Capacity Constraints and Inefficient Service Delivery: Theory and Evidence from Nursing Facilities*

Hiroki Saruya† Masaki Takahashi‡

December 10, 2024

Abstract

This paper studies frictions and inefficiencies in healthcare delivery under capacity constraints. We develop an economic model where a healthcare facility's admission/discharge decisions depend on bed occupancy through capacity constraints and demand inducements. The model generates behavioral and efficiency implications: (1) Capacity constraints imply that admissions/discharges respond to occupancy fluctuations more intensely at higher baseline occupancy, whereas demand inducements imply that the responses are more intense at lower baseline occupancy, so the relative importance of the mechanisms is testable. (2) If capacity constraints are more important, then smoothing occupancy across homogeneous facilities can increase aggregate service provision. Applying the framework to Japanese nursing facilities, with patient deaths as occupancy shocks, we find that admission responses to occupancy fluctuations are mainly driven by capacity constraints. Our simulation shows that smoothing occupancy across facilities can increase aggregate admissions without expanding capacity, suggesting inefficient access to facilities in the status quo.

Keywords: capacity constraint, misallocation, 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 EuHEA Conference 2024, JHEA Conference 2023, 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, Cabinet Office, Government of Japan. Email: hiroki.saruya@aya.yale.edu

‡Sophia University. Email: masaki-takahashi@sophia.ac.jp

1 Introduction

The efficient delivery of products and services under capacity constraints is a central issue in the economy. In many markets such as health care, housing, and transportation, governments regulate the supply capacity for a variety of reasons, including preventing unnecessary services and public spending, controlling externalities, and ensuring comfort and safety. However, such regulations tighten the capacity constraints faced by suppliers and can delay or prevent valuable market transactions, thereby reducing efficiency. Improving market efficiency while imposing capacity constraints is a common challenge faced by governments in a wide range of public policies.

Efficiency in the provision of capacity-constrained services depends on the allocation of capacity utilization across providers, as well as on the overall capacity. If one provider faces a high congestion cost (or a binding capacity constraint) while another has spare capacity, efficiency can be improved by shifting some transactions from the former to the latter without changing overall capacity. Such inefficient allocations may arise and persist if providers need time to adjust their service provision to demand shocks. To consider an efficiency-improving policy, we need to know how capacity affects heterogeneous providers: in the above example, we need to verify that a (marginal) increase in capacity does increase the provision of valuable services more intensely for the more congested provider than for less congested one, in order to argue for the (marginal) reallocation of service delivery.

We study frictions and inefficiencies in healthcare access under capacity constraints. 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: healthcare productivity may be reduced when providers are congested. The existing literature is limited in two ways. First, the positive and negative aspects of capacity constraints are analyzed separately, and their trade-offs are not examined in a unified manner. Second, they typically focus on the behavior of individual suppliers, and do not discuss the efficiency implications at the market level. We fill these gaps and explore how to efficiently manage capacity utilization 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 an economic model that shows how a facility's admission and discharge decisions depend on bed occupancy. In the model, the marginal cost of serving an additional patient[2] increases with occupancy, possibly reflecting the altruistic facility's concern about quality deterioration, as well as monetary costs such as diminishing returns and higher input costs. The model also allows for (loose) income targeting or occupancy targeting, which incentivizes the facility to induce demand in response to a reduction in occupancy. 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 a frictionless case, the facility adjusts the net admissions to exactly offset the occupancy reduction. In contrast, in a 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 allows us to empirically assess the relative importance of the

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

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

mechanisms. 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 valuable care provision.

We then discuss the efficiency implications of occupancy variation across facilities. We consider two homogeneous facilities that make admission decisions while facing different occupancy rates due to idiosyncratic shocks, such as emergency admissions, patient deaths, spatial and information frictions, and staff shortages. The model predicts that aggregate admissions will increase if patients at the more occupied facility is moved to the less occupied one. Thus, an occupancy-smoothing policy is useful to achieve efficient provision of care.

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 quality or operational efficiencies that also affect admissions. We address this problem by exploiting patient deaths as exogenous occupancy shocks.[3] The identification assumptions are that the exact timing of patient deaths (but not necessarily the longer-run volume of deaths) 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. These assumptions are plausible in our setting because (i) nursing facilities do not have advanced technologies to manipulate the timing of patient deaths, (ii) applicants or in-facility patients are unlikely

[3]People's deaths have been exploited as an exogenous variation in several settings, such as fatal shocks to family members (Fadlon and Nielsen, 2021), worker or executive exit from a firm (Jäger and Heining, 2019; Sauvagnat and Schivardi, 2023), entrepreneur exit from a start-up company (Becker and Hvide, 2022), changes in collaborative networks of inventors (Jaravel et al., 2018), changes in national leaders (Jones and Olken, 2005), and changes in academic co-authorship (Azoulay et al., 2010).

to respond quickly to daily patient deaths, and (iii) the extra work for facilities due to deaths seems irrelevant. We also show the absence of a pre-trend in admissions or discharges in an event study design.

We empirically test the above theoretical predictions in the context of Japanese nursing facilities for rehabilitation and transitional care, similar to skilled nursing facilities (SNFs) in the US. 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. The high demand for institutional care is reflected in high occupancy rates, which creates a congestion problem because staffing cannot be adjusted flexibly. 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 facility-by-date panel data on bed occupancy and the number of admissions, discharges, and deaths, we first implement an event study design to examine the effect of death-induced occupancy declines on admissions and discharges, investigating detailed dynamic effects. We then estimate regressions of weekly/monthly/quarterly admissions and discharges on daily occupancy, using patient deaths as an instrumental variable (IV).

We find that patient deaths immediately increase subsequent admissions, whereas they have a much weaker effect on live discharges. Admissions increase as early as the day after patient deaths, and the increase persists for over 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 76.7% of the 1-week admission response at baseline occupancy strictly between 95% and 100% is explained by the cost effect (0.23pp out of 0.30pp). 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 the mechanical effect of binding capacity constraints.

Our estimates suggest that aggregate admission can be increased by reallocating patients to smooth occupancy across facilities, a policy tool potentially useful in markets where capacity investment is regulated or otherwise difficult to adjust in the short run. Our most conservative estimate implies that marginally smoothing occupancy by moving a patient from the most occupied facility to the least occupied facility within each city-fiscal year-facility size bin leads to a 8.1% increase in the total 4-week admissions of the treated facilities. To the extent that the dispersion in occupancy is not explained by facility heterogeneity, the dispersion is indicative of spatial misallocation of patients.

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 (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 which emphasize financial incentives,[4] we show that capacity constraints can be a key mechanism in our context. We also show that the OLS estimate of the regression of admissions on occupancy is biased upward relative to the IV estimate, suggesting the importance of accounting for endogeneity. Finally, unlike most previous studies, we discuss the market (in)efficiency of the variation in occupancy across providers.

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;

[4]See, e.g., Evans (1974), Gruber and Owings (1996), Freedman (2016) and Ikegami et al. (2021).

Buchmueller et al., 2015; Gandhi, 2023; Alexander and Schnell, 2024; Cabral et al., 2024). Previous studies have focused on increasing provider payments (Alexander and Schnell, 2024) or expanding capacity (Gandhi, 2023) as tools to improve access. Our contribution is to document heterogeneous responses across differentially constrained facilities and to show that smoothing occupancy can improve overall access to facilities. Given the high costs of paying providers or expanding capacity, coordinating capacity utilization could be a useful policy tool to improve access to care.

Finally, this study contributes to the broad literature on how demand fluctuations affect market efficiency in the presence of capacity constraints (Baker et al., 2004; Collard-Wexler, 2013; Butters, 2020; Shurtz et al., 2022; Boehm and Pandalai-Nayar, 2022; Ilzetzki, 2024), or adjustment or matching frictions (see Gavazza and Lizzeri, 2021, for a review). 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. Researchers have studied how aggregate production in the manufacturing sector is affected by micro-level capacity constraints (Boehm and Pandalai-Nayar, 2022) or (factor) misallocation (Hsieh and Klenow, 2009). Another strand of research (e.g., Fréchette et al., 2019) has shown that short-run demand fluctuations can have important efficiency consequences in frictional markets. Our work extends these lines of research to the healthcare sector.[5]

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 implications of our results. Section 8 concludes.

[5]Our framework can potentially be modifiable to apply to other industries with partially altruistic providers, such as hospitals, childcare facilities, and schools.

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 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 certain level. 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 12.7 trillion JPY in 2021 (about 127 billion USD at the then exchange rate, 2.3% of Japan's GDP). The cost of institutional care, including nursing facilities, accounts for half of the total cost.

We focus on a type of nursing facilities called Geriatric Health Services Facilities (GHSFs).[6] 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.[7] 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 and fulfill their public purpose, 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 2019, there were 4,337 GHSFs nationwide, with approximately 360,000 patients admitted.

GHSFs provide care for two types of patients. "Long-stay" patients are admitted to the facility for rehabilitative care to return to the community. "Short-stay" patients, on the other hand, visit GHSFs to receive temporary assisted living services, typically for a respite or temporary unavailability of family caregivers. Stay types are identified by claims items rather than by length of stay. In our empirical

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

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

analysis, we primarily focus on the admissions and live discharges of the long-stay patients, because short-stay admissions and discharges are more likely to be influenced by exogenous factors.[8]

A variety of healthcare professionals work in GHSFs to provide appropriate care, including physicians, nurses, caregivers, 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.[9]

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

[8]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.

[9]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.

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 or, depending on their medical condition, to another GHSF.

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). Table S1 in Supplemental Appendix S.A describes the general health status for each care level. Only recipients 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. Table S2 in Supplemental Appendix S.A shows per-diem fixed and FFS payments by care levels, using our analysis sample described in Section 4.[10] 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

[10]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 Table S2 for more details.

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) Service delivery is negatively affected by congestion, partly due to labor shortages.

3 Conceptual Framework

3.1 Model Setup

To guide our empirical analysis, we present an economic model of admissions and discharges.[11] A representative facility chooses the number of new patients to admit, a, and the number of existing patients to discharge, d, to maximize its objective function given the number of patients currently in the facility, n. We fix capacity and express these variables as their ratio to capacity (e.g., n denotes occupancy rate).[12] The payoff of the facility is given by

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

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

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 $V'(\cdot) > 0$ and $V''(\cdot) \leq 0$. We also assume $V'''(\cdot) \geq 0$, which is not required for the propositions below but facilitates the interpretation of the results.[13] We allow V to express loose income targeting,

[11] Suitably modified versions of our model will be applicable to hospitals or other industries as well. For example, not-for-profit childcare facilities or schools may determine the number of admitting (and perhaps graduating) children/students by considering a profit or occupancy target, capacity constraint, and altruistic utility from service. Such decisions can possibly be expressed by a model similar to the one below.

[12] The parameters in Eq.(1) must be rescaled accordingly.

