Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Identifying the Bottleneck Unit: Impact of Congestion Spillover in Hospital Inpatient Unit Network

Song-Hee Kim, Fanyin Zheng, Joan Brown

To cite this article:
Song-Hee Kim, Fanyin Zheng, Joan Brown (2024) Identifying the Bottleneck Unit: Impact of Congestion Spillover in Hospital Inpatient Unit Network. Management Science 70(7):4200-4218. https://doi.org/10.1287/mnsc.2023.4887

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 7, July 2024, pp. 4200–4218
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Identifying the Bottleneck Unit: Impact of Congestion Spillover in Hospital Inpatient Unit Network

Song-Hee Kim,[a,*] Fanyin Zheng,[b] Joan Brown[c]

[a] SNU Business School, Seoul National University, Seoul 08826, South Korea; [b] Columbia Business School, Columbia University, New York, New York 10027; [c] Keck School of Medicine, University of Southern California, Los Angeles, California 90033

*Corresponding author

Contact: songheekim@snu.ac.kr, https://orcid.org/0000-0002-3106-5726 (S-HK); fanyin.zheng@columbia.edu, https://orcid.org/0000-0001-5215-8721 (FZ); joan.brown@med.usc.edu, https://orcid.org/0000-0003-3492-2858 (JB)

Received: August 5, 2020
Revised: March 7, 2022; August 3, 2022
Accepted: August 22, 2022
Published Online in Articles in Advance: August 22, 2023

https://doi.org/10.1287/mnsc.2023.4887

Abstract. Because a hospital is an interconnected, interdependent network of care units, allocating resources—beds, nurses, and improvement initiatives—to one unit to reduce its congestion may have spillover effects on other units. If such congestion spillover is substantial, ignoring it may lead to unintended consequences and missed opportunities. We use data collected over five years from a hospital with 16 inpatient units to empirically examine whether and how much congestion propagates through the network of inpatient units. Our estimation result suggests that the magnitude of the congestion spillover is indeed substantial in our study hospital. For example, increasing one inpatient unit's utilization by 10 percentage points today can increase its neighboring inpatient unit's utilization by up to 4.33 percentage points tomorrow. Using counterfactual analyses, we estimate the effect of adding a bed to each unit. We find that due to congestion spillover, adding one bed to the bottleneck unit can free up 4.14 beds in the hospital, which translates to 383.53 more hospital visits per year or a 3% increase in hospital throughput. This effect is about three times bigger in magnitude compared with what one can achieve by naively choosing which unit to add a bed to. Hospitals and other manufacturing and service systems with complex interdependence across resources can use our empirical framework to examine the spillover effect of resources on performance metrics and leverage such understanding to effectively improve their operations.

History: Accepted by Vishal Gaur, operations management.
Funding: This work was supported by the Creative-Pioneering Researchers Program through Seoul National University.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mnsc.2023.4887.

Keywords: empirical operations management • healthcare • network effect • bottleneck • congestion • congestion spillover • inpatient unit

1. Introduction

Research in the field of healthcare operations management over the last two decades has shown that operational characteristics can have a significant influence on care delivery and patient outcomes (KC et al. 2020). One of the key operational characteristics is bed utilization in hospitals. High bed utilization has been shown to decrease treatment duration (Forster et al. 2003, KC and Terwiesch 2012), decrease the likelihood of admission to certain inpatient units (Kim et al. 2015), increase the waiting time for admission to or transfer between inpatient units (Shi et al. 2016, Long and Mathews 2018), increase ambulance diversion (McConnell et al. 2005, Allon et al. 2013), and increase the likelihood of assigning a patient to a bed designated to a service that differs from the service the patient needs (Dong et al. 2019, Song et al. 2020). All these practices can have a serious impact on clinical outcomes, including an increased mortality rate, a higher readmission rate, and a longer hospital length of stay (KC and Terwiesch 2012, Kim et al. 2015, Kuntz et al. 2015, Berry Jaeker and Tucker 2016).

To prevent the detrimental effects of high bed utilization, hospital administrators may seek strategies to decrease bed utilization and improve bed availability. For example, hospital administrators may consider adding beds to an inpatient unit or targeting an inpatient unit to reduce patients' length of stay by streamlining the unit's care delivery process.

So, how should hospital administrators identify which unit to target to maximize the effect on hospital-wide performance? We argue that, for two reasons, this is not a trivial question. First, the key to improving a process is to identify and alleviate the bottleneck. In a process with multiple types of jobs, the resource (in our context, the inpatient unit or inpatient beds) with the highest implied utilization (demand for the resource divided by the

capacity of the resource) is the bottleneck (Cachon and Terwiesch 2012). However, this traditional bottleneck analysis applies only to systems with fixed demand and routing and a relatively simple process structure with no resource sharing. That is, *the traditional bottleneck analysis is likely to fail for hospitals* because resources such as beds are flexible and, in turn, are shared across units, especially when demand for some units becomes high. For example, when an inpatient unit is at or near its capacity and hence cannot admit new patients, hospitals do not let patients wait indefinitely until beds open up in that unit (Chan et al. 2012). Instead, patients' care pathways are altered by, for example, sending a patient to an off-service unit—one that is designated to a service that differs from the service the patient needs (Dong et al. 2019, Song et al. 2020). In other words, the true demand for each unit is not observed directly due to spillover effects across units, and the realized demand in each unit is likely to be endogenous to the hospital's congestion level. Furthermore, the provision of care for some patients needs to be completed in multiple units: Intensive care unit (ICU) patients are typically transferred to a floor unit to complete the provision of care before hospital discharge. The interdependence of patient needs across units further complicates the identification of the bottleneck in the hospital's inpatient unit network.

Second, a hospital is an interconnected entity. A 2017 Institute for Healthcare Improvement white paper states: "Making meaningful and sustainable changes to hospital operations, including patient flow, requires recognizing the interdependent nature of every facet of the hospital. Understanding hospital-wide patient flow requires looking at the whole system of care, not just individual patient care units or subgroups of patients" (Rutherford et al. 2017, p. 7). That is, *a change made to an inpatient unit is likely to have a spillover effect on other units in the hospital*. Hence, hospital administrators need to consider the interdependent nature of care units as they seek effective strategies to decrease bed utilization and improve bed availability.

In this paper, we (1) examine whether congestion spillover exists in a hospital inpatient unit network; (2) if it does, estimate the magnitude of the congestion spillover effect; and (3) use the estimation results to empirically identify the bottleneck unit.

We use patient movement data collected over five years from an urban teaching hospital in the United States. This hospital has 16 inpatient units, and our goal is to estimate the congestion spillover effect among these 16 units—that is, the effect of changes to one unit's utilization on the utilization of other units.[1] Working with observational data poses an important econometric challenge. As mentioned previously, realized patient flow is endogenous to the congestion level in the hospital. To address this endogeneity issue, we use the instrumental variable approach, whereby we construct instruments by exploiting the variation in the structure of the inpatient unit network (Bramoullé et al. 2009, Drakopoulos and Zheng 2017). Using counterfactual analyses, we show how our estimation results can be used to identify the bottleneck unit—the unit that has the biggest impact on system performance when its capacity is increased through an intervention.

We find that there is congestion spillover in our study hospital's inpatient unit network and that its magnitude is substantial. For example, we find that increasing one inpatient unit's utilization by 10 percentage points today can increase its neighboring inpatient unit's utilization by up to 4.33 percentage points tomorrow. In addition, by comparing our results, obtained with two-stage least squares regressions using instrumental variables, with the results obtained by ordinary least squares, we find evidence for the existence of significant endogeneity bias. This highlights the need to correctly identify and estimate the congestion spillover effects in our setting. Via counterfactual analyses, we estimate that adding a bed to a unit can free up 1.02 to 4.14 beds in the long run, depending on which unit the change is applied to—We emphasize our finding that adding one bed frees up more than one bed in the long run due to the congestion spillover effects. Freeing up 4.14 beds can translate to an increase in hospital throughput by 385.53 more hospital visits per year, or a 3% increase in hospital throughput. Importantly, our analyses illustrate that our study hospital may miss the opportunity to more than triple the benefit that the additional bed can generate if it does not take congestion spillover into account when making bed management decisions.

In summary, we make the following key contributions:

- We show that the magnitude of the congestion spillover is substantial in our study hospital, implying that if they do not account for the congestion spillover effect, hospitals might miss the target when evaluating which unit should receive additional resources. More generally, our results highlight the need for taking hospital-wide interdependency and endogenous patient routing into account and taking a network-conscious approach to effectively understanding and improving hospital operations.
- On the methodological side, we adapt existing econometric tools and identify suitable instrumental variables to estimate the causal spillover effect in the inpatient unit network without bias. To the best of our knowledge, this study is the first to use scheduling information to construct instruments in the healthcare operations management and the empirical operations management literature.
- We show that using our empirical framework and limited observational data, hospitals can identify their bottleneck unit, taking into account the complex and endogenous patient flows and the congestion spillover

effect in the hospital inpatient unit network. Hospital administrators can combine our results with their hospitals' cost information to answer important capacity-related questions, such as whether to add a bed and to which unit they should allocate the additional bed.

- Our empirical framework can be easily applied to study other hospital resources and performance metrics. Beyond hospitals, our empirical framework can be applied to manufacturing and service processes that have flexible resources and/or a complex network of resources with endogenous routing of jobs within the network, and that produce highly customized products. One example can be found in the classic Donner Company case (HBS 1998), which discusses a manufacturing process that produces highly customized printed circuit boards; other examples include service operations with similar features, such as management consulting firms, law firms, and architecture companies.

The remainder of the paper is structured as follows. We provide an overview of the related literature in Section 2. Section 3 describes the data set for our analyses. We introduce our model and describe the econometric challenge and our approach to address it in Section 4. Section 5 presents our empirical findings. We conduct counterfactual analyses and discuss the implications of our findings in Section 6. Finally, we conclude in Section 7.

2. Related Literature

2.1. Examining Network Effects in Hospital Settings

Almost two decades of active research has yielded a growing body of knowledge about understanding and improving hospital operations. Much of the work to date focuses on individual patient care units or the interactions between a small set of patient care units.

Some studies focus on understanding the impact of the inpatient unit operations on the emergency department (ED). They have found that higher utilization in inpatient units is associated with shorter length of stay in the ED (Forster et al. 2003), longer boarding time (waiting for admission to inpatient units) in the ED (Shi et al. 2016), and more ambulance diversion (Allon et al. 2013). McConnell et al. (2005) find that increased ICU capacity decreases both ambulance diversion and ED length of stay. Chan et al. (2017) find that delays in receiving ICU care because of delays in ED-ICU transfer is associated with longer length of stay in the ICU.

Another body of the literature has focused on the interactions among inpatient units. Long and Mathews (2018) find that high utilization in downstream units creates substantial delay in patient transfers from the ICU to downstream units. Dong et al. (2019) and Song et al. (2020) show that the off-service placement of a patient—that is, the patient is assigned to a bed designated to a service that differs from the service the patient needs—occurs frequently. In their paper examining patients receiving coronary artery bypass graft, Clark and Huckman (2012) study the performance spillovers among different specialized units within the same hospital and show that specialization in areas related to cardiovascular care directly improves the clinical outcomes for cardiovascular patients.

