n),d^*(n);n)=0$. Also, the Jacobian matrix

$$\begin{aligned} &J_F(a,d;n) \\ &=\left[\begin{array}{cc} MB^{A\prime}(p)-MC^{P\prime}(p)-MC^{A\prime}(a) & -MB^{A\prime}(p)+MC^{P\prime}(p) \\ MB^{D\prime}(p)+MC^{P\prime}(p) & -MB^{D\prime}(p)-MC^{P\prime}(p)-MC^{D\prime}(d) \end{array}\right], \end{aligned}$$

where $p=n+a-d$, is invertible if $MC^{A\prime}(\cdot)$ and $MC^{D\prime}(\cdot)$ are positive and $-V''(\cdot)$ and $MC^{P\prime}(\cdot)$ are non-negative: the determinant of J_F is

$$\begin{aligned} &D_{J_F}(a,d;n) \\ &=\left(-MB^{A\prime}(p)+MC^{P\prime}(p)\right)\left(MC^{A\prime}(a)+MC^{D\prime}(d)\right)+MC^{A\prime}(a)MC^{D\prime}(d)>0. \end{aligned}$$

(i-a) By assumption, J_F is invertible at $(a,d,n)=(\bar{a},\bar{d},\bar{n})$. Then, by the implicit function theorem, we have

$$\begin{aligned} \left[\begin{array}{c} \frac{\partial a^*}{\partial n}\big|_{n=\bar{n}} \\ \frac{\partial d^*}{\partial n}\big|_{n=\bar{n}} \end{array}\right] &= -J_F(\bar{a},\bar{d};\bar{n})^{-1}\left[\begin{array}{c} MB^{A\prime}(\bar{p})-MC^{P\prime}(\bar{p}) \\ MB^{D\prime}(\bar{p})+MC^{P\prime}(\bar{p}) \end{array}\right] \\ &= \frac{-MB^{A\prime}(\bar{p})+MC^{P\prime}(\bar{p})}{D_{J_F}(\bar{a},\bar{d};\bar{n})}\left[\begin{array}{c} -MC^{D\prime}(\bar{d}) \\ MC^{A\prime}(\bar{a}) \end{array}\right]. \end{aligned} \tag{11}$$

Thus, we have $-\frac{\partial a^*}{\partial n}\geq 0$ and $-\frac{\partial d^*}{\partial n}\leq 0$ at $n=\bar{n}$. The inequalities are strict if and only if $-MB^{A\prime}(\bar{p})+MC^{P\prime}(\bar{p})>0$, i.e., if and only if $V''(r\bar{p})<0$ or $C^{P\prime\prime}(\bar{p})>0$.

(i-b) With $MC^{A\prime}(\bar{a}) = \kappa_2^A$ and $MC^{D\prime}(\bar{d}) = \kappa_2^D$, Eq.(11) can be expressed as

$$\begin{bmatrix} \frac{\partial a^*}{\partial n}(\bar{n}) \\ \frac{\partial d^*}{\partial n}(\bar{n}) \end{bmatrix} = \frac{-MB^{A\prime}(\bar{p}) + MC^{P\prime}(\bar{p})}{\left(-MB^{A\prime}(\bar{p}) + MC^{P\prime}(\bar{p})\right)\left(\kappa_2^A + \kappa_2^D\right) + \kappa_2^A \kappa_2^D} \begin{bmatrix} -\kappa_2^D \\ \kappa_2^A \end{bmatrix}.$$