[13] $V''' \geq 0$ implies that the income effect (defined below) shrinks with occupancy. It is satisfied by most of the common candidates for V, e.g., a CARA utility function, a CRRA utility function, and $V(R) = x^k$ for $k \in (0,1)$. A nonnegative third derivative of utility function is also a common assumption in macroeconomic consumption models, to derive the concavity of the consumption

including an approximate non-negativity constraint on profit.[14] A large 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 set a target occupancy level, and the incentive to induce demand may increase if the occupancy rate falls below the target level.

Following the literature on non-profit organizations (Lakdawalla and Philipson, 1998; Gaynor and Vogt, 2003), we assume that the facility's payoff 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 strictly convex "congestion cost" that reduces per-patient utility as the number of patients increases. 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. Also, C^P may approximate the capacity constraint, since admissions in excess of capacity can severely degrade the quality of the patient experience. Finally, with an appropriate reinterpretation of r, C^P may capture the monetary cost of service that increases nonlinearly with volume.[15] 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 perfor-

function (e.g., Carroll and Kimball, 1996).

[14] E.g., $V(R) = v(R - FC)$ for some concave v and a fixed cost FC, where $v'(R)$ is high at $R < 0$ and low at $R > 0$.

[15] 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.

mance 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 monetary cost.[16]

The model abstracts from two features. First, it omits the facility's choice of care quality to influence patient health. Instead, the facility is concerned with congestion-dependent service quality (e.g., patient satisfaction), and it adjusts admissions and discharges directly, taking into account their effects on service quality. In other words, discharge in our model represents a short-run tool for managing congestion, rather than a long-run product of care. Second, the model omits patient heterogeneity, so that it isolates the role of occupancy from, e.g., selection incentives. To the extent that heterogeneity is controlled for by observables (including a proxy for care needs), our empirical results in Section 6 can be linked to the theoretical predictions below.

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 Theoretical Prediction

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.[17] 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 simplify by assuming 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

[16] Admission cost may reflect the cost of assessing patients' needs, coordinating the admission process, and moving patients to the facility. Discharge cost may consist of similar factors. These costs are likely to increase more rapidly as the volume increases, for example, due to higher labor costs of workers in charge of discharges.

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

assumption is imposed on MC^P or MB^A, except that MB^A is weakly decreasing and MC^P is strictly increasing.[18]

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 $n = \bar{n}$.*

(i) (Responses to exogenous discharges)

(a) $-\frac{\partial a^*}{\partial n} > 0$ and $-\frac{\partial d^*}{\partial n} < 0$.

(b) Holding $MC^A(\bar{a})$ and $MC^D(\bar{d})$ constant, $\left|\frac{\partial a^*}{\partial n}\right|$ decreases in κ_2^A and increases in κ_2^D, and $\left|\frac{\partial d^*}{\partial n}\right|$ decreases in κ_2^D and increases in κ_2^A.

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

(iii) (Covariation with occupancy) $-\frac{\partial^2 a^*}{\partial n^2} > (<)\ 0$ and $\frac{\partial^2 d^*}{\partial n^2} > (<)\ 0$ hold if $MC^{P\prime\prime}(\bar{p}) > (<)\ MB^{A\prime\prime}(\bar{p})$.

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 $-\frac{\partial d^*}{\partial n} = 0$.

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

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

where $MB^{A\prime}(\bar{p}) = r^2 V''(r\bar{p}) \leq 0$ and $D_{J_F}(\bar{a},\bar{d};\bar{n})$ 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 in order to compensate for the lost income. The second term represents *a cost effect*, whereby a decrease

[18] The conclusions of Propositions 1 and 2 hold under much weaker conditions: they hold if $-V''(r\bar{p}), C^{P\prime\prime}(\bar{p}) \geq 0$ and at least one is positive. If both are zero (e.g., V and C^P are linear), the optimal admissions and discharges are invariant to occupancy around $n = \bar{n}$.

in occupancy reduces the marginal cost of service and allows the facility to admit more patients. Proposition 1-(i) also shows that, holding marginal costs constant (hence holding a^* and d^* constant), the response of admission (discharge) to occupancy shocks is larger when the adjustment of admission is less (more) "frictional" than the adjustment of discharge, in the sense that the marginal cost of admission (discharge) does not increase fast or the marginal cost of discharge (admission) rises rapidly.[19] This is because the facility attempts to adjust occupancy in the less costly way.

Next, 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.

Crucially, Proposition 1-(iii) shows how the magnitude of the admission and discharge responses to occupancy shocks varies with occupancy. Eq.(3) shows that $MC^{P\prime\prime} > 0$ governs the variation in the cost effect and $MB^{A\prime\prime} \geq 0$ governs the variation in the income effect. The cost effect increases as occupancy increases, making the magnitude of the admission response *larger* at higher occupancy rates. In contrast, the income effect decreases as occupancy increases (recall $MB^{A\prime} \leq 0$), making the magnitude of the response *smaller* at higher occupancy rates. Intuitively, with a negative occupancy shock, the incentive to admit more patients due to reduced capacity constraints is greater when the facility is more occupied, while the incentive to admit extra patients for higher income is greater when it is less occupied. If the former mechanism dominates the latter, then the magnitude of the response increases with occupancy ($\frac{\partial}{\partial n}\left|\frac{\partial a^*}{\partial n}\right| > 0$), and it decreases with occupancy otherwise.[20]

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

[20] With general MC^A and MC^D, the signs of $\frac{\partial}{\partial n}\left|\frac{\partial a^*}{\partial n}\right|$ and $\frac{\partial}{\partial n}\left|\frac{\partial d^*}{\partial n}\right|$ are not solely determined by $MC^{P\prime\prime}$ and $MB^{A\prime\prime}$. Even in such cases, the size of admission and discharge responses increases (decreases) with n if $MC^{P\prime\prime}$ is sufficiently larger (smaller) than $MB^{A\prime\prime}$.

Although Proposition 1-(iii) concerns whether the *variation* in the cost effect or the income effect mainly explains the variation in the admission/discharge responses, the results can be used to bound the *levels* of the responses attributable to each effect, at a given n. To illustrate, we first note that D_{J_F} in Eq.(3) increases with n, and thus the level of the income effect decreases with n, if $-\frac{\partial^2 a^*}{\partial n^2} > 0$ and $\frac{\partial^2 d^*}{\partial n^2} > 0$.[21] Now, suppose, e.g., that the admission response to a 1pp decrease in occupancy is 0.4pp at $n = 0.5$ and 0.7pp at $n = 0.9$. If we assume that the income effect explains 0.2pp (50%) of the response at $n = 0.5$, then the income effect at $n = 0.9$ is in [0pp, 0.2pp], so the cost effect at $n = 0.9$ is in [0.5pp, 0.7pp]. Without such an assumption on the income effect at $n = 0.5$, we still know that its upper bound is the entire response, 0.4pp, so the cost effect (income effect) at $n = 0.9$ is at least 0.3pp (at most 0.4pp). This approach is useful when, e.g., we discuss the value of the admission responses by examining the fraction of induced demand in them.

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

3.3 Optimality of Occupancy Smoothing

Now, consider a situation where there are multiple facilities, as in the real world. If the cost effect is a more important driver of admission responses than the income effect, then Proposition 1-(iii) implies that smoothing occupancy between homogeneous facilities will increase total admissions. This is because the admission function is concave in occupancy. In Supplemental Appendix S.B, we formalize this intuition by extending the above model to a two-facility setting. We assume that two homogeneous facilities 1 and 2, with possibly different occupancy rates n_j ($j = 1, 2$), make an admission decision a_j, and we examine how total admissions depend on the distribution of occupancy. Variation in n_j can arise from idiosyncratic payoff and occupancy

[21] See Appendix A. Without this property, we know that the *fraction* of the income effect among the response decreases with n, which is useful for bounding exercises such as below.

shocks, such as emergency admissions, deaths, spatial or information frictions, and staffing shortages.[22]

We suppose that the government can reallocate $N = n_1 + n_2$ patients to facility 1 or 2 before the facilities make an admission decision. The government is concerned with total access to care, or the quantity of service production, given by $A(n_1; N) = n_1 + a_1^*(n_1) + n_2 + a_2^*(n_2) = N + a_1^*(n_1) + a_2^*(N - n_1)$, where $a_j^*(n_j)$ is facility j's optimal admission given occupancy n_j.

Proposition 3 in Supplemental Appendix S.B shows that, given the number of initial in-facility patients N, an occupancy-smoothing policy that moves patients from the more congested facility to the less congested facility increases aggregate access to care. Thus, aggregate access is maximized by setting $n_1 = n_2 = N/2$.

Note that this result describes a static or short-run effect. In a dynamic setting, occupancy may converge to a steady-state level even in the absence of reallocation. In such a case, reallocation can still improve short-run admissions as long as the direct effect on current admissions dominates the indirect effect on future admissions due to changes in future occupancy rates.

A caveat to using the above framework in data analysis is that it assumes away persistent heterogeneity across facilities. Even if higher occupancy leads to fewer admissions within the same facility, higher-occupancy facilities may admit more patients than lower-occupancy facilities due to unobserved heterogeneity in quality or operational efficiency. In what follows, we combine the above framework with empirical approaches to account for heterogeneity.

[22] If a facility cannot completely refuse emergency admissions at will after making the initial admission decision, this will shift its occupancy upward. If a facility faces increased demand because of a temporary improvement in transportation access or awareness, it may shift the admission cost downward. A temporary staffing shortage may shift both the service cost and the admission cost upward.

4 Data

4.1 Data Sources and Sample Selection

The primary data source is LTCI claims data. The sample period is from April 2011 to March 2018.[23] 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 patients, and the discharge date and destination for those who are discharged in our sample period. If a patient dies in a facility, the discharge destination is recorded as death. We also use the Survey of Long-Term Care Service Facilities to obtain annual information on the characteristics of each GHSF, such as the number of beds. 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.

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 1 shows summary statistics for our main sample at facility-date level. The average number of beds is 83, with most facilities having a capacity between 50 and 100. On average, a facility employs 0.72 full-time equivalent physicians and 8.96 full-

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

[24] Our main results are unchanged when we remove the sample restrictions based on the maximum occupancy rate. See Figure S4 and Table S4 in Supplemental Appendix S.A.

Table 1: Summary Statistics

	Mean (1)	SD (2)	p10 (3)	p90 (4)
Facility-date (Obs. = 6,759,468, #Facilities = 3,087)				
Capacity	83.44	30.45	48	100
Physician (Full-time equivalent)	0.72	0.56	0	1
Nurse (Full-time equivalent)	8.96	3.76	5	13
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	15.69	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

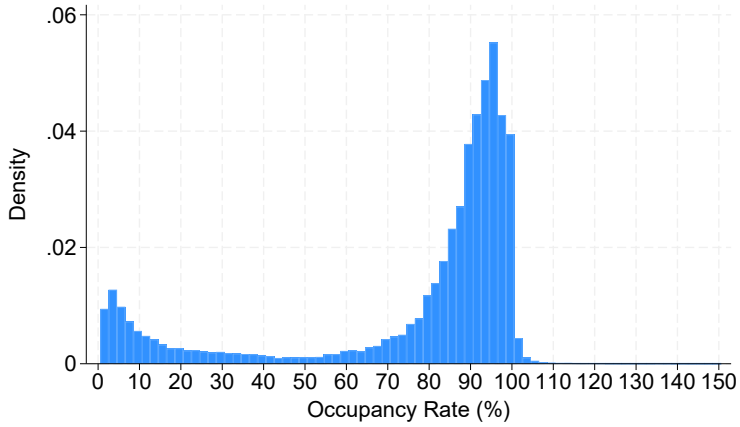
Notes: The table presents summary statistics for the facility-date panel. The last two columns present 10th and 90th percentiles. The occupancy rate (overall and by care levels) is the number of patients in the category divided by the capacity, expressed in pp. The other variables are expressed in level.