Some studies focus on patient transfers from inpatient units to locations outside the hospital. Zychlinski et al. (2020) examine bed blocking in inpatient units due to lack of beds at geriatric institutions for elderly patients and develop a mathematical model to offer insights into bed allocation decisions. Copenhaver et al. (2019) find that the hospital discharge process has a significant impact on congestion in inpatient units and redesign the discharge process by predicting hospital discharges in the next 24 hours.

A handful of prior studies take hospital-wide interdependency into account in examining hospital operations. For example, Armony et al. (2015) conduct exploratory data analysis of patient flow and provide a data-based queueing-network view of patient flow in hospitals; they conclude that their findings underscore the need for an integrative view of hospital units. Freeman et al. (2021) investigate volume-cost spillover effects between elective and emergency admissions and across specialties. They find that increased elective volume is associated with increased costs of emergency care and that increased emergency activity in one specialty is associated with decreased costs of emergency care in other specialties. Thomas et al. (2013) develop hospital-wide bed-patient assignment decision support tools that take interdependencies between units into account.

Jweinat et al. (2013) recognize "inter/intradepartmental and interdisciplinary collaboration" as one of the five key components of capacity management in their paper, in which they summarize a one-year hospital-wide effort to improve hospital patient flow. All the aforementioned findings highlight the need for more research on the effect of interdependencies between hospital units, so that practitioners and researchers can be better equipped with insights to effectively understand and improve hospital operations. Our findings in this paper improve the understanding of the interdependencies between hospital care units by examining whether and how much congestion in one inpatient unit spills over to other inpatient units.

2.2. Estimation of Network Effects in Other Application Settings

The need to understand and estimate network effects has been emerging in other application settings in operations management. For example, businesses are often part of a supply chain network through which information and materials flow across organizations. As such,

understanding the network effects is crucial when designing and maintaining supply chain networks. Serpa and Krishnan (2018) show that a firm's productivity is affected by its customers and that this productivity spillover depends on the structure of the supply chain network. Osadchiy et al. (2016) examine how the supply network structure mediates the effect of the state of the economy on firm sales, and Wang et al. (2021) quantify the degree of financial risk propagation in a supply chain network. Other settings in which network effects have been estimated include the propagation of financial shocks in a network of financial institutions (Acemoglu et al. 2015), spillover effects on customer experience in a network of restaurants (Yu et al. 2017, Mankad et al. 2019), and spillover effects on customer demand in a network of bike share stations (He et al. 2021). To the best of our knowledge, we are the first in the literature to estimate network spillover effects in a hospital inpatient unit network setting.

2.3. Methods to Estimate Network Effects

The difficulty of estimating causal network effects using observational data was first introduced by Manski (1993). Researchers have developed various methods to correct for the bias in estimating causal network effects due to various factors such as contextual and correlated effects (Manski 1993, Bramoullé et al. 2009), simultaneity (Van den Bulte and Lilien 2001, Godes and Mayzlin 2004), and homophily (Aral et al. 2009). The method we adopt is based on Bramoullé et al. (2009) and Drakopoulos and Zheng (2017), who construct instrumental variables by exploiting the variation in the network weight matrix. The key difference between our method and those based on the spatial econometrics literature (Anselin 2001, Mankad et al. 2019) is that all covariates in our model can be potentially correlated through the network, which is not allowed in their models.

3. Setting and Data

We use data from a high-volume academic hospital in a large metropolitan area of the United States. We collected data for every patient who received care from January 2014 to December 2018. For each hospitalization, the data include all the units the patient visited, including the operating room and inpatient units, along with the time stamps for when the patient entered and exited the unit. The data also include patient-level information such as admission type and surgery-level information such as the surgeon in charge.

3.1. Inpatient Units

Inpatient units at a hospital can be broadly divided into four groups according to their nurse-to-patient ratios, which reflect the treatment and monitoring levels. Generally, the ICUs have a nurse-to-patient ratio of 1:1 to 1:2; step-down units have a ratio of 1:3; telemetry units have a ratio of 1:4; and medical-surgical units have a ratio of 1:5 (Coffman et al. 2002). Our study hospital has 16 inpatient units, 7 of which are ICUs; 1 is a step-down unit; 7 are telemetry units; and 1 is a medical-surgical unit. The seven ICUs are specialized ICUs and hence are not perfect substitutes for each other. For instance, one is a cardiothoracic surgical ICU, whereas another is a pulmonary medical ICU. Similarly, each of the seven telemetry units has its specialty and is not a perfect substitute for any other. The unit sizes vary from 8 to 34 beds. Table 1 presents each inpatient unit's level of care, size, and median patient length of stay. During our study period, there were no major changes in unit sizes and staffing ratios.

Table 1. Summary of the Inpatient Units at Our Study Hospital

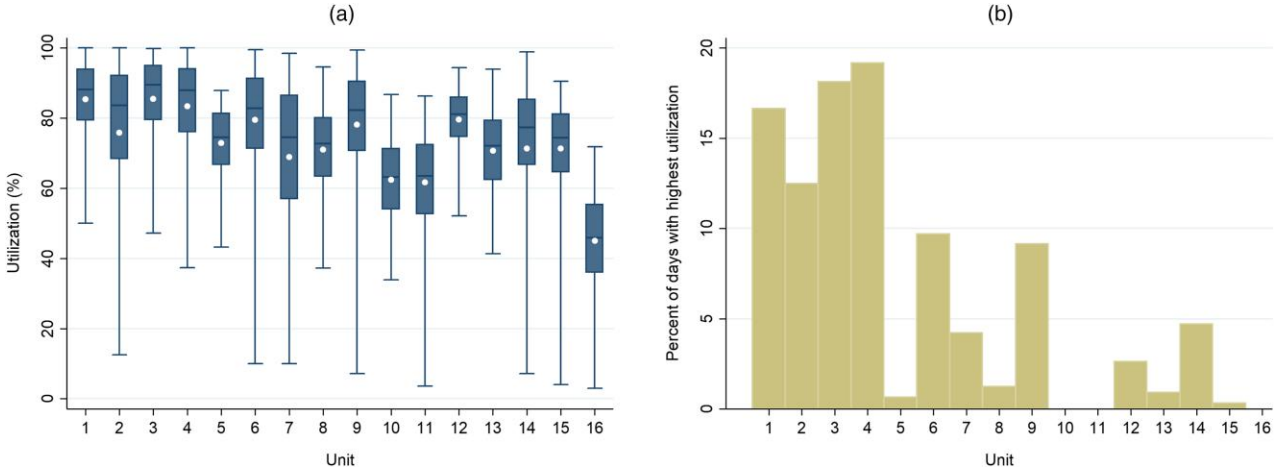
Unit	Level of care	Size (number of beds)	Median length of stay (days)
1	Intensive care	18	2.7
2	Intensive care	8	1.9
3	Intensive care	10	1.9
4	Intensive care	10	2.5
5	Intensive care	18	1.7
6	Intensive care	10	1.8
7	Intensive care	10	1.7
8	Step-down	20	2.8
9	Telemetry	14	2.9
10	Telemetry	34	2.9
11	Telemetry	28	2.6
12	Telemetry	31	3.9
13	Telemetry	34	2.3
14	Telemetry	14	3.7
15	Telemetry	25	3.2
16	Medical-surgical	34	2.0

3.2. Utilization Measures

In this paper, we examine whether and how much congestion in one inpatient unit spills over to other inpatient units. Therefore, our dependent variable is $Util_{it+1}$, which represents unit i's daily time-weighted measure of occupancy divided by the number of beds in the unit (hereafter referred to as utilization) in day $t+1$. Figure 1(a) summarizes the utilization of each unit during the five-year sample period: from top to bottom, it shows the 99th percentile, the 75th percentile, the median, the mean (white dot), the 25th percentile, and the 1st percentile. We observe substantial variation in the utilization across different units (the mean utilization ranges from 43.7% to 85.0%), as well as within units (the standard deviation of utilization for each unit ranges from 9.0% to 23.4%).

A hospital administrator may attempt to identify the bottleneck unit by examining the utilization levels of the inpatient units. For example, if there is an inpatient unit whose utilization level clearly dominates the utilization

Figure 1. (Color online) Utilization of the Inpatient Units



Notes. (a) Distribution of utilization in each inpatient unit. (b) Percent of days each unit has the highest utilization among all 16 inpatient units. Utilization is measured as a time-weighted utilization for each day. The statistics are for the five-year sample period, 1,826 days from 2014 to 2018. (a) Distributions of utilization for each unit are shown using boxplots. Each boxplot provides a summary of six statistics: (from top to bottom) 99th percentile, 75th percentile, median, mean (white dot), 25th percentile, and 1st percentile.

levels of other units, that unit may be the bottleneck unit. However, Figure 1(a) shows that it is difficult to pinpoint one unit with the highest utilization given the variation of utilization over time. In addition, the percent of days on which each unit has the highest utilization, illustrated in Figure 1(b), shows that there is no one unit whose share exceeds 20%. Except for three units (units 10, 11, and 16), there was at least one day in the five-year sample period when each unit had the highest utilization. Perhaps most importantly, when inpatient units have high utilization, patients' care pathways are often altered (see the discussion in Section 2), which makes the realized demand for each unit endogenous to the hospital's congestion level and different from the true demand (i.e., true patient needs) for each unit. This means that the unit with the highest realized utilization may *not* necessarily be the bottleneck unit. Moreover, even if one could estimate the true patient needs, traditional bottleneck analysis is likely to fail because it does not account for the complex nature of the care pathways and patient arrival patterns. Thus, one can attempt to identify the bottleneck unit through simulations, but those are also likely to fail because patient routing decisions vary with the congestion level and over time. In other words, it is difficult to fully take into account the complex patient routing decisions to simulate the traffic in the inpatient unit network and identify the bottleneck.

These observations suggest that it is challenging for hospital administrators to identify the bottleneck unit directly from the data. They also highlight the need to empirically estimate the congestion spillover effects across inpatient units and identify the bottleneck unit as we do in this paper.

3.3. Patient Transfers Between Inpatient Units and Network Matrix *G*

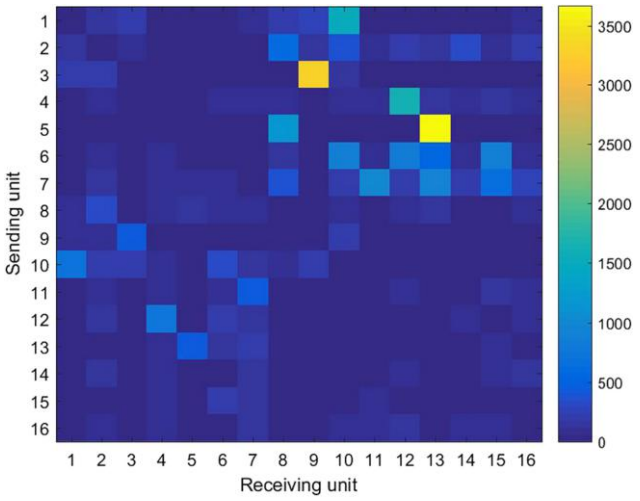
During each hospitalization, patients get transferred between different inpatient units as their conditions and needs change. For example, if the condition of a patient in an ICU improves, then the patient may be transferred to a telemetry unit and receive a lower level of care before being discharged. On the other hand, if the condition of a patient in a medical-surgical unit deteriorates, then the patient may be transferred to an ICU to receive a higher level of care. It is also possible for a patient to be transferred to a unit that has the same level of care as the originating unit. Moreover, if some units in the inpatient network are close to full capacity, then off-service placement might occur, and the destination units of some patient transfers might be affected.