Because we hold $MC^A(\bar{a})$ and $MC^D(\bar{d})$ constant, we have $\frac{\partial \bar{p}}{\partial \kappa_2^A} = \frac{\partial \bar{p}}{\partial \kappa_2^D} = 0$.[47] Therefore, when $V''(r\bar{p}) < 0$ or $C^{P\prime\prime}(\bar{p}) > 0$, we have $\frac{\partial}{\partial \kappa_2^A}\left(-\frac{\partial a^*}{\partial n}(\bar{n})\right) < 0$, $\frac{\partial}{\partial \kappa_2^D}\left(-\frac{\partial a^*}{\partial n}(\bar{n})\right) > 0$, $\frac{\partial}{\partial \kappa_2^A}\left(\frac{\partial d^*}{\partial n}(\bar{n})\right) > 0$, and $\frac{\partial}{\partial \kappa_2^D}\left(\frac{\partial d^*}{\partial n}(\bar{n})\right) < 0$. Otherwise, we have $-\frac{\partial a^*}{\partial n}(\bar{n}) = \frac{\partial d^*}{\partial n}(\bar{n}) = 0$, regardless of the values of κ_2^A and κ_2^D. □

(ii) If $V''(r\bar{p}) < 0$ or $C^{P\prime\prime}(\bar{p}) > 0$, then by Eq.(11),

$$-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) = \frac{1}{1 + \mathcal{E}(\bar{a}, \bar{d}; \bar{n})}$$

where $\mathcal{E}(a, d; n) = \frac{MC^{A\prime}(a)MC^{D\prime}(d)}{(-MB^{A\prime}(p)+MC^{P\prime}(p))(MC^{A\prime}(a)+MC^{D\prime}(d))} > 0$. Thus, $-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) \in (0, 1)$. Otherwise, $-\frac{\partial a^*}{\partial n} - \left(-\frac{\partial d^*}{\partial n}\right) = 0$ by (i-a). □

Proof of Proposition 2

If κ_2^A or κ_2^D is positive (and the other is zero), the conclusion follows from Eq.(11). If both are zero, Eq.(10) implies $n + a^* - d^*$ is constant, yielding the conclusion. □

C Model with Heterogeneity

C.1 Setting

To illustrate how persistent heterogeneity affects the empirical application of our theoretical framework, we consider a simple two-period extension of our baseline model introduced in Section 3. We make the following simplifying assumptions: (1) No discharge occurs. (2) The facility is myopic. (3) $b^P = b^A = 0$.[48]

[47] κ_1^A and κ_1^D need adjusting to hold the marginal costs constant.

[48] Alternatively, we may assume that V is linear and b^P is included in r.

Period 1. The facility chooses initial occupancy n. The utility from initially admitting n patients is $v_1 = V(rn, \eta_1) - C^P(n, \eta_1) - C^A(n, \xi_1, \eta_1) \equiv V_1(rn) - C_1^P(n) - C_1^A(n, \xi_1)$. Utility components are similar to those defined in Section 3, but they now have two types of heterogeneity. First, there are idiosyncratic shocks η_1. Second, the admission cost depends on persistent shocks ξ_1 such that $\frac{\partial MC_1^A}{\partial \xi_1}(a, \xi_1) < 0$ at each a. ξ may be interpreted as supply shocks or a reduced-form expression for demand shocks. For example, higher demand shocks attract more patients, making admission easier. ξ is positively correlated between the two periods: $\mathrm{Cov}(\xi_1, \xi_2) > 0$. In contrast, η is an idiosyncratic shock that is serially uncorrelated and uncorrelated to ξ.

Period 2. Given occupancy n, the facility chooses new admission a to maximize its utility $v_2 = V(r(n+a), \eta_2) - C^P(n+a, \eta_2) - C^A(a, \xi_2, \eta_2) \equiv V_2(r(n+a)) - C_2^P(n+a) - C_2^A(a, \xi_2)$.

We assume that the solution n^* for period 1 satisfies the first-order condition $\frac{\partial v_1}{\partial n} = 0$, and that the solution a^* for period 2 satisfies $\frac{\partial v_2}{\partial a} = 0$.[49] Then, based on a linear approximation of n^* and a^* around a fixed value of shocks $(\xi_1, \eta_1, \xi_2, \eta_2)$, we obtain the following result.