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. Occu-

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

Figure 1: Distribution of Occupancy Rates



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

pancy is mostly concentrated in the range of 80-100%. A small number of observations have occupancy rates higher than one, possibly due to the temporary use of a makeshift bed, although the reimbursement rate is reduced if such excess utilization persists for a period of time.

Figure 2 shows the binscatter of the patient-to-nurse ratio against occupancy rate, using facility-fiscal year observations.[26] The average number of patients per nurse is just over 8 at 80% occupancy, and it rises to about 9.5 as occupancy approaches 100%. This suggests that nurse staffing does not adjust to maintain the patient-to-nurse ratio, implying less nurse time per patient at higher occupancy rates.

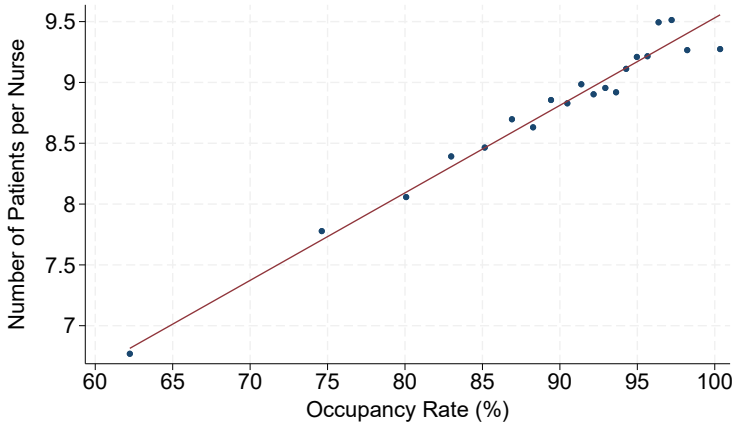
5 Empirical Strategy

5.1 Effect of Occupancy on Admissions and Discharges

A naive way to test the predictions of Propositions 1 and 2 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 quality or operational efficiencies that also affect admissions/discharges. We address this problem by exploiting patient deaths. When a patient dies in a facility, she is

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

Figure 2: Occupancy and Patient-to-Nurse Ratio



Notes: Figure 2 shows the binscatter of the patient-to-nurse ratio versus occupancy rate, using facility-fiscal year observations with at least 50% occupancy. The number of nurses is calculated by counting full-time equivalent nurses.

discharged from the facility, which reduces occupancy. 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:

$$Y_{jt} = \sum_{k=-84}^{84} \beta_k Deaths_{jt-k} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{4}$$

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.[27] We estimate β_k from 84 days (12 weeks) before and after patient deaths. The regression model includes fiscal year-specific facility fixed effects γ_{jy} 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 (4) 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 (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.(4). 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:

$$Y_{jw(t)} = \beta OC_{jt} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{5}$$

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.[28] 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

[27] 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).

[28] 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).

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 (Propositions 1-(i)(ii) 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

$$OC_{jt} = \alpha Deaths_{jt-1} + \tilde{\lambda}' X_{jt} + \tilde{\gamma}_{jy} + \tilde{\gamma}_t + \tilde{\varepsilon}_{jt}, \tag{6}$$

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

5.1.3 Effects by Baseline Occupancy

Proposition 1-(iii) 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. (5) 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:

$$Y_{jw(t)} = \beta_1 OC_{jt} + \beta_2 OC_{jt} \times I\left\{OC_{jt} \geq L^o\right\} + \lambda' X_{jt} + \gamma_{jy} + \gamma_t + \varepsilon_{jt}, \tag{7}$$

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 $t-1$.[29] 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 $t-1$.

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.[30]

5.2 Identification Concerns

The key identification assumption for Eq. (4) and (5) 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. In addition, patients are unlikely to respond quickly to short-run fluctuations in occupancy, in part because information on patient deaths is difficult to collect, while longer-term trends in occupancy are captured by provider-by-fiscal year fixed effects. The exclusion of short-run occupancy from patient preferences is especially plausible for admissions, because of the time lag between application and admission, which can be more than a month.

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 con-

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

[30] 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.

cerns 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.

Another concern in estimating Eq.(5) is that patient deaths may directly affect admissions and discharges, violating the exclusion restriction. For example, the deaths may cause live patients to avoid admission or to seek sooner discharge because they signal the poor quality of the facility. Such effects are likely to be negligible for admissions, because it is difficult for applicants to gather information about the daily deaths of in-facility patients. Patients are also unlikely to hasten discharge in response to death events, especially in the short run, because discharge requires advance planning. It is also unlikely that patients update their beliefs about facility quality based on daily (not long-run aggregate) deaths residualized by detailed covariates and fixed effects. Alternatively, deaths may create extra work, inducing facilities to reduce workload by deferring new admissions or expediting discharges. This is unlikely because the outcomes are admissions/discharges *after* dead patients' discharges and associated work have already been completed. In Section 6.1.1, we show that $\hat{\beta}_0$ in Eq. (4), the effect of patient deaths on admissions/discharges on the same day, is almost the same as the overall pre-trend. This result indicates that the death-related extra work has little impact on admissions/discharges. Moreover, the direct effects via patient preference or extra work, if any, would make our estimates conservative.

As described in Section 5.1.3, when we estimate Eq. (5) by occupancy levels, still 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: (i) estimate an alternative regression Eq. (7), which uses patient deaths in the past two weeks as a source of exogenous variation in baseline occupancy,[31] and (ii) estimate Eq. (5) and (7) by facility-specific occupancy percentiles rather than

[31]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.

levels, to exploit within-facility variation in assignment to high vs. low occupancy.

6 Results

6.1 Effects of Occupancy on Admissions and Discharges

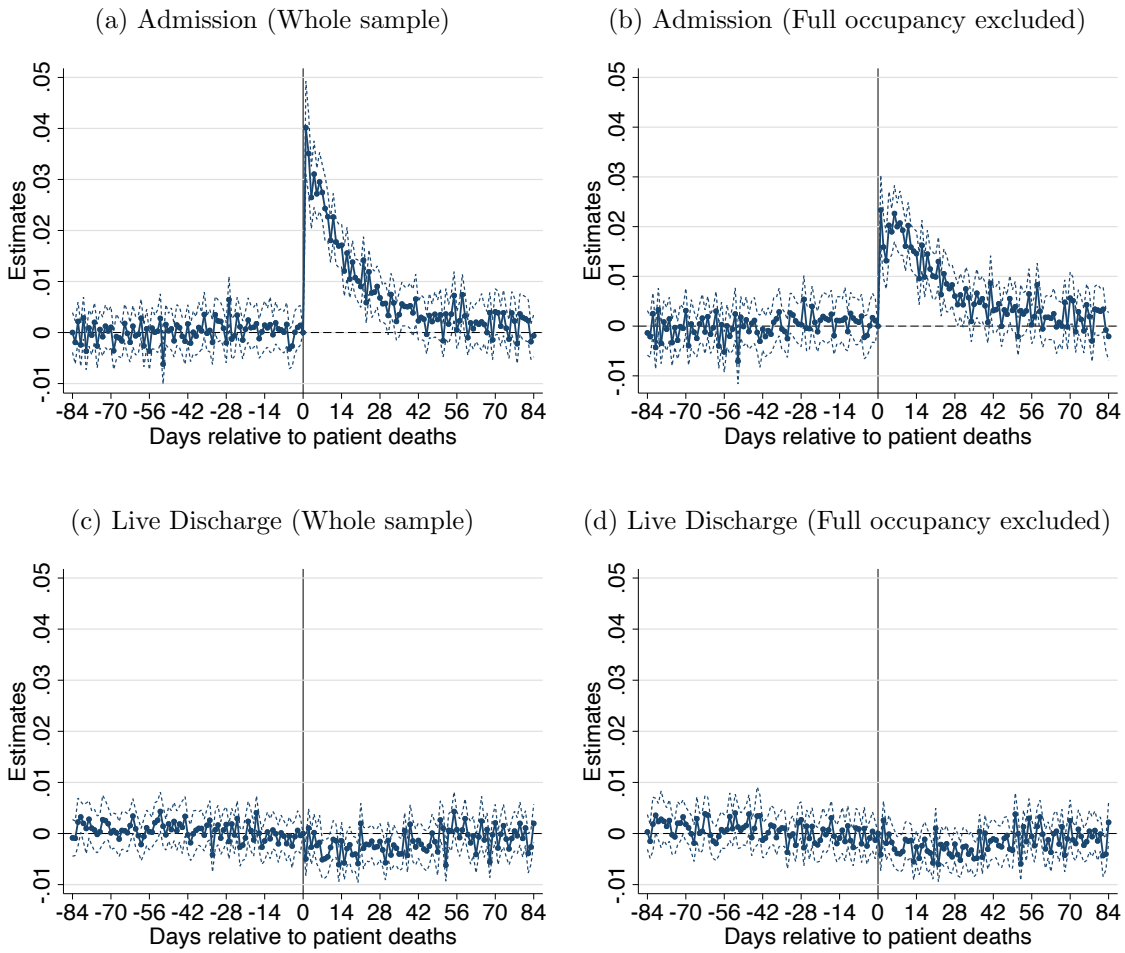
6.1.1 Event Study Results

Figure 3 plots the estimates of β_k in the event study regression (4), which represents the effects of patient deaths on the number of admissions or live discharges measured as a pp change in occupancy. The estimate for the day of patient deaths is normalized to zero ($\hat{\beta}_0 = 0$). 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. 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 decreases 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. It 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 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,

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



Notes: Figure 3 plots estimates of β_k coefficients from Eq. (4), 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.

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.