During the five-year sample period, there were a total of 36,461 transfers between the 16 inpatients units at our study hospital, which translates to about 20 transfers a day. Figure 2 illustrates the number of patient transfers between each pair of the 16 inpatient units. It shows that the majority of the transfers occurred between units with different levels of care—for instance, from ICUs (units 1–7) to non-ICUs (units 8–16) and vice versa. The highest number of transfers occurred between unit 5—an ICU—and unit 13—a telemetry unit.

To estimate the magnitude of congestion spillover, we need to compute the interaction intensities between different pairs of inpatient units. We define a network matrix G, where the interaction intensity between unit i and unit j is calculated as

$$G_{ij} = \frac{N_{ij} + N_{ji}}{\sum_k (N_{ik} + N_{ki})}. \quad (1)$$

Figure 2. (Color online) Number of Patient Transfers Between the 16 Inpatient Units at Our Study Hospital During the Five-Year Sample Period



Here, N_{ij} is the total number of patient transfers from unit i to unit j in a given period (by definition, $N_{ii} = 0$ for all i). That is, G_{ij} corresponds to the interaction intensity between unit i and unit j—from i to j as well as from j to i—relative to all interactions that unit i had with other units.

To identify the bottleneck unit, we need to know the *true* demand for each unit. Hence, it is important that our network matrix G reflects true patient needs. That is, we need G to represent how patients will flow through different inpatient units *when all inpatient units have ample capacity*. In such a setting, patients' route and length of stay will depend solely on their needs.

To ensure that G represents the interaction intensity between units determined by patient needs, we impose two restrictions on how we compute N_{ij}. First, N_{ij} may be endogenous to the hospital's congestion level. For example, if unit j is highly congested or temporarily closed, a potential patient transfer from unit i to unit j may be blocked (Kim et al. 2015). As another example, if unit i needs to transfer its patient to unit k, but the transfer is not possible because of high congestion in unit k, unit i may transfer the patent to unit j instead (Dong et al. 2019, Song et al. 2020). To capture the interactions that represent patient needs and are not affected by the congestion level in the hospital, we compute N_{ij} using data from nonbusy days only, where we define a day to be nonbusy if the daily hospital utilization is among the bottom 80% of each year—that is, if the daily hospital utilization is lower than the 80th percentile value of each year's daily hospital utilization values.[2] We also do not consider days on which any of the inpatient units is closed.

Second, the interaction intensity between each pair of units may change over time as care providers, hospital services, and patient needs evolve and change. To address this potential change in the interaction intensity over time, we construct G for each year. For example, Table B1 in the online appendix shows our network matrix G constructed using the patient movement data from days that were not busy and had no closed unit in year 2017. Note that we construct G at the yearly level and not at a higher frequency—for example, monthly or weekly level—because we are interested in estimating the long-term average congestion spillover effect, which reflects the long-term average patient needs instead of the effect measured at a higher frequency. This is consistent with the objective of the paper, which is to help hospitals make long-term capacity-planning decisions.

3.4. Control Variables

We now describe other variables that are included in our empirical analysis to control for their potential effects on our dependent variable $Util_{it+1}$.

3.4.1. Previous Days' Utilization. Because today's utilization of unit i, $Util_{it}$, is likely to affect tomorrow's utilization of unit i, $Util_{it+1}$, we control for $Util_{it}$. Furthermore, Table 1 shows that the median lengths-of-stay in the inpatient units are two to three days. This suggests that tomorrow's utilization $Util_{it+1}$ may be affected by not just today's utilization, $Util_{it}$, but also by yesterday's utilization, $Util_{it-1}$. Hence, we also control for $Util_{it-1}$.

3.4.2. Patient Arrivals from Outside. In addition to arrivals from other inpatient units, each unit has arrivals from outside of the inpatient unit network. At the study hospital, patient arrivals from outside can be broadly divided into four types: (1) scheduled admissions who get admitted to an inpatient unit after a scheduled surgery, which we call scheduled admissions; (2) transfers from other hospitals, which we call transfer admissions; (3) direct admissions from outpatient providers and other sources, which we call direct admissions; and (4) elective admissions without scheduled surgery, which we call elective admissions.

We let $SchedAdm_{it}$ denote the expected number of scheduled admissions to unit i on day t. To avoid endogeneity issues, we use the *expected* number of admissions rather than the *realized* number of admissions, which may be endogenous to the hospital's congestion level. To schedule surgery at our study hospital, surgeons need to submit electronic booking slips with surgery-related information, such as the surgeon in charge, the operating room, and the inpatient unit the patient is expected to go after surgery. Surgeons typically schedule surgery *several weeks in advance*. Thus, patients' expected postsurgery units reported in electronic booking slips are determined solely by patients' needs and surgeons' preferences and are exogenous to the hospital's congestion level.

Table 2. Summary Statistics of Control Variables by Unit

Unit	(1) *SchedAdm*	(2) *TransferAdm*	(3) *DirectAdm*	(4) *ElectiveAdm*	(5) *PatientMix*	(6) *Dsc*
1	0.00 (0.00)	3.49 (1.38)	0.73 (0.53)	0.46 (0.40)	76.17 (11.89)	1.97 (0.62)
2	0.00 (0.00)	3.37 (1.39)	1.80 (1.39)	0.72 (0.64)	65.37 (23.15)	2.74 (0.87)
3	6.67 (10.20)	4.39 (1.75)	0.94 (0.88)	0.21 (0.19)	60.52 (19.76)	1.10 (0.37)
4	0.00 (0.00)	2.85 (1.08)	0.81 (0.57)	0.27 (0.28)	62.42 (17.93)	1.84 (0.57)
5	6.10 (6.97)	8.55 (3.27)	0.99 (0.88)	0.56 (0.50)	65.65 (13.77)	2.49 (0.86)
6	9.12 (11.30)	3.46 (1.41)	1.11 (0.96)	0.23 (0.21)	59.78 (19.07)	1.79 (0.56)
7	0.00 (0.00)	3.53 (1.59)	1.50 (1.10)	0.47 (0.40)	52.80 (24.58)	1.91 (0.70)
8	3.45 (4.93)	1.45 (0.58)	1.53 (0.99)	0.50 (0.42)	56.47 (14.95)	10.20 (3.21)
9	0.04 (0.56)	2.04 (0.81)	1.36 (0.85)	0.66 (0.56)	38.22 (18.73)	14.49 (4.56)
10	2.09 (2.98)	2.16 (0.87)	2.80 (1.67)	1.44 (1.24)	48.42 (14.27)	9.41 (2.91)
11	10.23 (8.59)	2.02 (0.81)	2.90 (1.89)	0.76 (0.64)	36.66 (17.80)	17.50 (5.43)
12	0.13 (0.85)	2.61 (1.03)	3.75 (2.26)	1.45 (1.22)	42.84 (11.38)	11.78 (3.62)
13	5.56 (5.57)	2.10 (0.89)	1.67 (1.11)	1.22 (1.05)	49.99 (12.31)	16.40 (5.07)
14	0.00 (0.00)	2.33 (0.94)	4.39 (2.61)	1.68 (1.41)	48.98 (23.86)	9.50 (3.00)
15	8.33 (8.31)	1.45 (0.58)	2.00 (1.30)	0.49 (0.47)	44.92 (14.60)	13.42 (4.18)
16	9.79 (9.67)	0.97 (0.38)	2.42 (1.71)	0.65 (0.56)	49.33 (24.15)	12.83 (3.99)

Notes. Mean and standard deviations (in parentheses) are shown. Because we use patient movements from the previous years to construct *TransferAdm*, *DirectAdm*, *ElectiveAdm*, and *Dsc*, they cannot be computed for the first year of our data period. Hence, the dates in 2014 are removed, which leaves $N = 1{,}461$ days. *SchedAdm*, *TransferAdm*, *DirectAdm*, *ElectiveAdm*, and *Dsc* are normalized by dividing them by the respective unit sizes and multiplying by 100 to use in our estimation models.

Because of data limitation, we are able to extract the expected unit after surgery from the electronic booking slips for only a subset of our data period: of the 25,612 scheduled admissions[3] during the data period, we have the expected unit data for 45% of them. To overcome this data limitation, we use linear regression to predict the expected postsurgery unit using surgeon identifier and operating room information. For our in-sample data (i.e., surgery that had expected unit information), our prediction method correctly predicted the expected unit for 90% of the cases. The expected unit after surgery may not necessarily match the unit the patient is actually sent to after surgery because patient needs may change or because there may be congestion in the desired unit on the day of admission. For our in-sample data, the predicted unit matched the actual unit 77% of the time. For our out-of-sample data, the predicted unit matched the actual unit 72% of the time, which supports a reasonable performance of our prediction method. Hence, we use our prediction method to predict the expected unit for each surgery on day t and constructed $SchedAdm_{it}$ using the prediction result.

We let $TransferAdm_{it}$ denote the expected number of transfer admissions to unit i on day t. To construct this variable, for each year in our data period, we first compute the percentage of hospital-wide transfer admissions that were sent to unit i using data from nonbusy days with no closed units only (see Section 3.3 for the definition of a nonbusy day). We then compute $TransferAdm_{it}$ by multiplying the number of hospital-wide transfer admissions on day t by the previous year's percentage of transfer admissions sent to unit i. We note that $TransferAdm_{it}$ is exogenous to the hospital's congestion level due to the stochastic nature of the transfer patient arrivals, as well as to its construction based on the expected allocation to different units using the previous year's data.

We let $DirectAdm_{it}$ denote the expected number of unscheduled admissions to unit i on day t. It is similarly defined as $TransferAdm_{it}$. Similar to $TransferAdm_{it}$, $DirectAdm_{it}$ is exogenous to the hospital's congestion level due to the stochastic nature of the direct patient arrivals, as well as to its construction based on the expected allocation to different units using the previous year's data.

Last, we let $ElectiveAdm_{it}$ denote the expected number of elective admissions to unit i on day t. It is similarly defined as $TransferAdm_{it}$ and $DirectAdm_{it}$. However, elective admissions may be endogenous to the hospital's congestion level because physicians may decide whether to admit an elective patient depending on the congestion level. This differs from scheduled admissions that are scheduled weeks in advance and transfer and direct admissions whose arrivals are random and come from urgent patient needs that are not likely to be influenced by the hospital's congestion level.

Columns (1)–(4) of Table 2 provide the means and standard deviations of $SchedAdm_{it}$, $TransferAdm_{it}$, $DirectAdm_{it}$, and $ElectiveAdm_{it}$ by unit.

3.4.3. Patient Mix. We let $PatientMix_{it}$ denote the percentage of patients whose length of stay in unit i has exceeded the median length of stay (of unit i) at 11 p.m. on day t (see Table 1 for the median length of stay in each unit). If many patients in the unit have stayed longer than the median length of stay on day t, it is likely that more patients will be ready to be discharged on day $t+1$, lowering the utilization of unit i on day $t+1$.

Column (5) of Table 2 provides the means and standard deviations of $PatientMix_{it}$ by unit.