Proposition 4. *The following approximation holds around* $(\xi_1, \eta_1, \xi_2, \eta_2) = (\xi_1^0, \eta_1^0, \xi_2^0, \eta_2^0)$, *where* $n^0 = n^*(\xi_1^0, \eta_1^0)$*:*

$$\frac{\mathrm{Cov}(a^*, n^*)}{\mathrm{Var}(n^*)} \approx \underbrace{\frac{\partial a^*}{\partial n}\left(n^0, \xi_2^0, \eta_2^0\right)}_{\text{admission response}} + \underbrace{\frac{\partial a^*}{\partial \xi_2}\left(n^0, \xi_2^0, \eta_2^0\right)\frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\frac{\mathrm{Cov}(\xi_1, \xi_2)}{\mathrm{Var}(n^*)}}_{\text{endogeneity bias}}. \tag{12}$$

C.2 Proof of Proposition 4

Approximating $n^* = n^*(\xi_1, \eta_1)$ around $(\xi_1, \eta_1) = (\xi_1^0, \eta_1^0)$, we obtain

$$n^* \approx n^*(\xi_1^0, \eta_1^0) + \frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\left(\xi_1 - \xi^0\right) + \frac{\partial n^*}{\partial \eta_1}\left(\xi_1^0, \eta_1^0\right)\left(\eta_1 - \eta_1^0\right). \tag{13}$$

[49] Given these assumptions, it is easy to show that $\frac{\partial n^*}{\partial \xi_1} > 0$ holds if $\frac{\partial MC_1^A}{\partial \xi_1}(\cdot, \xi_1) < 0$, and that $\frac{\partial a^*}{\partial \xi_2} > 0$ holds for fixed (n, ξ_1) if $\frac{\partial MC_2^A}{\partial \xi_2}(\cdot, \xi_2) < 0$.

Similarly, approximating $a^* = a^*\left(n^*\left(\xi_1, \eta_1\right), \xi_2, \eta_2\right)$ around $(\xi_1, \eta_1, \xi_2, \eta_2) = (\xi_1^0, \eta_1^0, \xi_2^0, \eta_2^0)$ and letting $n^0 = n^*\left(\xi_1^0, \eta_1^0\right)$, we obtain

$$
\begin{aligned}
a^* \approx a^*\left(n^0, \xi_2^0, \eta_2^0\right) &+ \frac{\partial a^*}{\partial n}\left(n^0, \xi_2^0, \eta_2^0\right)\left\{\frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\left(\xi_1 - \xi_1^0\right) + \frac{\partial n^*}{\partial \eta_1}\left(\xi_1^0, \eta_1^0\right)\left(\eta_1 - \eta_1^0\right)\right\} \\
&+ \frac{\partial a^*}{\partial \xi_2}\left(n^0, \xi_2^0, \eta_2^0\right)\left(\xi_2 - \xi_2^0\right) + \frac{\partial a^*}{\partial \eta_2}\left(n^0, \xi_2^0, \eta_2^0\right)\left(\eta_2 - \eta_2^0\right).
\end{aligned}
$$

Then, around $(\xi_1, \eta_1, \xi_2, \eta_2) = (\xi_1^0, \eta_1^0, \xi_2^0, \eta_2^0)$, we have

$$
\begin{aligned}
\operatorname{Cov}\left(a^*, n^*\right) \approx & \frac{\partial a^*}{\partial n}\left(n^0, \xi_2^0, \eta_2^0\right) \operatorname{Var}\left(\frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\xi_1 + \frac{\partial n^*}{\partial \eta_1}\left(\xi_1^0, \eta_1^0\right)\eta_1\right) \\
&+ \frac{\partial a^*}{\partial \xi_2}\left(n^0, \xi_2^0, \eta_2^0\right) \operatorname{Cov}\left(\xi_2, \frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\xi_1 + \frac{\partial n^*}{\partial \eta_1}\left(\xi_1^0, \eta_1^0\right)\eta_1\right) \\
&+ \frac{\partial a^*}{\partial \eta_2}\left(n^0, \xi_2^0, \eta_2^0\right) \operatorname{Cov}\left(\eta_2, \frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right)\xi_1 + \frac{\partial n^*}{\partial \eta_1}\left(\xi_1^0, \eta_1^0\right)\eta_1\right) \\
= & \frac{\partial a^*}{\partial n}\left(n^0, \xi_2^0, \eta_2^0\right) \operatorname{Var}\left(n^*\right) + \frac{\partial a^*}{\partial \xi_2}\left(n^0, \xi_2^0, \eta_2^0\right) \frac{\partial n^*}{\partial \xi_1}\left(\xi_1^0, \eta_1^0\right) \operatorname{Cov}\left(\xi_1, \xi_2\right).
\end{aligned}
$$