6.1.2 IV and OLS Estimates

Figure 4 plots the IV estimates of β in Eq. (5), 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) 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.[32]

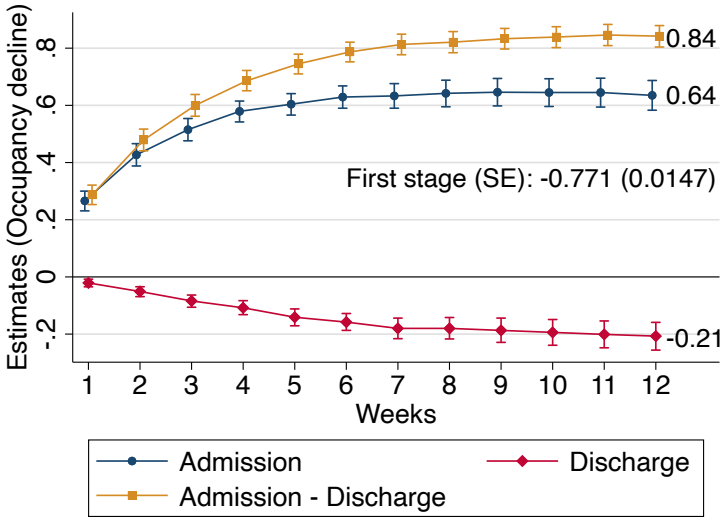
Figure 5 plots the OLS and IV estimates of the coefficients on occupancy in the regression Eq. (5). Although both (reversed) estimates are positive for admissions and negative for discharges, the IV estimates are larger than the OLS estimates, especially for admissions. This suggests that the OLS under-estimates the effect of an occupancy reduction, probably because congested facilities tend to admit more patients. This attenuation bias is reminiscent of the attenuation bias of the OLSE in the regression of quantity on price.

6.2 Heterogeneous Effects by Occupancy Levels

We now estimate Eq. (5) separately by occupancy rates on day $t-1$. Figure 6a 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

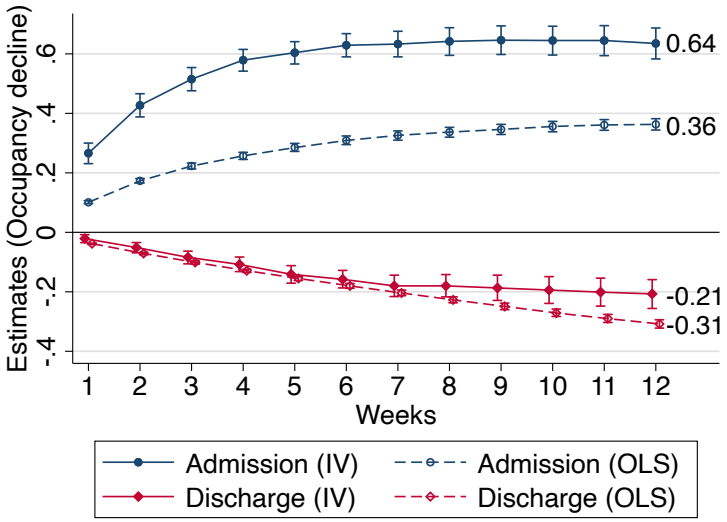
[32]The magnitude of responses is not monotonically increasing in weeks, because the admissions and discharges in the "control group" change over time as well.

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



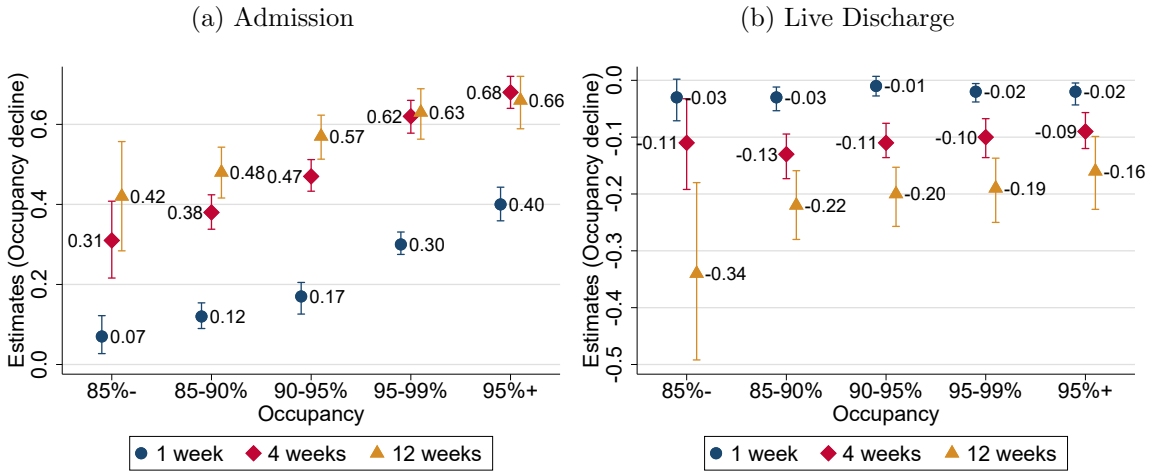
Notes: Figure 4 plots the estimates of the coefficient on occupancy in Eq. (5) and their 95% confidence intervals, using as outcomes long-stay admissions (blue curve), live discharges (red), and admissions minus live discharges (orange) for the following T week(s), $T = 1, ..., 12$. The sign is reversed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

Figure 5: OLS vs IV Estimates of the Coefficient on Occupancy



Notes: Figure 5 plots OLS (dotted curves) and IV (bold) estimates of the coefficient on occupancy in Eq. (5) 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 6: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Level



Notes: Figure 6 plots the IV estimates of the coefficient on occupancy in Eq. (5) 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 sign is reversed to represent the pp effect of a 1pp reduction in occupancy. Standard errors are clustered at the municipality level.

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. In contrast, Figure 6b 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 Figure S1).

Note that the differences in admission responses between occupancy levels shrink for longer-run outcomes, especially 12 weeks. This is possibly because admissions return to the steady-state level in the long run (recall the discussion in Section 3.3). 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 2 presents the estimates from Eq. (7). 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

Table 2: Effect of Empty Beds on Admissions, with Nonlinear Terms

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.305***	0.289***	0.259***	0.612***	0.599***	0.574***
	(0.0263)	(0.0215)	(0.0162)	(0.0243)	(0.0213)	(0.0185)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.0507***			0.0415***		
	(0.0125)			(0.0134)		
I(Occupancy≥ 90pp)		0.0592***			0.0484***	
		(0.0134)			(0.0154)	
I(Occupancy≥ 85pp)			0.106***			0.0868***
			(0.0256)			(0.0277)
Cragg-Donald F-stats	169.1	142.2	83.88	169.1	142.2	83.88
N	6,673,739	6,673,739	6,673,739	6,673,739	6,673,739	6,673,739

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.311***	0.300***	0.282***	0.617***	0.608***	0.593***
	(0.0264)	(0.0232)	(0.0202)	(0.0246)	(0.0226)	(0.0204)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0313***			0.0256***		
	(0.00752)			(0.00824)		
I(Occupancy≥ 50pctile)		0.0288***			0.0236***	
		(0.00672)			(0.00757)	
I(Occupancy≥ 25pctile)			0.0445***			0.0364***
			(0.0104)			(0.0118)
Cragg-Donald F-stats	256.5	303.5	214.5	256.5	303.5	214.5
N	6,673,739	6,673,739	6,673,739	6,673,739	6,673,739	6,673,739

Notes: The table shows the estimates of the coefficients in the regression (7), 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.

to an occupancy reduction are larger when the baseline occupancy is higher. The same pattern holds when we exclude observations with binding capacity constraints, although some interactions are statistically insignificant (see Panel A of Table S3 in Supplemental Appendix S.A). Again, the results are qualitatively similar when we

use occupancy groups based on percentiles instead of levels (see Panel B of Tables 2/S3).

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 Figure S4 and Table S4.[33]

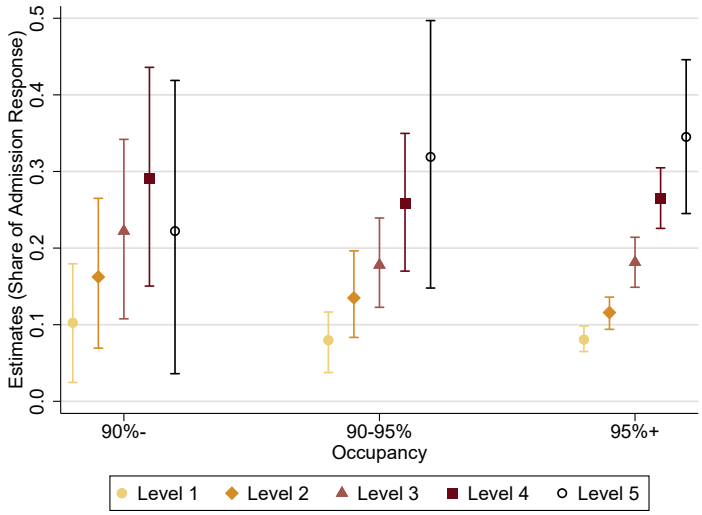
6.3 Occupancy and Patient Selection

Next, we test for selective admissions (Gandhi, 2023) by examining whether the composition of admission responses differs at different occupancy levels. Figure 7 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 response accounted for by the particular care level. The figure suggests that, although admissions of higher care levels 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 Figure S2).

Similarly, Figure S3 in Supplemental Appendix S.A plots the coefficients for the 1-week or 4-week admissions of long- vs. short-stay patients. Similar to Figure 7, the numbers represent the share of long or short stays in the admission response. The share of long-stay admissions in 1 week or 4 weeks increases slightly as occupancy increases (although the increase is not significant). The larger share of long stays at higher occupancy rates is likely because providers prefer patients who fill in their beds for longer periods of time. All of the empirical results except Figure S3 focus on (the more preferred) long-stay patients, so the dynamic incentive for selection along this dimension is likely unimportant.

[33]The results are also similar when we instead strengthen the sample restriction, by dropping facilities with the maximum occupancy below 50pp from our main sample (results available upon request).

Figure 7: Effect of Empty Beds on Admissions, By Occupancy and Care Levels



Notes: Figure 7 plots the IV estimates of the coefficient on occupancy in Eq. (5) 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.

7 Discussions

7.1 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 the cost effect. The response pattern is unlikely to be explained by variation in the income effect, because responses would then decrease with occupancy under plausible utility functions for income.

The cost effect is also likely to explain a large part of the overall level of admission responses. The discussion of Proposition 1-(iii) suggests that the income effect for 1-week admissions is at most 0.07pp (i.e., the admission response at occupancy below 85%), so the cost effect at occupancy strictly between 95% and 100% is at least 0.23pp

(= 0.30pp − 0.07p), or 76.7% (= $\frac{0.23}{0.30}$) of the total response.[34] Moreover, the income effect is likely less than 0.07pp, because demand inducement is difficult in the short run. Thus, the cost effect will account for some part of admission responses at all occupancy levels. Even if we use the results of 4-week admissions, for which demand inducement is likely easier, the cost effect accounts for at least 0.31pp (= 0.62pp − 0.31pp) or 50% (=$\frac{0.31}{0.62}$).

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 6a). 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 6a), 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 of the 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 7). Length of stay, another possible dimension of patient selection, is also not important, because we focus on long-stay patients. They are probably preferred to short-stay patients because of their longer lengths of stay, reducing demand uncertainty and admission/discharge costs.[35]

Several reasons may explain why the marginal cost of service increases with occupancy. First, congestion reduces the quality of service, thereby reducing the altruistic payoff of the facility. Lower service quality can also result from worker burnout or management slowdowns that prevent efficient care delivery. Second, medical resources

[34]The bound for the income effect is even larger if we include observations with binding capacity constraints.