3.4.4. Discharges to Outside of the Inpatient Unit Network. We let Dsc_{it} denote the expected number of discharges from unit i to outside of the hospital inpatient unit network on day t. We first compute the percentage of hospital-wide discharges to outside from unit i using data from nonbusy days only (see Section 3.3 for the definition of a nonbusy day) for each year in our data period. We then compute Dsc_{it} by multiplying the number of hospital-wide discharges on day t by last year's percentage of discharges from unit i. We note that Dsc_{it} may be endogenous to the hospital's congestion level; because the congestion level can influence the timing of hospital discharges (Chan et al. 2012, KC and Terwiesch 2012), the number of discharges from the hospital are potentially endogenous to the hospital's congestion level. Column (6) of Table 2 provides the means and standard deviations of Dsc_{it} by unit.

3.4.5. Unit and Time-Related Controls. We include unit fixed effects. We also include day-of-week and year-month fixed effects to control for the variation in utilization by day-of-week, month, and year.

4. Model and Estimation

In this section, we first introduce our model in Section 4.1. We then discuss the identification challenges in Section 4.2 and the identification strategy in Section 4.3. We explain our instrumental variables in detail in Section 4.4, and we illustrate how we slice the network matrix G to study potentially heterogeneous congestion spillover effects between different types of inpatient unit pairs in Section 4.5. Last, we discuss how we select our instrumental variables with least absolute shrinkage and selection operator selection in Section 4.6.

4.1. Model

We model unit i's utilization in day $t+1$, $Util_{it+1}$, as follows:

$$Util_{it+1} = \alpha Util_{it-1} + \beta_1 Util_{it} + \beta_2 G_i Util_t + \theta_1 Z_{it} + \theta_2 G_i Z_t + \gamma_1 W_{it} + \gamma_2 G_i W_t + c_i + \rho_t + \nu_{it}. \quad (2)$$

Here, $Util_{it-1}$ is unit i's utilization on day $t-1$, and $Util_{it}$ is unit i's utilization on day t. G_i is the exogenous network vector indicating the expected patient flows between unit i and the other units (i.e., G_i is the ith row of the network matrix G). As described in Section 3.3, we construct G for each year to address the potential change in the expected patient flows between units over time. Furthermore, to ensure that the expected patient flows between units are not influenced by the hospital's congestion level on day t, we use the previous year's G to compute the network terms for our variables for day t.[4] For instance, to compute $G_i Util_t$, if t is in year 2018, we use G constructed using the data from nonbusy days with no closed units in year 2017. $Util_t$ is a 16 by 1 vector of the unit-specific utilization $Util_{it}$ for all units i. β_2 captures the congestion spillover effect we are interested in estimating.

Section 3.4 provides a detailed description of the control variables we include in our empirical analysis. We divide the control variables into two groups, Z_{it} and W_{it}, according to their type. We use Z_{it} to denote a vector of *exogenous covariates* that affect the utilization of unit i on day $t+1$ directly and indirectly through the network G_i. Z_{it} includes $SchedAdm_{it}$, $TransferAdm_{it}$, and $DirectAdm_{it}$. We let W_{it} denote a vector of other covariates that affect the utilization of unit i on day $t+1$ directly and indirectly through the network G_i. W_{it} includes $ElectiveAdm_{it}$, $PatientMix_{it}$, and Dsc_{it}. Last, c_i is the unit i fixed effect, and ρ_t is the time fixed effect of day t and includes day-of-week fixed effects and year-month fixed effects. ν_{it} is the unobserved determinant of unit i's utilization.

Following the classic panel data literature, we take the difference between $Util_{it+1}$ and $Util_{it}$ to remove potential time trends in the data. Equation (2) becomes

$$\Delta Util_{it+1} = \alpha \Delta Util_{it-1} + \beta_1 \Delta Util_{it} + \beta_2 G_i \Delta Util_t + \theta_1 \Delta Z_{it} + \theta_2 G_i \Delta Z_t + \gamma_1 \Delta W_{it} + \gamma_2 G_i \Delta W_t + \omega_t + \varepsilon_{it}, \quad (3)$$

where Δ denotes the time difference for each variable; ω_t is the difference between ρ_t and ρ_{t-1}; and ε_{it} is the difference between ν_{it} and ν_{it-1}.

4.2. Identification Challenges

The right-hand side of Equation (2) contains $G_i Util_t$, as well as $G_i Z_t$ and $G_i W_t$. In other words, our model allows for both the other units' utilization to affect unit i's utilization through the network G and the determinants of utilization, Z_{it} and W_{it}, to be correlated through the network G. This is the main difference between our model, which fits the literature on causal network effect estimation in economics and operations management (Bramoullé et al. 2009, Drakopoulos and Zheng 2017) and the models in the spatial econometrics literature (Anselin 2001). This difference makes the identification and estimation more complicated than the models in Anselin (2001) and the related applications (Mankad et al. 2019).

Directly estimating Equation (2) leads to biased estimates due to the reflection problem in the identification of the network effect (Manski 1993). In our setting, the reflection problem can be interpreted as follows. First, if units i and j are connected through the network G, there are patients transferred between the two units. As a result, any unobserved considerations in the patient transfer decisions between i and j can be correlated with

the utilization of unit i and unit j. These unobservables, captured by v_{it} in Equation (2), introduce bias in the estimate of β_2. Second, because i and j are connected through the network G, the determinants of the utilization of the two units—including determinants of the patient arrivals and discharges of the two units—can also be correlated. In other words, it is challenging to separately identify β_2 from θ_2 and γ_2, which leads to additional bias when estimating Equation (2) directly.

4.3. Identification Strategy

We use the instrumental variable (IV) approach to account for the endogeneity of $Util_{it-1}$, $Util_{it}$, and G_iUtil_t and estimate the congestion spillover effect β_2 without bias. Our network matrix G measures the expected amount of patient flows between each pair of units and, hence, is a weighted exogenous graph. Intuitively, the heterogeneity in weights across different links in the graph provides the main source of identification to our model.

The IVs we use are G^2Z_{t-1} and G^2Z_{t-2}, which are the network-lagged and time-lagged measurements of the exogenous covariates. G_i^2 is the ith row of G^2, and Z_{it} is defined as a vector of exogenous covariates that affect $Util_{it+1}$ directly and indirectly through the network G. They are valid IVs under two conditions (see proposition 3 on p. 16 in Drakopoulos and Zheng (2017) for a detailed argument): (1) Z_{it} are exogenous—that is, $E[\varepsilon|Z_{it}] = 0$; and (2) the matrices I (identity matrix), G, and G^2 are linearly independent. If the two conditions hold, $G_i^2Z_{t-1}$ measures the indirect effect of unit i's neighbor's exogenous covariates Z_{t-1} on $Util_{it+1}$ through the network G.

A concrete example is as follows. Suppose that we are interested in measuring how the utilization of an ICU—unit i—in period $t+1$ (i.e., $Util_{it+1}$) is affected by the utilization of a non-ICU unit—unit j—in period t (i.e., $Util_{jt}$) through the patient movements between the two units. Because the ICU and the non-ICU unit (units i and j) are connected through the network G, the observed and unobserved determinants of their utilization are likely to be correlated through G as well. As a result, we look for IVs, which introduce exogenous variations to the determinants of the utilization in unit i. Suppose that there is another non-ICU unit (unit k) that has patient transfers with unit j (and potentially with unit i). Let Z_{kt-1} represent the exogenous patient arrivals to unit k in period $t-1$ (i.e., Z_{kt-1} is an exogenous determinant of unit k's utilization). Then, $G_i^2Z_{t-1}$ can be our instrumental variable. The kth element of $G_i^2Z_{t-1}$, $G_{ik}^2Z_{kt-1}$, captures the impact of exogenous patient arrivals to unit k on the utilization of the ICU (unit i) through its impact on the utilization of unit j. Because I, G, and G^2 are linearly independent, the effect (the impact of exogenous patient arrivals to unit k on the utilization of unit i through its impact on the utilization of unit j) can be separately identified from any direct correlation between the observed and unobserved determinants of the utilization of the ICU (unit i) and the non-ICU unit (unit j), which causes the endogeneity issue in the estimation. Similar arguments can be made for G^2Z_{t-2}.

4.4. Constructing Our IVs

Recall that we use two sets of IVs, G^2Z_{t-1} and G^2Z_{t-2}, and that Z_t includes $SchedAdm_t$, $TransferAdm_t$, and $DirectAdm_t$. Thus, our IVs are $G_i^2SchedAdm_{t-1}$, $G_i^2TransferAdm_{t-1}$, and $G_i^2DirectAdm_{t-1}$, $G_i^2SchedAdm_{t-2}$, $G_i^2TransferAdm_{t-2}$, and $G_i^2DirectAdm_{t-2}$.

As mentioned previously, our IVs need to satisfy two conditions. First, $SchedAdm_{it}$, $TransferAdm_{it}$, and $DirectAdm_{it}$ need to be exogenous. As described in Section 3.4, because surgeons typically plan scheduled surgery several weeks in advance, the number of scheduled admissions is unlikely to be correlated with any unobserved determinants of the inpatient units' realized utilization in the focal time period. Furthermore, we use patients' expected postsurgery units reported in the electronic booking slips (which are filled out when surgeons schedule a surgery) to derive $SchedAdm_{it}$. Because these expected units are determined solely by patients' needs and surgeons' preferences in advance, they are exogenous to the hospital's congestion level in the focal time period.

$TransferAdm_{it}$ and $DirectAdm_{it}$ are also unlikely to be affected by the inpatient units' utilization due to the stochastic nature of the transfer and direct patient arrivals, as well as to their construction based on the expected number using the previous year's pattern of patient allocation. Our study hospital does not have an emergency room. When our framework is applied to a hospital with an emergency room, arrivals from the ER can be treated as transfer/direct admissions because they are also close to random and are not likely to be affected by the hospital's congestion level.[5]

Second, the network matrices I, G, and G^2 need to be linearly independent. In other words, there needs to be enough variation across the weights on the links of the network. We test this assumption directly and find that it is satisfied. Section B in the online appendix provides the test details. Also, as discussed in detail in Sections 3.3 and 4.1, the construction of matrix G ensures that G reflects the true patient needs and is exogenous to the realized congestion level in the inpatient units of the hospital.

4.5. Slicing the Network Matrix *G*

Estimating β_2 in Equation (2) gives the average congestion spillover effect across different pairs of units, regardless of the types of interactions between each pair. However, the magnitude of the congestion spillover effect may depend on the type of interactions. For example, for an ICU, the congestion spillover from another

ICU may be different from the congestion spillover from a non-ICU, even if the respective interaction intensities are the same. Motivated by this observation, we divide our inpatient units into two groups, ICUs and non-ICUs, because the level of care provided in ICUs is much higher than that provided in non-ICUs. We then examine whether the congestion spillover effect from units that have different levels of care (i.e., between an ICU and a non-ICU) is different from the congestion spillover effect from units that have the same level of care (i.e., between an ICU and another ICU, as well as between a non-ICU and another non-ICU) by estimating the following model:

$$\begin{aligned} &Util_{it+1} \\ &= \alpha Util_{it-1} + \beta_1 Util_{it} + \beta_2 G_i^{DiffLevel} Util_t + \beta_3 G_i^{SameLevel} Util_t \\ &\quad + \theta_1 Z_{it} + \theta_2 G_i^{DiffLevel} Z_t + \theta_3 G_i^{SameLevel} Z_t \\ &\quad + \gamma_1 W_{it} + \gamma_2 G_i^{DiffLevel} W_t + \gamma_3 G_i^{SameLevel} W_t \\ &\quad + c_i + \rho_t + \nu_{it}. \end{aligned} \tag{4}$$

Here, the network matrix G is sliced into two matrices, $G^{DiffLevel}$ and $G^{SameLevel}$:

$$G_{ij}^{DiffLevel} = \begin{cases} 0, & \text{if } \mathbb{1}_{i=ICU} = \mathbb{1}_{j=ICU} \\ G_{ij}, & \text{otherwise} \end{cases} \quad \text{and}$$

$$G_{ij}^{SameLevel} = \begin{cases} G_{ij}, & \text{if } \mathbb{1}_{i=ICU} = \mathbb{1}_{j=ICU} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathbb{1}_{i=ICU}$ is equal to one if unit i is an ICU and zero otherwise. Then, β_2 captures the congestion spillover effect between units that have different levels of care, and β_3 captures the congestion spillover effect between units that have the same level of care.