Thus, we have obtained the desired expression. □

D Insight on Market-level Healthcare Capacity

Under some assumptions, our results can provide insights into policies on market-level healthcare capacity. In Section 7, we outline two illustrative ideas to compare markets in terms of the returns to a capacity expansion, which can then be used to determine which markets should be prioritized for capacity assignment. In what follows, we further illustrate these ideas.

Specifically, we consider hypothetical capacity expansions that reduce occupancy rates by 1pp for all facilities in each market, defined by municipalities, within the prefecture of Tokyo,[50] and discuss which market(s) should be the target of the expansion policy. Capacity expansions for existing facilities leave most demand factors (e.g., distance) unaffected, allowing us to focus on the supply-side behavior. We focus on

[50] Japan consists of 47 prefectures, of which Tokyo is the largest in terms of population and economy. Each prefecture is divided into municipalities, including cities (shi), towns (chō or machi), villages (son or mura), and wards (ku).

a single prefecture because prefectures are the jurisdictions that issue permits for the establishment of nursing facilities and also because facilities will be facing relatively homogeneous patient preferences and input markets within a prefecture, making cost and income effects relatively comparable across markets.

The market-level cost and income effects are computed by summing up the facility-level counterparts. To do so, we first assume that the income effect of a facility at any occupancy level is equal to the admission response at the occupancy level below 85%, and that the remaining part of the admission response is attributed to the cost effect.[51] Because the income effects decrease with occupancy, this amounts to setting the (low-benefit) income effect at its upper bound and the (high-benefit) cost effect at its lower bound, making the benefits of a capacity expansion conservative.

We first consider a capacity targeting rule by only assuming that the net social benefit of admissions due to cost effects is positive and weakly higher than the (positive or negative) net social benefit of admissions due to income effects.[52] Now, define that Market A *dominates* Market B if Market A has a larger admission response and a smaller income effect, than Market B (implying that Market A has a larger cost effect). Compared to Market B, Market A can increase admissions more by more high-benefit admissions and fewer low-benefit admissions. Because the net benefits of cost effects (excluding the costs of the capacity expansion) are assumed to be positive, Market A has a larger total benefit of a capacity expansion. Moreover, because we assume that the per-facility income effect is constant (equal to the admission response of facilities with occupancy below 85%), Market A has a smaller number of facilities than Market B. Thus, if we additionally assume that the cost of the capacity expansion program is increasing in the number of facilities, then Market A has a smaller program cost than Market B. In sum, Market A dominates Market B if Market A offers a larger net benefit of admission responses with a lower cost of a capacity expansion. We can then consider a *non-dominance criterion*, which requires that policymakers select markets that are not dominated by any other market. The non-dominance criterion does not assume specific values of the net benefits of the two

[51] The total admission responses are those in Figure 5a. The estimates for 95%+ are used if the occupancy rate is equal to or greater than 95%.

[52] See footnote 42 in the main text for a discussion of this assumption. Also, at this point, we do not make any assumptions on the cost of capacity expansion.

effects.