[35]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.

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 quite consistent with some of the explanations: the patient-to-nurse ratio (a measure of inverse quality and that of the harshness of the work environment) tends to be higher at higher occupancy rates.

7.2 Potential Gains from Reallocation

Our results suggest that aggregate access to care (total admissions) can be improved by smoothing occupancy across similar facilities (Proposition 3). Such a policy may be of interest because short-run adjustment of capacity may be difficult due to regulations, such as certificate of need (CON) laws, or capital adjustment frictions. Below, we simulate a simple policy which reallocates a patient from a high-occupancy facility to a low-occupancy facility.[36]

Specifically, we first define a market by a unique combination of city, fiscal year, and facility size, where a facility's size is defined to be large if its bed capacity is above the city-fiscal year-specific median and it is small otherwise. We then consider a one-time reallocation of a patient from the most occupied facility (denoted $j = h$) to the least occupied facility ($j = l$), among facilities with occupancy rate strictly below 100%, within each market. Within-market reallocation leaves patient-facility distance little affected, and it confines the simulation to reallocation between facilities that face similar input and output markets.[37] We focus on facilities with an empty

[36] Although reallocation does not change primitive parameters such as bed counts, our estimates may be insufficient for simulation under some conditions, e.g., when admission depends on the occupancy of competing facilities via application decisions. We assume that there is excess demand so that facilities' acceptance decisions are the main driver of admissions.

[37] To further focus on similar facilities, we drop facilities with fewer than 50 beds, before imposing the above restrictions. We also drop (i) markets where reallocation reverses the congestion group (i.e., facility h belongs to a lower-occupancy group than facility l after reallocation) and (ii) markets where the capacity of facilities h and l differs by more then 20. Condition (i) is imposed to focus on markets where reallocation does smooth occupancy. Reallocation may still *reduce* total admissions if facilities are so asymmetric in bed count that the increase in occupancy of facility l is dominantly larger than the decrease in occupancy of facility h. Condition (ii) mitigates this concern.

bed to eliminate the mechanical effect of binding capacity constraints; in this sense, our result is conservative. Also, we consider marginal reallocation rather than perfect smoothing, because our estimates are only marginal effects. We then compute the effect of reallocation on admissions, using the IV regression results above. Our main analysis simulates reallocation on the first day of each fiscal year (April 1st).[38] In Supplemental Appendix S.A, we also show the simulation results using average occupancy at the facility-fiscal year level, rather than occupancy on a single day, to focus on the discrepancy of occupancy over a longer term.

We measure the effect of reallocation on admissions at facility j by

$$\begin{aligned}\Delta_j^a &= E\left[Y_j^{\text{admission}} \mid OC_j^{\text{post}}\right] - E\left[Y_j^{\text{admission}} \mid OC_j^{\text{pre}}\right] \\ &= \beta^{a,OC_j^{\text{post}}} OC_j^{\text{post}} - \beta^{a,OC_j^{\text{pre}}} OC_j^{\text{pre}}\end{aligned} \tag{8}$$

where $\beta^{a,OC}$ is the coefficient from admission regression Eq. (5) for facilities with occupancy level OC, depicted in Figure 6a.[39] OC_j^{post} (OC_j^{pre}) is the occupancy rate before (after) reallocation. We also show the results using regressions by occupancy percentiles instead of levels.[40]

Table 3 summarizes the effects of reallocation. In an average market, reallocation results in a net increase in admissions in the next week which corresponds to a 1.3pp, or a 36.0%, increase in occupancy.[41] The effect becomes smaller when we examine four-week admissions, or when we use estimates based on occupancy percentiles rather than levels. Still, the most conservative result suggests an average increase of 8.1% in total 4-week admissions. The results are qualitatively similar when we use average occupancy over a fiscal year rather than occupancy on the first day of a fiscal year, though the effects based on occupancy percentiles become smaller (see Table S5). This is likely because averaging reduces discrepancies in occupancy percentiles across facilities. To the extent that reallocation is simulated between homogeneous facilities, this exercise illustrates some degree of spatial misallocation (cf. Hsieh and Klenow,

[38] The choice of the specific day of a fiscal year is likely innocuous, because occupancy is highly correlated within a fiscal year.

[39] For facilities with occupancy at or above 95%, we use the coefficient for 95-99%.

[40] In this case, we move a patient from the facility in the highest occupancy percentile to the facility in the lowest percentile.

[41] The effect is large because even a marginal reallocation of a patient induces a large change: recall that the average long-stay admission per day is 0.2 (Table 1).

2009) of patients, and potential efficiency gains from reallocation.

Table 3: Simulated Results of Reallocation (Facilities with $\geq$ 50 Beds)

	Mean	Std. Dev.	Median	Obs.(Market)
	(1)	(2)	(3)	(4)
Panel A: Reallocation based on occupancy level				
(i) Change in 1-Week Admission				
Percentage point	1.34	4.97	0.07	2,056
Percentage change	35.98	133.33	1.93	2,056
(ii) Change in 4-Week Admission				
Percentage point	1.42	5.86	0.12	2,056
Percentage change	9.58	39.27	0.82	2,056
Panel B: Reallocation based on occupancy percentile				
(i) Change in 1-Week Admission				
Percentage point	1.48	5.13	0.12	1,930
Percentage change	39.57	137.42	3.23	1,930
(ii) Change in 4-Week Admission				
Percentage point	1.20	5.15	0.12	1,930
Percentage change	8.05	34.53	0.80	1,930

Notes: The table summarizes the net changes in admissions after reallocation, at the market level (sum of Δ_j^a in Eq. (8) within each market), where a market is defined by a unique combination of city, first day of the fiscal year, and facility size (large or small). "Percentage point" shows the effect on admissions represented as a pp change in occupancy (so the effect of 1 means a 1pp increase). "Percentage change" shows the effect on admissions as a fraction of pre-intervention admissions of the two intervened facilities (so the effect of 1 means a 1% increase).

7.3 Are Increased Admissions Good?

We do not directly examine whether the increased admissions due to reallocation are valuable to the admitted patients. The discussion in Section 7.1 suggests the importance of the cost effect in determining admissions, so the increased admissions may be those that are valuable but would be deterred without the capacity created

by an occupancy reduction. Some results are consistent with this hypothesis. First, the immediate increase in admissions following patient deaths, as shown in Figure 3, suggests that the "marginal patients" (those admitted in response to a reduction in occupancy) are likely to come from a waiting list. In-facility care is plausibly valuable to such waiting patients. Second, Figures 7 and S2 indicate that the marginal patients tend to be high-needs patients for whom in-facility care may be more valuable.

This paper also abstracts from the effect of occupancy on the health outcomes of patients already admitted, and how the outcomes are affected by occupancy-smoothing reallocation of patients. Analyzing health effects will require major changes to the framework of this paper. First, we will need to control for the characteristics of individual patients, which will require a patient-level analysis rather than a facility-level analysis. Second, the effect of daily occupancy on health outcomes will be negligible, so we will need to study the effect of longer-run occupancy, such as the average occupancy during an episode of stay. Finally, patient deaths are unlikely to be a good instrument for occupancy, because they are directly correlated with health. In a companion paper (Saruya and Takahashi, 2024), we address these issues to study the effect of congestion on patient outcomes measured by discharge outcomes (home discharge, hospitalization, and death), and use the estimates to simulate occupancy smoothing. The paper finds a negative effect of occupancy on outcomes (i.e., congestion leading to fewer home discharges and more hospitalizations). Moreover, it highlights an important tradeoff between congestion and quality: more congested facilities tend to be of higher quality. Thus, occupancy-smoothing reallocation reduces the average congestion on one hand, but it reduces the average quality on the other hand. Still, the paper finds that such reallocation can improve patient outcomes.

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. Using Japanese long-term care claims data, we find

that a reduction in occupancy from baseline leads facilities to increase admissions, and the admission response is larger at higher baseline occupancy, consistent with capacity constraints driving the admission response to occupancy variation. A simulation shows that reallocating a patient between relatively homogeneous facilities to smooth occupancy rates can improve aggregate access to institutionalized care, suggesting the potential importance of spatial misallocation of patients.

This study has some limitations. 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 of this paper, as discussed in Section 7.3. We study the effect of congestion on patient outcomes in a companion paper (Saruya and Takahashi, 2024).

We conclude by discussing policy implications. First, despite the widespread use of policies to constrain supply capacity for preventing wasteful service provision (e.g., CON laws) and other purposes, such policies have an adverse effect of worsening access to services. CON laws in health care can increase healthcare spending if capacity constraints prevent people from accessing necessary care. Second, our results show that policies can focus on the allocation of capacity in addition to or instead of the overall level of capacity. Given the substantial cost of adjusting service capacity, policymakers can stir service users to improve the efficiency in service provision, without adding capacity. Possible policy tools include providing information (on the availability or wait times of facilities) to patients and increasing out-of-pocket costs ("congestion tax") for using congested facilities. Analysis of the effectiveness of these policies requires information on some primitive parameters (e.g., patients' responsiveness to financial incentives), and is an interesting topic for future research. Finally, labor market reforms to enable more flexible adjustment of staffing can mitigate the negative effect of congestion if inflexible staffing is the main source of the effect. We leave analysis of staffing to future research.

A Appendix

Proof of Proposition 1

Define a function

$$F(a,d;n)=\left[\begin{array}{c} MB^{A}(n+a-d)-MC^{P}(n+a-d)-MC^{A}(a) \\ MB^{D}(n+a-d)+MC^{P}(n+a-d)-MC^{D}(d) \end{array}\right] \tag{9}$$

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 as long as $MC^{P\prime}(\cdot)$, $MC^{A\prime}(\cdot)$, and $MC^{D\prime}(\cdot)$ are positive and $-V''(\cdot)$ is 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

$$\begin{aligned} \left[\begin{array}{c} \frac{\partial a^{*}}{\partial n}\big|_{n=\bar{n}} \\ \frac{\partial d^{*}}{\partial n}\big|_{n=\bar{n}} \end{array}\right] &=-J_{F}(\bar{a},\bar{d};\bar{n})^{-1}\left[\begin{array}{c} MB^{A\prime}(\bar{p})-MC^{P\prime}(\bar{p}) \\ MB^{D\prime}(\bar{p})+MC^{P\prime}(\bar{p}) \end{array}\right] \\ &=\frac{-MB^{A\prime}(\bar{p})+MC^{P\prime}(\bar{p})}{D_{J_F}(\bar{a},\bar{d};\bar{n})}\left[\begin{array}{c} -MC^{D\prime}(\bar{d}) \\ MC^{A\prime}(\bar{a}) \end{array}\right]. \end{aligned} \tag{10}$$

Thus, we have $-\frac{\partial a^{*}}{\partial n}>0$ and $-\frac{\partial d^{*}}{\partial n}<0$ at $n=\bar{n}$.