Furthermore, because the magnitude of congestion spillover from an ICU to an ICU may differ from that from a non-ICU to a non-ICU, we estimate the following model:

$$\begin{aligned} &Util_{it+1} \\ &= \alpha Util_{it-1} + \beta_1 Util_{it} + \beta_2 G_i^{DiffLevel} Util_t \\ &\quad + \beta_3 G_i^{SameLevel-ICU} Util_t + \beta_4 G_i^{SameLevel-NonICU} Util_t \\ &\quad + \theta_1 Z_{it} + \theta_2 G_i^{DiffLevel} Z_t + \theta_3 G_i^{SameLevel-ICU} Z_t \\ &\quad + \theta_4 G_i^{SameLevel-NonICU} Z_t \\ &\quad + \gamma_1 W_{it} + \gamma_2 G_i^{DiffLevel} W_t + \gamma_3 G_i^{SameLevel-ICU} W_t \\ &\quad + \gamma_4 G_i^{SameLevel-NonICU} W_t \\ &\quad + c_i + \rho_t + \nu_{it}. \end{aligned} \tag{6}$$

Here, $G^{SameLevel}$ is further sliced into two matrices, $G^{SameLevel-ICU}$ and $G^{SameLevel-NonICU}$:

$$G_{ij}^{SameLevel-ICU} = \begin{cases} G_{ij}, & \text{if } i = j = ICU \\ 0, & \text{otherwise} \end{cases} \quad \text{and}$$

$$G_{ij}^{SameLevel-NonICU} = \begin{cases} G_{ij}, & \text{if } i = j = Non-ICU \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In Equation (6), β_2 captures the congestion spillover effect between ICUs and non-ICUs; β_3 captures the congestion spillover effect among ICUs; and β_4 captures the congestion spillover effect among non-ICUs.

4.6. Selecting Instrumental Variables with LASSO Regression

We have a large set of potential IVs: six IVs—G^2Z_{t-1} and G^2Z_{t-2}—for the model in Equation (2); 12 IVs—$(G^{DiffLevel})^2 Z_{t-1}$, $(G^{DiffLevel})^2 Z_{t-2}$, $(G^{SameLevel})^2 Z_{t-1}$, and $(G^{SameLevel})^2 Z_{t-2}$—for the model in Equation (4); and 18 IVs—$(G^{DiffLevel})^2 Z_{t-1}$, $(G^{DiffLevel})^2 Z_{t-2}$, $(G^{SameLevel-ICU})^2 Z_{t-1}$, $(G^{SameLevel-ICU})^2 Z_{t-2}$, $(G^{SameLevel-NonICU})^2 Z_{t-1}$, and $(G^{SameLevel-NonICU})^2 Z_{t-2}$—for the model in Equation (6). We use the recent tools in econometrics to select the optimal IVs through LASSO regressions (Belloni et al. 2012, Chernozhukov et al. 2015). For the postselection inference to be correct and to maintain the causal interpretation of the coefficient estimates, in the second stage of the estimation, we need to take into account the selection procedure in the first stage when computing the standard errors of the estimated coefficients. We follow Belloni et al. (2012, 2014) to select the IVs and derive the correct standard errors in the estimation. We implement this approach using the *ivlasso* command in STATA (Ahrens et al. 2018).

5. Results

In this section, we present our main estimation results of identifying the congestion spillover effect in Section 5.1 and conduct several robustness checks in Section 5.2.

5.1. Estimated Effect of Congestion Spillover

Table 3 presents the estimation results obtained by two-stage least squares regressions. The network matrices used for the columns are G for column (1), $G^{DiffLevel}$ and $G^{SameLevel}$ for column (2), and $G^{DiffLevel}$, $G^{SameLevel-ICU}$, and $G^{SameLevel-NonICU}$ for column (3). The first-stage F statistics for the significance of the instruments for our endogenous variables are all well over the commonly used benchmark of 10, which indicates that our instruments are quite strong. The p values from the Hansen J tests are also provided. The Hansen J test examines overidentifying restrictions when robust standard errors are estimated; the null hypothesis is that the instruments are valid instruments. All the p values are greater than 0.05,

Table 3. Effect of Congestion Spillover

	(1)	(2)	(3)
$\Delta Util_{it-1}$	−0.075	−0.007	−0.020
	(0.048)	(0.045)	(0.044)
$\Delta Util_{it}$	−0.082	−0.028	−0.058
	(0.087)	(0.065)	(0.074)
$G_i\Delta Util_{it}$	0.571***		
	(0.118)		
$G_i^{DiffLevel}\Delta Util_{it}$		0.400***	0.448***
		(0.094)	(0.097)
$G_i^{SameLevel}\Delta Util_{it}$		1.310***	
		(0.310)	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			1.392**
			(0.469)
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			1.413***
			(0.302)
N	23,312	23,312	23,312
R^2	0.188	0.229	0.216
Hansen J statistic p value	0.529	0.340	0.088
First-stage F statistics			
$\Delta Util_{it-1}$	87.844	43.012	30.679
$\Delta Util_{it}$	86.319	41.696	30.780
$G_i\Delta Util_{it}$	131.167		
$G_i^{DiffLevel}\Delta Util_{it}$		54.416	44.417
$G_i^{SameLevel}\Delta Util_{it}$		94.848	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			109.382
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			157.101

Notes. $N = 23{,}312$ unit-days. Robust standard errors in parentheses.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

and hence we fail to reject the null hypotheses that our instruments are valid.

The estimation results show that unit i's neighbors' utilization today have statistically significant and positive effects on unit i's utilization tomorrow, regardless of the type of network matrix used. For example, column (1) of Table 3 shows that a one-unit increase in G_iUtil_t increases $Util_{it+1}$ by 0.571 ($p < 0.001$). One way to interpret the result is to examine the effect of increasing the utilization of all inpatient units, except for the focal unit, by x percentage points. Using the coefficient value 0.571, the effect can be written as $0.571\sum_j G_{ij}x$ for focal unit i ($G_{ii} = 0$). Because of the way we construct G (see Equation (1)), $\sum_j G_{ij} = 1$ for all unit i, which means that the effect can be written as $0.571x$. That is, if the utilization levels of all inpatient units, except for unit i, increase by ten percentage points today, unit i's utilization will increase by 5.71 percentage points tomorrow.

In addition, we can examine the congestion spillover effect between each inpatient unit pair. For instance, the effect of increasing unit j's utilization by 10 percentage points today on unit i's utilization tomorrow can be written as $0.571 \times G_{ij} \times 10$. We compute this value for all 240 inpatient unit pairs (i, j) where $i \neq j$ using the network matrix G constructed from the patient movement data in year 2017 (see Table B1 in the online appendix). The biggest pairwise congestion spillover effect is observed when $G_{i,j} = G_{3,9}$: If unit 9's utilization increases by 10 percentage points today, it is predicted that unit 3's utilization will increase by 4.328 percentage points tomorrow. The smallest pairwise congestion spillover effect is zero because some inpatient pairs had no patient transfers between them. The average pairwise congestion spillover effect is 0.381 percentage points with a standard deviation of 0.673.

Column (2) of Table 3 shows the congestion spillover effect between units that have different levels of care and between units that have the same level of care are both positive and statistically significant. We find that, although the magnitude of the coefficients is different, the actual predicted average effects are similar. For example, increasing unit j's utilization by 10 percentage points today leads to an average pairwise congestion spillover of 0.402 percentage points (maximum, 3.031; standard deviation, 0.612) to units whose level of care differs from that of unit j and an average pairwise congestion spillover of 0.382 percentage points (maximum, 2.193; standard deviation, 0.351) to units whose level of care is the same as that of unit j.

Probing the congestion spillover effect from units with the same level of care deeper, we again find that the congestion spillover effect among ICUs and the congestion spillover effect among non-ICUs are both positive and statistically significant. We find that the predicted average effects are slightly larger for that among non-ICUs. For example, if unit j is an ICU, increasing unit j's utilization by 10 percentage points today leads to an average pairwise congestion spillover of 0.380 percentage points (maximum, 1.539; standard deviation, 0.310) to other ICUs. If unit j is not an ICU, increasing unit j's utilization by 10 percentage points today leads to an average pairwise congestion spillover of 0.427 percentage points (maximum, 2.366; standard deviation, 0.411) to other non-ICUs. In Section 6, we conduct counterfactual analyses using the estimation results in column (3) of Table 3 to draw managerial insights.

Importantly, the magnitude of the estimated congestion spillover effects in Table 3 differs substantially from the estimation results obtained without instrumental variables. For comparison purposes, we provide the estimation results obtained by ordinary least squares in Table A1 of the online appendix. We observe that although the estimated congestion spillover effect are all positive and statistically significant, the magnitudes of the effects are much smaller compared with those estimated in Table 3. The coefficients on our endogenous controls, $\Delta Util_{it-1}$ and $\Delta Util_{it}$, also differ significantly. In fact, the coefficients on $\Delta Util_{it-1}$ and $\Delta Util_{it}$ tend to lose statistical significance when estimated by two-stage least squares regressions instead of by ordinary least squares. These differences support the existence of substantial endogeneity bias and thus

the need to adjust for them to correctly identify the congestion spillover effects.

Last, the number of IVs we use to obtain the estimation results in Table 3 was quite high: We used 6 IVs for the model in column (1), 12 for column (2), and 18 for column (3). As discussed in Section 4.6, we used the LASSO estimator to examine whether the model suffers from overfitting and if so to efficiently drop some of our IVs. Somewhat surprisingly, all our IVs were retained, which support the strength and validity of our IVs. We conclude that all our IVs play an important role in identifying the congestion spillover effects and that our models do not suffer from overfitting.

5.2. Robustness Checks

5.2.1. Effect of Congestion Spillover on Nonbusy vs. Busy Days. The congestion spillover effect we estimate in Section 5.1 is the *average* effect across different days. However, the magnitude of the congestion spillover effect may depend on the hospital's congestion level. For example, if the composition of patients admitted on nonbusy days versus busy days are different, the differences in patient needs of the admitted patients may lead to differences in congestion spillover. Indeed, we find some evidence in our data that the composition of patients admitted on nonbusy days versus busy days may be different. For instance, among the Wednesdays in 2018, there were an average of 18.3 scheduled admissions on nonbusy Wednesdays and an average of 22.2 scheduled admissions on busy Wednesdays. One possible way this could influence the effect of congestion spillover is that because scheduled admissions involve surgery that may take up more resources throughout the hospital, the congestion spillover effect when there is a higher number of scheduled admissions may be greater. Thus, to examine whether the magnitude of the congestion spillover effect depend on hospital's congestion level, we estimate our model separately on nonbusy days and on busy days.