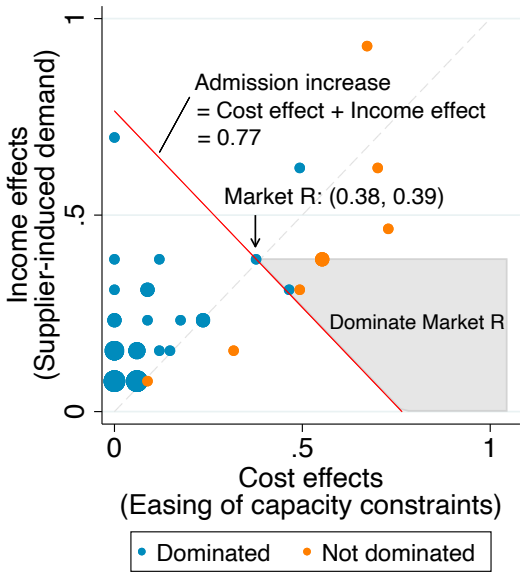
Figure A7a graphically illustrates the non-dominance criterion using the scatter plots of market-level cost and income effects in Tokyo. Now, consider Market R in Figure A7a. The red line is the "iso-quant" line of the total admission increases in response to a 1pp capacity expansion at each existing facility, corresponding to the same total admission increases as Market R. Then, Market R is dominated by markets in the gray area, so it is not selected by the non-dominance criterion. Using the same criterion, markets shown in orange are not dominated by any other markets, making them favorable candidates for a capacity expansion.

Next, we can further narrow down the priority markets for capacity expansion, if we assume the relative social values of cost and income effects. Suppose that the policymaker attaches a positive value to cost effects and a negative value to income effects. Then, she is willing to tolerate an increase in income effects to have a 1pp increase in cost effects. Thus, her preference is expressed by a positively sloped indifference curve. If we assume that the trade-off rate between the two effects is constant, then the indifference curve is represented by a straight line, and the objective of the policymaker becomes to select the market that maximizes a weighted sum of cost effects and income effects. We call this rule of market selection a *weighted sum criterion*.

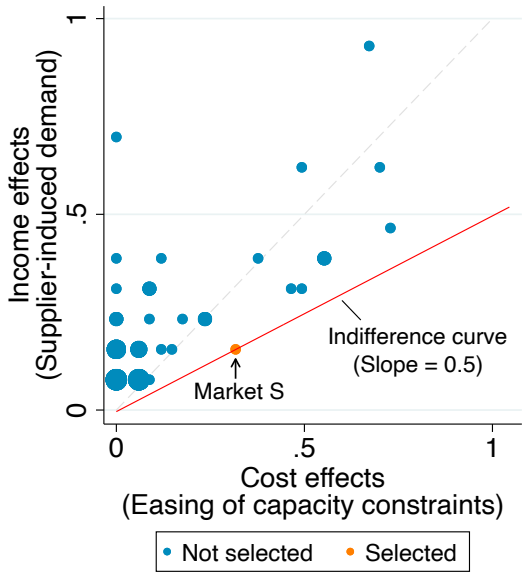
Figure A7b shows the same scatter plots as Figure A7a and illustrates the situation where the policymaker tolerates a 0.5pp increase in admission due to income effects to achieve a 1pp increase in admission due to cost effects. Market S in Figure A7b maximizes the weighted sum of cost and income effects ($u = 0.5 * \text{cost effects} - \text{income effects}$).

Figure A7: Visual Illustration of Capacity Assignment Rules

(a) Non-Dominance Criterion



(b) Weighted Sum Criterion



Notes: Figures A7a and A7b illustrate capacity assignment rules based on the "non-dominance criterion" and "weighted sum criterion" outlined in the text, respectively. Both figures show scatter plots of market-level cost and income effects in Tokyo Prefecture using 1-week admission responses to a 1pp decrease in occupancy. The size of the plot reflects the number of markets with the same cost and income effects. In Figure A7a, the negative 45-degree line represents the combinations of cost and income effects with the same level of total admission responses as Market R, which is equal to 0.77. Markets in the gray area are those that dominate Market R (i.e., they have a larger admission response and a smaller income effect than Market R). Figure A7b illustrates the situation where the policymaker tolerates a 0.5pp increase in admission due to income effects to achieve a 1pp increase in admission due to cost effects. The policymaker's objective (represented by $u = 0.5 * \text{cost effects} - \text{income effects}$) is maximized at Market S.