(i-b) With $MC^{A\prime}(\bar{a}) = \kappa_2^A$ and $MC^{D\prime}(\bar{d}) = \kappa_2^D$, Eq. (10) 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 $\frac{\partial \bar{p}}{\partial \kappa_2^A} = \frac{\partial \bar{p}}{\partial \kappa_2^D} = 0$.[42] Therefore, we have $\frac{\partial}{\partial \kappa_2^A}\left(-\frac{\partial a^*}{\partial n}(\bar{n})\right) < 0$, $\frac{\partial}{\partial \kappa_2^D}\left(-\frac{\partial a^*}{\partial n}(\bar{n})\right) > 0$, $\frac{\partial}{\partial \kappa_2^D}\left(\frac{\partial d^*}{\partial n}(\bar{n})\right) < 0$, and $\frac{\partial}{\partial \kappa_2^A}\left(\frac{\partial d^*}{\partial n}(\bar{n})\right) > 0$. □

(ii) By Eq. (10),

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

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

(iii) Let $p^* = n + a^* - d^*$. Then,

$$\begin{aligned} \frac{\partial}{\partial n}\begin{bmatrix} \frac{\partial a^*}{\partial n} \\ \frac{\partial d^*}{\partial n} \end{bmatrix} &= \frac{\partial}{\partial n}\left\{ \frac{-MB^{A\prime}(p^*) + MC^{P\prime}(p^*)}{\left(-MB^{A\prime}(p^*) + MC^{P\prime}(p^*)\right)\left(\kappa_2^A + \kappa_2^D\right) + \kappa_2^A \kappa_2^D} \begin{bmatrix} -\kappa_2^D \\ \kappa_2^A \end{bmatrix} \right\} \\ &= \left(1 + \frac{\partial a^*}{\partial n} - \frac{\partial d^*}{\partial n}\right) \\ &\quad \times \frac{\partial}{\partial p}\left\{ \frac{-MB^{A\prime}(p^*) + MC^{P\prime}(p^*)}{\left(-MB^{A\prime}(p^*) + MC^{P\prime}(p^*)\right)\left(\kappa_2^A + \kappa_2^D\right) + \kappa_2^A \kappa_2^D} \right\} \begin{bmatrix} -\kappa_2^D \\ \kappa_2^A \end{bmatrix} \\ &= \left(1 + \frac{\partial a^*}{\partial n} - \frac{\partial d^*}{\partial n}\right) \frac{\left(-MB^{A\prime\prime}(p^*) + MC^{P\prime\prime}(p^*)\right) \kappa_2^A \kappa_2^D}{\left\{D_{J_F}(a^*, d^*)\right\}^2} \begin{bmatrix} -\kappa_2^D \\ \kappa_2^A \end{bmatrix}. \end{aligned}$$

Because $1 + \frac{\partial a^*}{\partial n} - \frac{\partial d^*}{\partial n} > 0$ by (ii), $-\frac{\partial^2 a^*}{\partial n^2} > 0$ and $\frac{\partial^2 d^*}{\partial n^2} > 0$ at $n = \bar{n}$ if and only if $MC^{P\prime\prime}(\bar{p}) > MB^{A\prime\prime}(\bar{p})$. □

Note that $\frac{\partial D_{J_F}(a^*, d^*)}{\partial n} = \left(1 + \frac{\partial a^*}{\partial n} - \frac{\partial d^*}{\partial n}\right)\left(-MB^{A\prime\prime}(p^*) + MC^{P\prime\prime}(p^*)\right)\left(\kappa_2^A + \kappa_2^D\right)$, which is positive at $n = \bar{n}$ if and only if $MC^{P\prime\prime}(\bar{p}) > MB^{A\prime\prime}(\bar{p})$. This, together with (iii), establishes the statement in Section 3 that D_{J_F} increases with n if $-\frac{\partial^2 a^*}{\partial n^2} > 0$ and $\frac{\partial^2 d^*}{\partial n^2} > 0$.

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

Proof of Proposition 2

If κ_2^A or κ_2^D is positive (and the other is zero), the conclusion follows from Eq. (10). If both are zero, Eq. (9) implies $n + a^* - d^*$ is constant, yielding the conclusion. □

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.

Azoulay, Pierre, Jialan Wang, and Joshua Graff Zivin. 2010. "Superstar Extinction." *Quarterly Journal of Economics* 125 (2): 549–589.

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.

Becker, Sascha O, and Hans K Hvide. 2022. "Entrepreneur Death and Startup Performance." *Review of Finance* 26 (1): 163–185.

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. 2024. "The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid." Working Paper 29471, National Bureau of Economic Research.

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." http://www.kaigo-center.or.jp/report/pdf/h28_chousa_kekka.pdf (Japanese).

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

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*.

Evans, Robert G. 1974. "Supplier-Induced Demand: Some Empirical Evidence and Implications." In *The Economics of Health and Medical Care: Proceedings of a Conference held by the International Economic Association at Tokyo*, edited by Perlman, Mark 162–173, London: Palgrave Macmillan UK.

Fadlon, Itzik, and Torben Heien Nielsen. 2021. "Family Labor Supply Responses to Severe Health Shocks: Evidence from Danish Administrative Records." *American Economic Journal: Applied Economics* 13 (3): 1–30.

Fréchette, Guillaume R., Alessandro Lizzeri, and Tobias Salz. 2019. "Frictions in a Competitive, Regulated Market: Evidence from Taxis." *American Economic Review* 109 (8): 2954–92.

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.

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

Gavazza, Alessandro, and Alessandro Lizzeri. 2021. "Chapter 6 - Frictions in product markets." In *Handbook of Industrial Organization, Volume 4*, edited by Ho, Kate, Ali Hortaçsu, and Alessandro Lizzeri Volume 4. of Handbook of Industrial Organization 433–484, Elsevier.

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.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *The Quarterly Journal of Economics* 124 (4): 1403–1448.

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.

Jaravel, Xavier, Neviana Petkova, and Alex Bell. 2018. "Team-Specific Capital and Innovation." *American Economic Review* 108 (4-5): 1034–1073.

Jones, Benjamin F., and Benjamin A. Olken. 2005. "Do Leaders Matter? National Leadership and Growth Since World War II." *Quarterly Journal of Economics* 120 (3): 835–864.

Jäger, Simon, and Jörg Heining. 2019. "How Substitutable Are Workers? Evidence from Worker Deaths." *Working Paper*.

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.

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

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).

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. 2024. "Congestion-Quality Tradeoff: Evidence from Japanese Long-Term Care Facilities." Working paper.

Sauvagnat, Julien, and Fabiano Schivardi. 2023. "Are Executives in Short Supply? Evidence from Death Events." *The Review of Economic Studies* 91 (1): 519–559.

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.

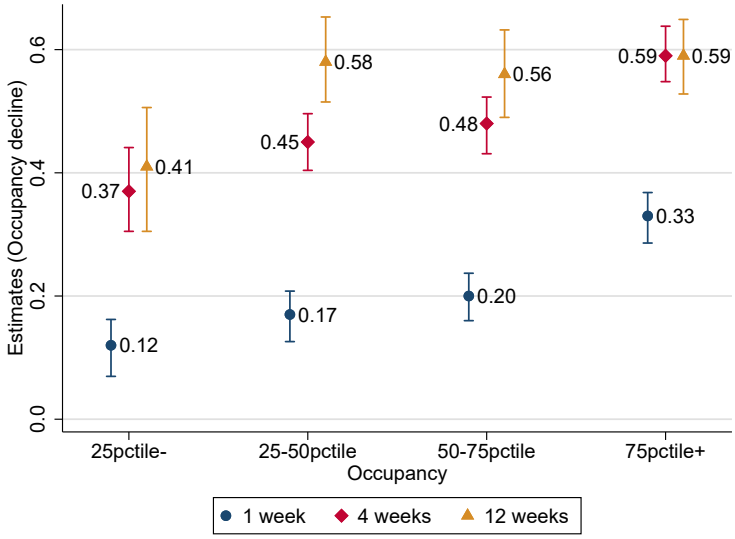
United Nations. 2019. "World Population Aging 2019: Highlights." https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf, Accessed December 5, 2022.

Supplemental Appendix

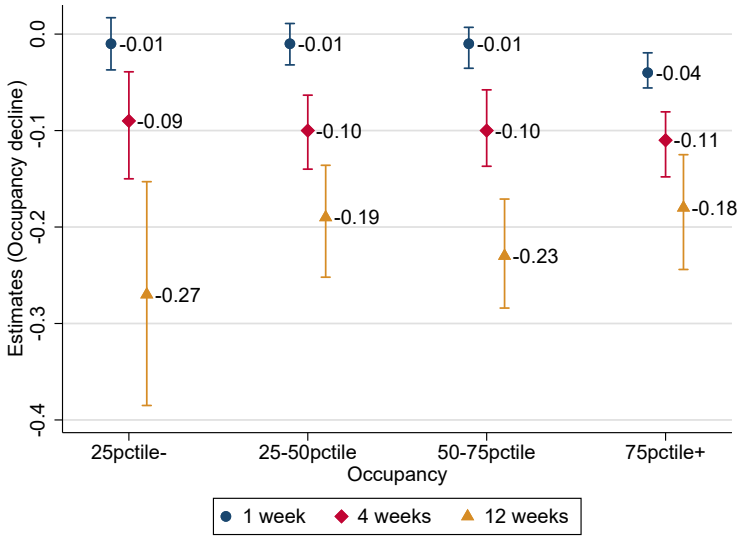
S.A Additional Figures and Tables

Figure S1: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Percentile

(a) Admission

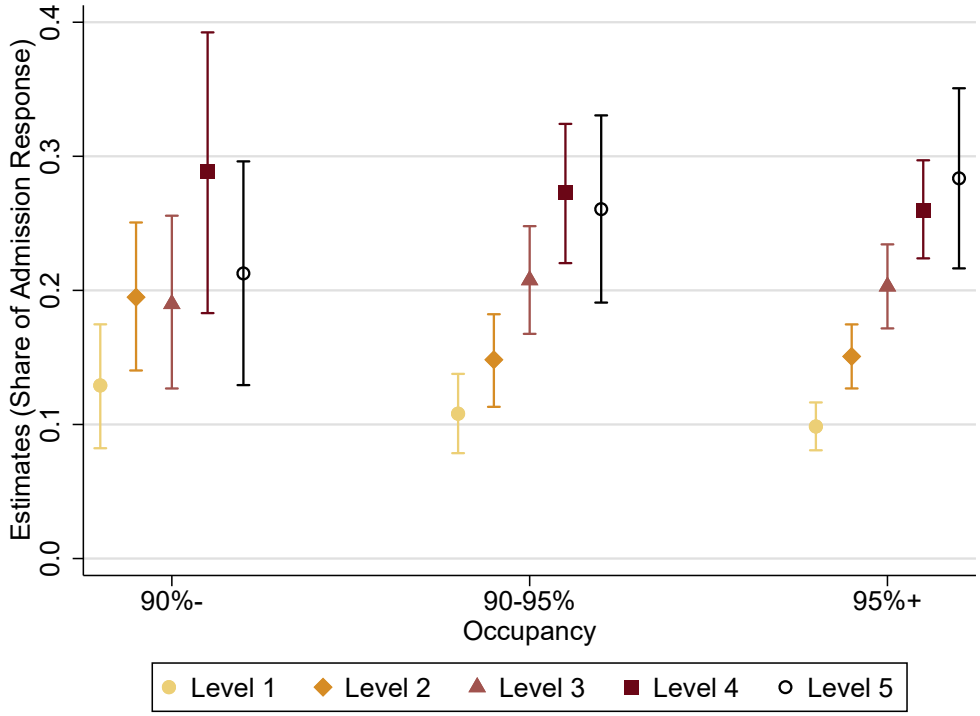


(b) Live Discharge



Notes: Figure S1 plots the IV estimates of the coefficient on occupancy in Eq. (5) 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.