Table 4 shows the estimation results. Recall that we define a day to be nonbusy if the daily hospital utilization is among the bottom 80%. Thus, we use 80% of our observations to estimate congestion spillover effect on nonbusy days and use the remaining 20% to estimate congestion spillover effect on busy days. The first-stage F statistics for the significance of the instruments and the Hansen J test results suggest that our instruments tend to lose strength when the model is estimated on busy days, especially when we have a greater number of endogenous variables. This is because of the much

Table 4. Effect of Congestion Spillover on Nonbusy Days vs. Busy Days

	Nonbusy days			Busy days		
	(1)	(2)	(3)	(4)	(5)	(6)
$\Delta Util_{it-1}$	−0.079	−0.029	−0.042	−0.142	0.043	−0.064
	(0.050)	(0.047)	(0.045)	(0.169)	(0.129)	(0.130)
$\Delta Util_{it}$	−0.034	−0.028	−0.017	−0.275	−0.022	−0.179
	(0.090)	(0.067)	(0.077)	(0.292)	(0.181)	(0.183)
$G_i\Delta Util_{it}$	0.481***			0.911*		
	(0.121)			(0.461)		
$G_i^{DiffLevel}\Delta Util_{it}$		0.331***	0.340***		0.473^{+}	0.681*
		(0.096)	(0.102)		(0.258)	(0.273)
$G_i^{SameLevel}\Delta Util_{it}$		1.400***			1.665**	
		(0.353)			(0.582)	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			1.060*			2.115^{+}
			(0.517)			(1.132)
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			1.404***			1.880**
			(0.338)			(0.632)
N	18,640	18,640	18,640	4,672	4,672	4,672
R^2	0.233	0.253	0.257	−0.099	0.148	0.051
Hansen J statistic p value	0.317	0.269	0.474	0.843	0.218	0.018
First-stage F statistics						
$\Delta Util_{it-1}$	74.309	38.583	27.110	13.229	5.654	4.742
$\Delta Util_{it}$	75.671	36.546	27.211	11.643	6.468	4.754
$G_i\Delta Util_{it}$	114.153			13.054		
$G_i^{DiffLevel}\Delta Util_{it}$		48.691	41.306		6.218	4.485
$G_i^{SameLevel}\Delta Util_{it}$		77.675			23.803	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			90.402			16.451
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			115.562			47.068

Notes. N = 18,640 unit-days for columns (1) to (3). N = 4,672 unit-days for columns (4) to (6). Robust standard errors in parentheses.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

smaller sample size we have for busy days, which causes the variation in the instruments to decrease.

Nonetheless, we observe very strong evidence that the congestion spillover effect is indeed larger on days when the hospital is more congested. Specifically, the estimated congestion spillover effects on busy days (columns (4) to (6) in Table 4) are all larger in magnitude than the estimated congestion spillover effects on nonbusy days (columns (1) to (3) in Table 4). Furthermore, the estimated congestion spillover effects on busy days (columns (4) to (6) in Table 4) are all larger in magnitude than the estimated congestion spillover effects on all days (columns (1) to (3) in Table 3).

5.2.2. Different Thresholds for Defining a Nonbusy Day. For our analysis to identify congestion spillover effects to be valid, it is important that the network matrix G and the covariates included in Z are exogenous. One of the ways we ensured that G and Z are exogenous was to use patient flows from nonbusy days with no closed units only when constructing G and Z, where we defined a day to be nonbusy if the daily hospital utilization is among the bottom 80%. In this section, we use different thresholds for defining a day to be nonbusy and examine whether our results are consistent.

Columns (1) to (3) in Table 5 show the estimation results when we define a day to be nonbusy if the daily hospital utilization is among the bottom 70%. Columns (4) to (6) in Table 5 show the estimation results when we define a day to be nonbusy if the daily hospital utilization is among the bottom 90%. The estimated congestion spillover effects are similar to those in Table 3, which shows that our results are robust to different thresholds for defining a day as nonbusy.

5.2.3. Excluding Weekends. If how patients are transferred between units is different on weekends than on weekdays, then the congestion spillover effects on weekends versus weekdays may also differ. As such, we consider excluding data from weekends altogether to conduct our analysis. Columns (7) to (9) in Table 5 show the estimation results. Again, we find that the results are similar to our main results.

6. Counterfactual Analyses

We use our estimation results to evaluate the potential effect of changing the utilization of each unit on the rest of the system in our study hospital. We consider adding a bed to each unit, which mimics a resource allocation problem often faced by hospital administrators—when you have an extra bed to add, to which unit should you add that bed? We use our estimates to predict how the change to one unit caused by adding a bed to that unit will propagate across the inpatient unit network and affect hospital-wide bed availability over time.

We emphasize that we leverage our estimation results to understand the effect of *small* changes in utilization. Big changes in utilization likely result from significant structural changes in the hospital—such as changing the size of some units considerably or adding or removing service lines of substantial size—which may influence the routing decisions and expected length of stay in different units. That is, such big changes are likely to alter the expected patient flows between inpatient units and to invalidate our estimates because they are based on the expected patient flows between inpatient units observed in the data. In other words, our results can be used to estimate the potential effect of utilization changes *if* the expected patient flows between inpatient units will not be affected much by the change. If the particular change alters the expected patient flows between inpatient units (or vice versa), then performing counterfactual analyses using our results is likely to lead to erroneous conclusions.

6.1. Setup of Counterfactual Analyses

To estimate the effect of decreasing the utilization of unit k by adding an extra bed to unit k on day p, we replace $Util_{kp}$ (unit k's utilization on day p) by $Util'_{kp} = Util_{kp} \times \frac{Size_k}{Size_k+1}$, where $Size_k$ is the original number of beds in unit k. We use $Util'_{kp}$ (instead of $Util_{kp}$) for unit k, $Util_{k'p}$ for all units $k' \neq k$, and our estimated model presented in column (3) of Table 3 to predict the counterfactual utilization of all inpatient units on day $p+1$—that is, $Util'_{ip+1}$ for all i. We then use $Util'_{ip+1}$ to predict $Util'_{ip+2}$ and $Util'_{ip+2}$ to predict $Util'_{ip+3}$, and we repeat the process to predict $Util'_{it}$ for $t = p, \ldots, p+30$ for all i. In computing the predictions, we (1) calculate the spillover effect by sampling from the estimated distribution of our parameters (i.e., a trivariate standard normal distribution with mean 0.448, 1.392, and 1.413 (see column (3) of Table 3) and the variance-covariance matrix given by the estimated model) for each starting day p, to take into account the uncertainty in the estimated coefficients, and (2) use the observed values for all other variables.

We predict up to $t = p+30$, 30 days after the change, based on the observation that 30 days is long enough for the effect of the change to stabilize. We perform this prediction for each unit k, $k = 1, \ldots, 16$, for 300 different days (ps) in 2018. We report our results based on the average performance across the 300 different sample paths, along with its 95% confidence interval.

6.2. Results of Counterfactual Analyses

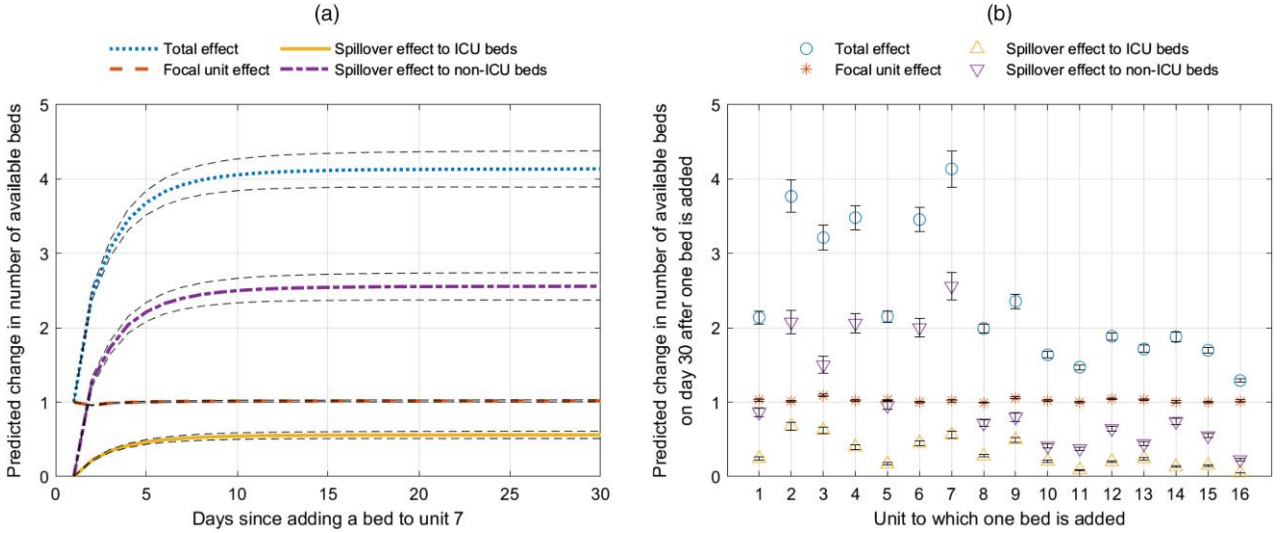
6.2.1. Effect on Hospital-Wide Bed Availability. We use the original utilization $Util_{it}$ and the counterfactual utilization $Util'_{it}$ to quantify the effect of our interventions in terms of the change in the number of available beds in the study hospital. If a bed is added to unit k on

Table 5. Effect of Congestion Spillover, Using Alternate Thresholds for Defining a Nonbusy Day and Excluding Weekends

	Nonbusy day if hospital utilization is bottom 70%			Nonbusy day if hospital utilization is bottom 90%			Excluding weekends		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\Delta Util_{it-1}$	−0.074	−0.008	−0.026	−0.076	−0.008	−0.020	−0.085	0.004	−0.026
	(0.048)	(0.044)	(0.044)	(0.048)	(0.045)	(0.044)	(0.061)	(0.056)	(0.052)
$\Delta Util_{it}$	−0.072	−0.027	−0.045	−0.085	−0.032	−0.066	−0.222^{+}	−0.071	−0.083
	(0.086)	(0.065)	(0.073)	(0.088)	(0.065)	(0.074)	(0.124)	(0.084)	(0.091)
$G_i\Delta Util_{it}$	0.549***			0.579***			0.734***		
	(0.115)			(0.119)			(0.170)		
$G_i^{DiffLevel}\Delta Util_{it}$		0.379***	0.434***		0.397***	0.448***		0.395**	0.423***
		(0.091)	(0.095)		(0.094)	(0.096)		(0.123)	(0.122)
$G_i^{SameLevel}\Delta Util_{it}$		1.288***			1.371***			1.288***	
		(0.304)			(0.318)			(0.319)	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			1.139*			1.526**			0.954^{+}
			(0.472)			(0.468)			(0.524)
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			1.318***			1.488***			1.222***
			(0.295)			(0.308)			(0.316)
N	23,312	23,312	23,312	23,312	23,312	23,312	16,640	16,640	16,640
R^2	0.196	0.233	0.224	0.185	0.229	0.213	0.079	0.222	0.221
Hansen J statistic p value	0.518	0.464	0.191	0.532	0.331	0.062	0.399	0.024	0.008
First-stage F statistics									
$\Delta Util_{it-1}$	87.703	43.241	30.903	88.210	43.186	30.919	70.742	30.669	23.599
$\Delta Util_{it}$	86.029	41.604	31.021	86.611	42.015	31.122	61.477	28.622	21.480
$G_i\Delta Util_{it}$	130.867			129.990			93.566		
$G_i^{DiffLevel}\Delta Util_{it}$		53.986	43.503		54.832	44.268		38.740	31.205
$G_i^{SameLevel}\Delta Util_{it}$		93.652			97.003			76.042	
$G_i^{SameLevel-ICU}\Delta Util_{it}$			112.101			109.755			78.054
$G_i^{SameLevel-NonICU}\Delta Util_{it}$			157.338			158.711			123.887

Notes. N = 23,312 unit-days for columns (1) to (6). N = 16,640 unit-days (weekdays only) for columns (7) to (9). Robust standard errors in parentheses.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Figure 3. (Color online) Predicted Effect of Adding a Bed



Notes. (a) Predicted effect over time when a bed is added to unit 7 with 95% confidence intervals. (b) Predicted effect of adding a bed to each unit with 95% confidence intervals.

day p, the number of available beds in unit k due to the change s days after can be written as $(1 - Util'_{k,p+s})(Size_k + 1) - (1 - Util_{k,p+s})Size_k$. We call this the *focal unit effect* s days after the change. Adding a bed to unit k on day p will free up beds in other units because of congestion spillover. We define the *spillover effect to ICU beds* s days after the change as the number of freed-up ICU beds: $\sum_{i\neq k, i=ICU}(Util_{i,p+s} - Util'_{i,p+s}) \times Size_i$. Similarly, we define the *spillover effect to non-ICU beds* s days after the change as the number of freed-up non-ICU beds, $\sum_{i\neq k, i=non-ICU}(Util_{i,p+s} - Util'_{i,p+s}) \times Size_i$. We refer to their sum as the *total effect* s days after change.

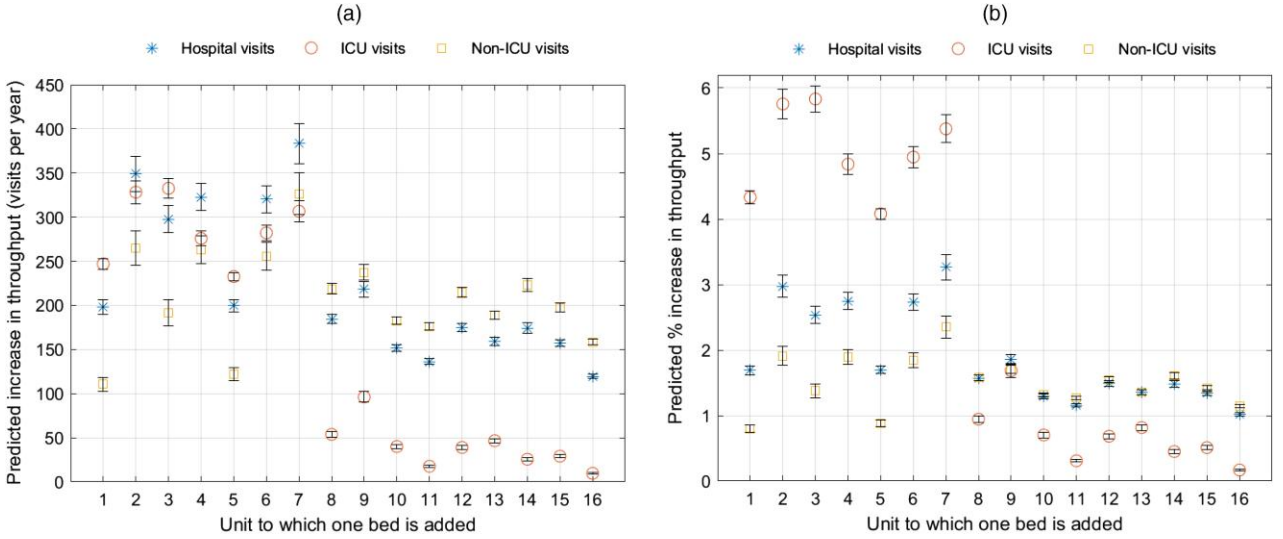
Figure 3(a) illustrates how the focal unit effect, the spillover effect to ICU beds, the spillover effect to non-ICU beds, and the total effect evolve over time when a bed is added to unit 7. The figure shows that the effect of the change stabilizes after about eight days. It shows that adding a bed to unit 7 frees up 4.14 beds in the hospital in the long run, which can be decomposed to 1.02 ICU beds in unit 7 (the focal unit), 0.56 ICU beds in units 1 to 6, and 2.56 non-ICU beds in units 8 to 16. Figure 3(b) summarizes such predictions from our counterfactual analyses for each of the 16 units. For example, the four dots that correspond to unit 7 are the effects that we can read off of Figure 3(a). Figure 3(b) illustrates that adding a bed to unit 7 leads to the biggest total effect. At the same time, it shows that adding a bed to unit 16 results in the smallest total effect.

We make three observations based on Figure 3(b). First, the magnitudes of the total effects are greater than one for all units, and by a substantial amount for the majority of the units, which suggests that *the spillover effect of adding a bed can be substantial*. Second, the focal unit effects, represented by the stars, are similar across the different units and are slightly larger than one bed. Third, the spillover effects, represented by the upward-pointing triangles for ICU beds and downward-pointing triangles for non-ICU beds, vary substantially across different units and are the main driver of the differences in the total effects.

In sum, these observations highlight the key role that the congestion spillover effect plays in identifying the bottleneck of a hospital inpatient unit network. Specifically, in our example of adding a bed, without taking the congestion spillover effect into account, the hospital administrator may choose to allocate an additional bed to a unit, thinking that other units will not be affected by the change. As a result, the hospital may miss the opportunity to more than the triple the benefit that the additional bed can generate in improving hospital-wide bed availability.

6.2.2. Effect on Throughput and Number of Critically Busy Unit-Days. We can quantify the predicted effect of adding a bed to each unit using other performance metrics, such as the change in throughput. For instance, adding a bed to unit 7 is predicted to free up 4.14 beds—1.58 ICU beds and 2.56 non-ICU beds—per day (Figure 3(b)). These numbers can translate to 1511.10 bed-days per year, 576.70 ICU bed-days per year, and 934.40 non-ICU bed-days per year, respectively. In our study hospital, the median length of a hospital visit is 3.94 days; the median length of an ICU visit is 1.88 days; and the median length of a non-ICU visit is 2.86 days. A back-of-the-envelope calculation then shows that adding a bed to unit 7 can increase the study hospital's throughput by $1{,}511.10/3.94 = 383.53$ more hospital visits per year,

Figure 4. (Color online) Predicted Increase in Throughput After Adding One Bed to Each Unit



Notes. (a) Predicted increase in throughput measured in visits per year with 95% confidence intervals. (b) Predicted % increase in throughput with 95% confidence intervals.

which is about a 3% increase in throughput measured by hospital visits. The change in ICU throughput and non-ICU throughput can be computed separately, as well: adding a bed to unit 7 can increase ICU throughput by 576.70/1.88 = 306.76 more ICU visits per year (a 5% increase) and increase non-ICU throughput by 934.40/2.86 = 326.71 more non-ICU visits per year (a 2% increase). Figure 4 summarizes these predicted increases in throughput when a bed is added to each unit.

Although increasing throughput is a common goal that many hospitals share, some hospitals may have different objectives. For instance, given the wealth of evidence that high bed utilization may lead to worse patient outcomes (Forster et al. 2003, KC and Terwiesch 2012, Kim et al. 2015, Shi et al. 2016, Long and Mathews 2018, Dong et al. 2019, Song et al. 2020), hospitals may want to avoid having critically busy units. This type of performance metric can also be computed from our counterfactual analysis. Figure 5 shows the predicted average change in unit-days with unit utilization greater than 90%[6] measured over 30 days (recall that we simulate the effect of each change for 30 days). For example, if a bed is added to unit 3, it is predicted that, for the next 30 days, the number of days on which the utilization of unit 3 will be greater than 90% will decrease by 8.69 days, and the number of days on which the utilization of other units will be greater than 90% will decrease by 6.49 days (3.78 days for ICUs and 2.71 days for non-ICUs) due to the spillover effect.

Examining the results in Figures 4 and 5 together, we find that the best unit to target may change, depending on the decision maker's objective. To make the optimal decision for their hospitals, administrators can use the multi-dimensional aspect of the benefit of allocating resources to different units—which our analysis provides—and the cost of resources.

6.3. Congestion Spillover Effects and Unit Characteristics

In this section, we examine whether unit characteristics can explain the predicted total effect of adding a bed to a unit. We consider the following characteristics:

- Unit size: Because adding a bed to smaller units decreases utilization more than adding a bed to larger units, in general, it is expected that the total effect, ceteris paribus, will be larger for smaller units.

Figure 5. (Color online) Predicted Change in Unit-Days with Utilization >90% over 30 Days After Adding One Bed to Each Unit

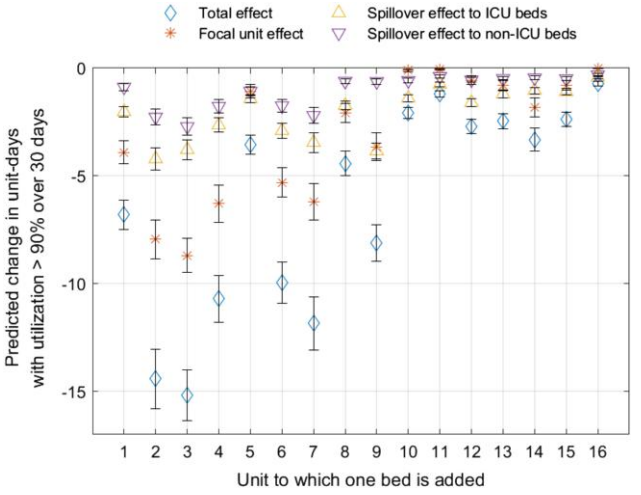


Table 6. Association Between the Effect of Adding a Bed to Each Unit and Unit Characteristics

	(1)	
Unit size	−0.078**	(0.022)
Average utilization	−0.030	(0.018)
Network centrality	0.221*	(0.081)
ICU indicator	0.961	(0.860)
Median LOS	0.260	(0.311)
Average no. of daily admissions normalized by unit size	−0.000	(0.039)
Average no. of daily discharges normalized by unit size	0.027	(0.048)
Constant	3.421*	(1.319)
N	16	
R^2	0.871	

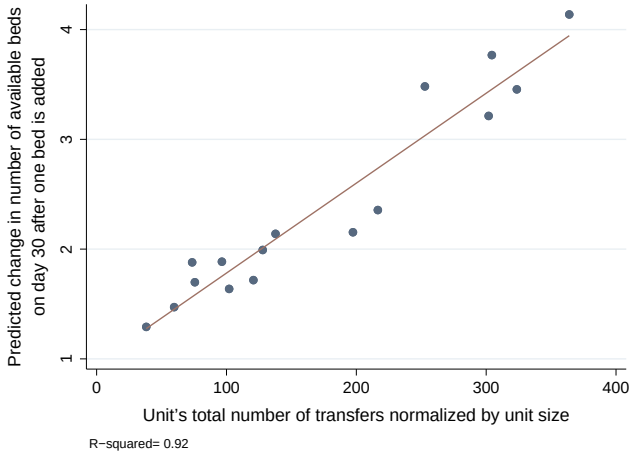
Notes. Column (1) is a linear regression model estimated at the unit level. The dependent variable is the total effect (measured 30 days after the change). Robust standard errors in parentheses.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

- Average utilization: Hospital administrators often use average utilization to identify bottleneck units. However, as discussed in Section 3.2, the unit with the highest observed utilization may not necessarily be the bottleneck unit.
- Network centrality: Network centrality indicates the importance of each unit in the inpatient unit network. We measure the network centrality of each unit in our network matrix G using the PageRank centrality measure—a commonly used centrality measure in the literature (Drakopoulos and Zheng 2017).
- ICU: The total effect may depend on whether the unit is an ICU versus a non-ICU.
- Median LOS: The total effect may depend on the speed of patient turnover, which we measure using median LOS.
- Average number of daily admissions and discharges normalized by unit size: The total effect may depend the number of admissions and discharges a unit needs to process each day.