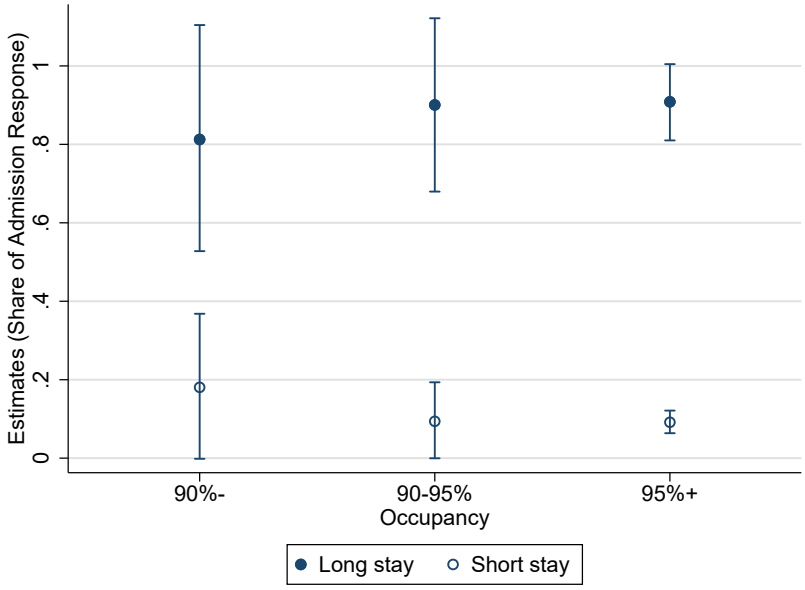
Figure S2: Effect of Empty Beds on Admissions, By Occupancy and Care Levels (4 Weeks)



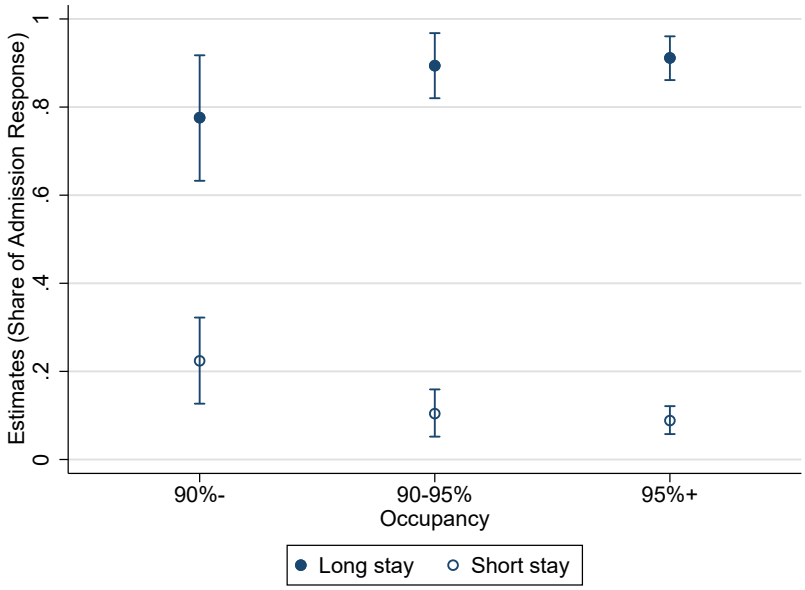
Notes: Figure S2 plots the IV estimates of the coefficient on occupancy in Eq. (5) and their 95% confidence intervals, using long-stay admissions in the following 4 weeks as the outcome, separately by occupancy and care levels. The coefficients are divided by the coefficient in the baseline regression that pools all care levels, so the numbers represent the share of each care level in the admission response.

Figure S3: Effect of Empty Beds on Admissions, By Occupancy and Long vs. Short Stays

(a) 1 Week



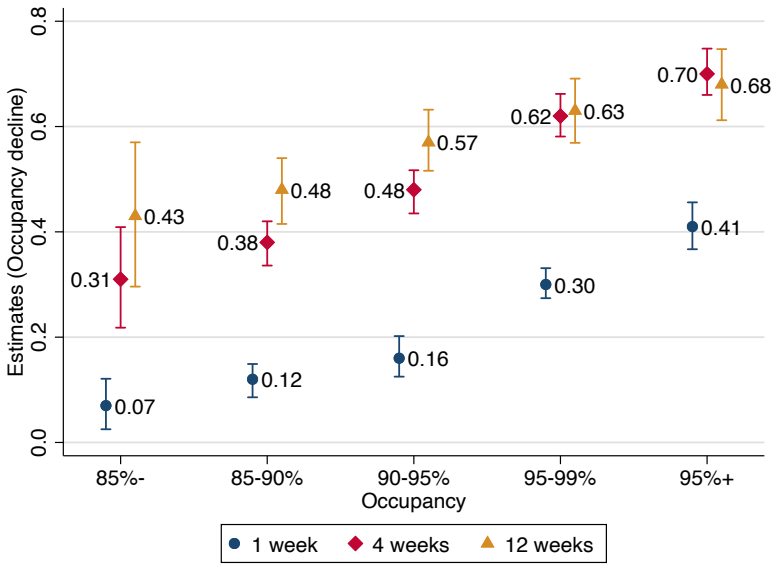
(b) 4 Weeks



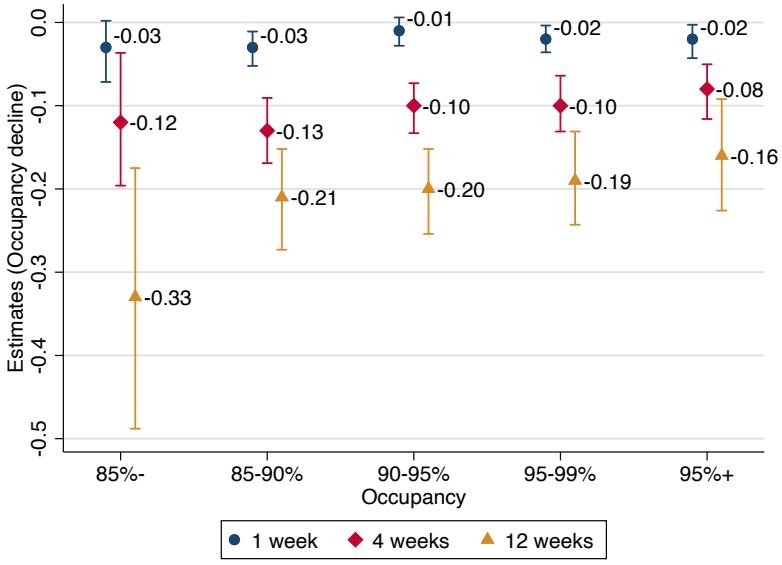
Notes: Figure S3 plots the IV estimates of the coefficient on occupancy in Eq. (5) and their 95% confidence intervals, using as outcomes long-stay (filled circle) or short-stay (open) admissions for the following 1 week (panel (a)) or 4 weeks (panel (b)). The coefficients are divided by the coefficient in the baseline regression that pools both long-stay and short-stay admissions, so the numbers represent the share of each type of admissions in the admission response.

Figure S4: Effect of Empty Beds on Admissions and Live Discharges, by Occupancy Level (Uncensored Sample)

(a) Admission



(b) Live Discharge



Notes: Figure S4 plots the IV estimates of the coefficient on occupancy in Eq. (5) 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.

Table S1: 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 S2: Per-diem Reimbursement

	Long stay			Short stay		
	Fixed (USD) (1)	FFS (USD) (2)	% of fixed pay (3)	Fixed (USD) (4)	FFS (USD) (5)	% of fixed pay (6)
Care level 1	78.0	10.3	88.3%	82.0	20.8	79.8%
Care level 2	83.1	10.4	88.9%	86.9	21.0	81.2%
Care level 3	89.4	10.5	89.5%	93.2	21.6	81.9%
Care level 4	95.2	10.6	90.0%	98.2	22.2	81.6%
Care level 5	101.3	10.4	90.4%	104.6	24.5	81.0%

Notes: The table presents daily averages of fixed and fee-for-service (FFS) payments in our analysis sample, separately by care levels and long/short stays. The averages are computed through the following steps. (1) Compute the daily averages of fixed and FFS payments within each patient-year-month-long/short 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-long/short level.

Table S3: 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)	1.201	0.270***	0.219***	1.814	0.637***	0.573***
	(2.948)	(0.0463)	(0.0230)	(3.951)	(0.0702)	(0.0273)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.287			0.363		
	(0.825)			(1.109)		
I(Occupancy≥ 90pp)		0.0488**			0.0618	
		(0.0229)			(0.0400)	
I(Occupancy≥ 85pp)			0.0576**			0.0729**
			(0.0285)			(0.0311)
Cragg-Donald F-stats	0.681	31.82	46.57	0.681	31.82	46.57
N	6,345,745	6,345,745	6,345,745	6,345,745	6,345,745	6,345,745

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.335***	0.291***	0.261***	0.720***	0.664***	0.626***
	(0.0891)	(0.0588)	(0.0428)	(0.106)	(0.0718)	(0.0561)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0206*			0.0261*		
	(0.0114)			(0.0144)		
I(Occupancy≥ 50pctile)		0.0178**			0.0225*	
		(0.00903)			(0.0118)	
I(Occupancy≥ 25pctile)			0.0285**			0.0360*
			(0.0136)			(0.0199)
Cragg-Donald F-stats	61.17	88.22	70.13	61.17	88.22	70.13
N	6,345,745	6,345,745	6,345,745	6,345,745	6,345,745	6,345,745

Notes: The table shows the estimates of the coefficients in the regression (7), 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 S4: Effect of Empty Beds on Admissions, with Nonlinear Terms (Uncensored Sample)

Panel A: Estimates based on Occupancy Levels

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.325***	0.308***	0.276***	0.656***	0.641***	0.613***
	(0.0307)	(0.0253)	(0.0198)	(0.0328)	(0.0292)	(0.0259)
Occupancy (decline) ×						
I(Occupancy≥ 95pp)	0.0539***			0.0477***		
	(0.0134)			(0.0146)		
I(Occupancy≥ 90pp)		0.0624***			0.0552***	
		(0.0141)			(0.0164)	
I(Occupancy≥ 85pp)			0.112***			0.0990***
			(0.0269)			(0.0297)
Cragg-Donald F-stats	157.4	135.4	80.56	157.4	135.4	80.56
N	6,786,594	6,786,594	6,786,594	6,786,594	6,786,594	6,786,594

Panel B: Estimates based on Percentiles of Occupancy

	1-Week Admission			4-Week Admission		
Occupancy (decline)	0.330***	0.318***	0.298***	0.661***	0.650***	0.633***
	(0.0303)	(0.0264)	(0.0231)	(0.0326)	(0.0298)	(0.0273)
Occupancy (decline) ×						
I(Occupancy≥ 75pctile)	0.0328***			0.0290***		
	(0.00783)			(0.00875)		
I(Occupancy≥ 50pctile)		0.0301***			0.0266***	
		(0.00691)			(0.00791)	
I(Occupancy≥ 25pctile)			0.0464***			0.0411***
			(0.0107)			(0.0123)
Cragg-Donald F-stats	244.3	293.3	209.2	244.3	293.3	209.2
N	6,786,594	6,786,594	6,786,594	6,786,594	6,786,594	6,786,594

Notes: Table S4 shows the estimates of the coefficients in the regression (7). 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 S5: Simulated Results of Reallocation (Facilities with ≥ 50 Beds, Using Average Occupancy)

	Mean	Std. Dev.	Median	Obs.(Market)
	(1)	(2)	(3)	(4)
Panel A: Reallocation based on occupancy level				
(i) Change in 1-Week Admission				
Percentage point	1.97	8.00	0.05	2,102
Percentage change	52.93	213.67	1.46	2,102
(ii) Change in 4-Week Admission				
Percentage point	1.61	7.44	0.11	2,102
Percentage change	10.80	49.91	0.72	2,102
Panel B: Reallocation based on occupancy percentile				
(i) Change in 1-Week Admission				
Percentage point	0.91	3.96	0.05	1,591
Percentage change	24.35	106.13	1.35	1,591
(ii) Change in 4-Week Admission				
Percentage point	0.46	4.40	0.07	1,591
Percentage change	3.11	29.52	0.48	1,591

Notes: The table summarizes the net changes in admissions after reallocation, at the market level (sum of Δ_j^a in Eq. (8) within each market), where a market is defined by a unique combination of city, fiscal year, and facility size (large or small). "Percentage point" shows the effect on admissions represented as a pp change in occupancy (so the effect of 1 means a 1pp increase). "Percentage change" shows the effect on admissions as a fraction of pre-intervention admissions of the two intervened facilities (so the effect of 1 means a 1% increase).

S.B Theory of Reallocation

This section provides a formal theory of reallocation intuitively discussed in Section 3.3.

S.B.1 Model

Consider the following two-period, two-facility setting. There are two homogeneous facilities, 1 and 2. In period 1, each of the two facilities chooses its initial occupancy, to maximize its payoff. In period 2, given the occupancy n_j, facility j chooses new admissions a_j to maximize its payoff. To allow for heterogeneity in n_j, we assume that facilities face idiosyncratic payoff and occupancy shocks in period 1. We make the following simplifying assumptions: (1) V is linear, $b^P = b$, and $b^A = 0$. (2) The facilities only choose admissions. (3) The facilities are myopic.[43]

Period 1. The utility of facility $j \in \{1, 2\}$ in period 1 is $U_{j1} = (r + b)n - C^P(n) - C^A(n)$, where r denotes per-patient profitability and b denotes per-patient altruistic utility. C^P and C^A are both strictly increasing and strictly convex. We assume that the optimal admission is an interior solution determined by the first-order condition. Facilities face transitory shocks to $\left(r, b, C^P, C^A\right)$ before choosing the initial occupancy, and there can also be direct shocks to the occupancy rate. Examples of such shocks are emergency admissions, deaths, spatial or information frictions, and staff shortages (see footnote 22).

Period 2. Given occupancy n_j, facility j chooses new admissions a_j to maximize $U_{j2} = (r + b)(n_j + a_j) - C^P(n_j + a_j) - C^A(a_j)$. We omit idiosyncratic shocks for this period to focus our analysis on the effect of heterogeneous initial occupancy. We assume the solution $(a_1^*(n_1), a_2^*(n_2))$ and resulting occupancy are interior.

The myopic facilities solve the period-1 problem to obtain $n_j = n_j^*(r, b, C^P, C^A)$. Due to idiosyncratic shocks, n_j can vary between the two facilities. In period 2, given n_j, facility j chooses a_j to maximize $U_{j2}(a_j; n_j, r, b, C^P, C^A)$.

The government can reallocate $N = n_1 + n_2$ patients to facility 1 or 2 before the facilities make an admission decision. The government is concerned with total access to care, or the quantity of service production, given by $A(n_1; N) = n_1 + a_1^*(n_1) + n_2 +$

[43] Dynamics will complicate the analysis without affecting the mechanisms in our static model.

$a_2^*(n_2) = N + a_1^*(n_1) + a_2^*(N - n_1)$. We now show that occupancy smoothing improves this objective. As in Section 3, we assume $MC^A(a) = \kappa_1^A + \kappa_2^A a$, $\kappa_1^A, \kappa_2^A \geq 0$.

Proposition 3. *Suppose $MC^{P\prime\prime}(p) > 0$ for all p and $\kappa_2^A > 0$. Then $A(n_1; N)$ is strictly increasing in n_1 if and only if $n_1 < n_2$, and is maximized at $n_1 = \frac{N}{2}$.*

The proof is in Appendix A. Given the total number of initial in-facility patients $N = n_1 + n_2$, an occupancy-smoothing policy that moves patients from the more congested facility to the less congested facility increases aggregate access to care. Thus, aggregate access is maximized by setting $n_1 = n_2 = \frac{N}{2}$.

S.B.2 Proof of Proposition 3

Let

$$G\left(a_1, a_2; n_1, N\right) = \begin{bmatrix} r + b - MC^P(n_1 + a_1) - MC^A(a_1) \\ r + b - MC^P(N - n_1 + a_2) - MC^A(a_2) \end{bmatrix}. \tag{11}$$

The optimal admission satisfies $G\left(a_1^*, a_2^*; n_1, N\right) = 0$. Note that if $n_1 = N - n_1$, then $a_1^* = a_2^*$ and if $n_1 < (>) N - n_1$, then $a_1^* > (<) a_2^*$. The Jacobian of G,

$$J_G = \begin{bmatrix} -MC^{P\prime}(n_1 + a_1) - \kappa_2^A & 0 \\ 0 & -MC^{P\prime}(N - n_1 + a_2) - \kappa_2^A \end{bmatrix}, \tag{12}$$

is full rank at all (n_1, a_1, a_2) because of strict convexity.

Let $p_j^* = n_j + a_j^*$. By the implicit function theorem, at any given n_1, we have

$$\begin{bmatrix} \frac{\partial a_1^*}{\partial n_1} \\ \frac{\partial a_2^*}{\partial n_1} \end{bmatrix} = -J_G^{-1} \begin{bmatrix} \frac{\partial G_1}{\partial n_1} \\ \frac{\partial G_2}{\partial n_1} \end{bmatrix} = \begin{bmatrix} \frac{-MC^{P\prime}(p_1^*)}{MC^{P\prime}(p_1^*) + \kappa_2^A} \\ \frac{MC^{P\prime}(p_2^*)}{MC^{P\prime}(p_2^*) + \kappa_2^A} \end{bmatrix}. \tag{13}$$

Differentiating both sides with respect to n_1 yields

$$\begin{aligned}\frac{\partial}{\partial n_1}\begin{bmatrix}\frac{\partial a_1^*}{\partial n_1}\\ \frac{\partial a_2^*}{\partial n_1}\end{bmatrix} &= \begin{bmatrix}\left(1+\frac{\partial a_1^*}{\partial n_1}\right)\frac{\partial}{\partial p_1}\left\{\frac{-MC^{P\prime}(p_1)}{MC^{P\prime}(p_1)+\kappa_2^A}\right\}\Big|_{p_1=p_1^*}\\ \left(-1+\frac{\partial a_2^*}{\partial n_1}\right)\frac{\partial}{\partial p_2}\left\{\frac{MC^{P\prime}(p_2)}{MC^{P\prime}(p_2)+\kappa_2^A}\right\}\Big|_{p_2=p_2^*}\end{bmatrix}\\ &= \begin{bmatrix}\left(1+\frac{\partial a_1^*}{\partial n_1}\right)\frac{-MC^{P\prime\prime}(p_1^*)\kappa_2^A}{\left\{MC^{P\prime}(p_1^*)+\kappa_2^A\right\}^2}\\ \left(-1+\frac{\partial a_2^*}{\partial n_1}\right)\frac{MC^{P\prime\prime}(p_2^*)\kappa_2^A}{\left\{MC^{P\prime}(p_2^*)+\kappa_2^A\right\}^2}\end{bmatrix}.\end{aligned}$$

By Eq.(13), we have $\frac{\partial a_1^*}{\partial n_1} > -1$ and $\frac{\partial a_2^*}{\partial n_1} < 1$. Therefore,

$$\begin{aligned}\frac{\partial\left(\frac{\partial a_1^*}{\partial n_1}+\frac{\partial a_2^*}{\partial n_1}\right)}{\partial n_1} &= \left(1+\frac{\partial a_1^*}{\partial n_1}\right)\frac{-MC^{P\prime\prime}(p_1^*)\kappa_2^A}{\left\{MC^{P\prime}(p_1^*)+\kappa_2^A\right\}^2}+\left(-1+\frac{\partial a_2^*}{\partial n_1}\right)\frac{MC^{P\prime\prime}(p_2^*)\kappa_2^A}{\left\{MC^{P\prime}(p_2^*)+\kappa_2^A\right\}^2}\\ &< 0.\end{aligned}$$

Thus, $\frac{\partial a_1^*}{\partial n_1}+\frac{\partial a_2^*}{\partial n_1}$ strictly decreases with n_1. Moreover, we have $\frac{\partial a_1^*}{\partial n_1}+\frac{\partial a_2^*}{\partial n_1}=0$ if $n_1 = N-n_1$ ($n_1=\frac{N}{2}$). Therefore, if $n_1 < n_2 = N-n_1$ ($n_1<\frac{N}{2}$), then we have $\frac{\partial a_1^*}{\partial n_1}+\frac{\partial a_2^*}{\partial n_1}>0$ and $A(n_1;N)=N+a_1^*(n_1)+a_2^*(N-n_1)$ is increasing in n_1, and if $n_1>n_2$ ($n_1>\frac{N}{2}$), then we have $\frac{\partial a_1^*}{\partial n_1}+\frac{\partial a_2^*}{\partial n_1}<0$ and $A(n_1;N)$ is decreasing in n_1. Thus, $A(n_1;N)$ is maximized at $(n_1,n_2)=\left(\frac{N}{2},\frac{N}{2}\right)$. □