Table 6 shows the result of regressing the total effect on the aforementioned unit characteristics. We find that the coefficients for unit size and network centrality are statistically significant at the 5% significance level. As expected, larger units have smaller total effects (note that ICUs are likely to be smaller, and we expect the unit size variable to capture some of the effect of a unit being an ICU), and units that are more connected in the inpatient unit network have larger total effects. Interestingly, all the other variables are not statistically significant. Importantly, we find that the average utilization is not a good predictor of the total effect of adding a bed, which supports our recommendation that hospital administrators should be cautious about using only average utilization to guide their capacity-related decision making (Section 3.2).

Our results show that hospitals should target units with higher network centrality when deciding on resource allocation to different units. However, network centrality may not be readily available or easy for hospital administrators to compute. Because network centrality measures the connectivity of each unit to the rest of the inpatient unit network—that is, units with more transfers to and from other units are more connected in the network—we construct a proxy for network centrality by computing the number of transfers of each unit to and from other units and normalizing it by unit size (because larger units naturally have more transfers). Figure 6 illustrates the remarkable performance of our proxy variable—It shows that our proxy explains about 92% of the variation in the total effect. This result suggests that the total number of transfers to and from each unit can be an important piece of information that hospital administrators can use as they try to identify the bottleneck in the complex inpatient unit network and thus better use their resources.

Figure 6. (Color online) Association Between the Effect of Adding a Bed to Each Unit and Total Number of Transfers Normalized by Unit Size



7. Conclusions

In this paper, we study the congestion spillover in the inpatient unit network. Using data collected from a large teaching hospital and econometrics tools, we estimate the causal effect of changes to one unit's utilization on the utilization of the other units in the inpatient unit network. We find that congestion spillover is substantial. Using counterfactual analyses, we demonstrate the importance of quantifying the spillover effect in identifying the bottleneck unit and how hospital administrators can use our framework to make resource allocation decisions for their hospitals.

The resource we focus on is inpatient beds, and our primary performance metric is bed availability and hospital throughput. We emphasize that hospitals can apply our empirical framework to understand the spillover effect of other resources—such as care providers and medical equipment—on other performance metrics—such as the percentage of blocked admissions, the percentage of off-service placement, delay in transfer, and hospital length of stay. Going beyond hospitals, to examine the spillover effect of resources on system performance, our empirical framework can be applied to other manufacturing and service processes that have flexible resources, have a complex network of resources with endogenous routing of jobs within the network, and produce highly customized products.

Our study has limitations that suggest future research directions. First, as discussed in our counterfactual analyses, we study small changes to the system while holding the patient routing policy fixed. Although the assumption of a fixed patient routing policy is reasonable for analyzing marginal changes, such an assumption fails when one wants to evaluate the effect of structural changes. As such, designing the optimal routing policy while taking into account the inpatient unit network structure is a promising direction for future research. Second, the focus of the paper is to quantify the magnitude of the congestion spillover, but our analysis cannot be used to identify the main drivers of congestion spillover. The existing literature suggests various reasons for congestion spillover, such as waiting for shared resources across different units, care providers slowing down due to mental strain and lower quality of care that leads to more work in other units (KC and Terwiesch 2012, Kuntz et al. 2015). Identifying the main drivers of congestion spillover may be a fruitful direction that could yield improvements in practice.

Acknowledgments

The authors gratefully acknowledge the editors and reviewers for many helpful suggestions and comments that greatly improved this paper. The authors also thank the participants at the Wharton Empirical Workshop in Operations Management and the INFORMS and MSOM meetings for very helpful comments.

Endnotes

[1] Our study hospital does not have an emergency room. However, our empirical framework can be easily applied to hospitals with emergency rooms (see Section 4.3).

[2] We also use different thresholds for defining a day to be nonbusy. We show that the results are qualitatively similar in the robustness checks in Section 5.2.

[3] Because scheduled admissions usually occur on weekdays, we removed 94 scheduled admissions on weekends.

[4] The previous year's G is a good proxy for the focal year's G. Specifically, the element-by-element correlation coefficient between G_y and G_{y+1} for year y is 0.975 for $y = 2014$, 0.983 for $y = 2015$, and 0.986 for $y = 2016$.

[5] The assumption of hospital admissions from the emergency room being exogenous to the hospital's congestion level may not hold when there is an extreme surge in demand for hospital care, for example, during a pandemic. In such situations, one needs to be more careful in constructing instrumental variables that are exogenous to the hospital's congestion level.

[6] Other high utilization levels might be of interest to hospitals as well. We use 90% without loss of generality.

References

Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2015) Systemic risk and stability in financial networks. *Amer. Econom. Rev.* 105(2):564–608.

Ahrens A, Hansen CB, Schaffer M (2018) Pdslasso and ivlasso: Progams for post-selection and post-regularization OLS or IV estimation and inference. http://ideas.repec.org/c/boc/bocode/s458459.html.

Allon G, Deo S, Lin W (2013) The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Oper. Res.* 61(3):544–562.

Anselin L (2001) Spatial econometrics. Baltagi BH, ed. *A Companion to Theoretical Econometrics* (Blackwell Publishing Ltd, Oxford, England), 310–330.

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. National Acad. Sci. USA* 106(51):21544–21549.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econom. Stud.* 81(2):608–650.

Belloni A, Chen D, Chernozhukov V, Hansen C (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6):2369–2429.

Berry Jaeker JA, Tucker AL (2016) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Sci.* 63(4):1042–1062.

Bramoullé Y, Djebbari H, Fortin B (2009) Identification of peer effects through social networks. *J. Econometrics* 150(1):41–55.

Cachon G, Terwiesch C (2012) *Matching Supply with Demand* (McGraw-Hill Publishing).

Chan CW, Farias VF, Escobar GJ (2017) The impact of delays on service times in the intensive care unit. *Management Sci.* 63(7): 2049–2072.

Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* 60(6):1323–1341.

Chernozhukov V, Hansen C, Spindler M (2015) Post-selection and post-regularization inference in linear models with many controls and instruments. *Amer. Econom. Rev.* 105(5):486–490.

Clark JR, Huckman RS (2012) Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* 58(4):708–722.

Coffman JM, Seago JA, Spetz J (2002) Minimum nurse-to-patient ratios in acute care hospitals in California. *Health Affairs* 21(5):53–64.

Copenhaver MS, Hu M, Levi R, Safavi K, Zenteno Langle AC (2019) Health system innovation: Analytics in action. *Operations Research and Management Science in the Age of Analytics* (INFORMS), 238–266.

Dong J, Shi P, Zheng F, Jin X (2019) Off-service placement in inpatient ward network: Resource pooling vs. service slowdown. Working paper, Columbia University, New York.

Drakopoulos K, Zheng F (2017) Network effects in contagion processes: Identification and control. Research Paper 18-8, Columbia Business School, New York.

Forster AJ, Stiell I, Wells G, Lee AJ, Van Walraven C (2003) The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad. Emergency Medicine* 10(2):127–133.

Freeman M, Savva N, Scholtes S (2021) Economies of scale and scope in hospitals: An empirical study of volume spillovers. *Management Sci.* 67(2):673–697.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Harvard Business School (1998) Donner company. *Harvard Bus. School* 9(689030):1–14.

He P, Zheng F, Belavina E, Girotra K (2021) Customer preference and station network in the London bike-share system. *Management Sci.* 67(3):1392–1412.

Jweinat J, Damore P, Morris V, D'Aquila R, Bacon S, Balcezak TJ (2013) The safe patient flow initiative: A collaborative quality improvement journey at Yale-New Haven Hospital. *Joint Comm. J. Quality Patient Safety* 39(10):447–AP9.

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing Service Oper. Management* 14(1):50–65.

KC DS, Scholtes S, Terwiesch C (2020) Empirical research in healthcare operations: Past research, present understanding, and future opportunities. *Manufacturing Service Oper. Management* 22(1):73–83.

Kim SH, Chan CW, Olivares M, Escobar G (2015) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Sci.* 61(1):19–38.

Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Sci.* 61(4):754–771.

Long EF, Mathews KS (2018) The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production Oper. Management* 27(12):2122–2143.

Mankad S, Shunko M, Yu Q (2019) How to find your most valuable outlets? Measuring influence in service and retail networks. Preprint, submitted April 4, http://dx.doi.org/10.2139/ssrn.3366127.

Manski CF (1993) Identification of endogenous social effects: The reflection problem. *Rev. Econom. Stud.* 60(3):531–542.

McConnell KJ, Richards CF, Daya M, Bernell SL, Weathers CC, Lowe RA (2005) Effect of increased ICU capacity on emergency department length of stay and ambulance diversion. *Ann. Emergency Medicine* 45(5):471–478.

Osadchiy N, Gaur V, Seshadri S (2016) Systematic risk in supply chain networks. *Management Sci.* 62(6):1755–1777.

Rutherford P, Provost L, Kotagal U, Luther K, Anderson A (2017) *Achieving Hospital-Wide Patient Flow* (Institute for Healthcare Improvement, Cambridge, MA).

Serpa JC, Krishnan H (2018) The impact of supply chains on firm-level productivity. *Management Sci.* 64(2):511–532.

Shi P, Chou MC, Dai J, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Sci.* 62(1):1–28.

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Sci.* 66(9):3825–3842.

Thomas BG, Bollapragada S, Akbay K, Toledano D, Katlic P, Dulgeroglu O, Yang D (2013) Automated bed assignments in a complex and dynamic hospital environment. *Interfaces* 43(5):435–448.

Van den Bulte C, Lilien GL (2001) Medical innovation revisited: Social contagion vs. marketing effort. *Amer. J. Sociol.* 106(5):1409–1435.

Wang Y, Li J, Wu D, Anupindi R (2021) When ignorance is not bliss: An empirical analysis of subtier supply network structure on firm risk. *Management Sci.* 67(4):2029–2048.

Yu Q, Shunko M, Mankad S (2017) A quality value chain network: Linking supply chain quality to customer lifetime value. Preprint, submitted June 2, https://dx.doi.org/10.2139/ssrn.2979592.

Zychlinski N, Mandelbaum A, Momčilović P, Cohen I (2020) Bed blocking in hospitals due to scarce capacity in geriatric institutions: Cost minimization via fluid models. *Manufacturing Service Oper. Management* 22(2):396–411.