This article was downloaded by: [162.194.116.104] On: 23 March 2025, At: 12:48
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

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

The Spillover Effects of Capacity Pooling in Hospitals

Jong Myeong Lim, Hummy Song, Julius J. Yang

To cite this article:
Jong Myeong Lim, Hummy Song, Julius J. Yang (2024) The Spillover Effects of Capacity Pooling in Hospitals. Management Science 70(11):7692-7711. https://doi.org/10.1287/mnsc.2022.02202

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2024, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

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

MANAGEMENT SCIENCE
Vol. 70, No. 11, November 2024, pp. 7692–7711
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

The Spillover Effects of Capacity Pooling in Hospitals

Jong Myeong Lim,[a,*] **Hummy Song,**[b] **Julius J. Yang**[c]

[a] Miami Herbert Business School, University of Miami, Coral Gables, Florida 33146; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215

*Corresponding author

Contact: jonglim@mbs.miami.edu, https://orcid.org/0000-0001-5058-7493 (JML); hummy@wharton.upenn.edu, https://orcid.org/0000-0003-1335-9314 (HS); jyang@bidmc.harvard.edu (JJY)

Received: March 8, 2021
Revised: March 19, 2023; June 19, 2023
Accepted: June 22, 2023
Published Online in Articles in Advance: January 16, 2024

https://doi.org/10.1287/mnsc.2022.02202

Copyright: © 2024 INFORMS

Abstract. Off-service placement is a common capacity-pooling strategy that hospitals utilize to address mismatches in supply and demand that arise from the day-to-day variation in patient demand. This strategy involves placing patients in a bed in a unit that is designated for another specialty service. Building on prior work that documents the negative first order effects of off-service placement on patients who are placed off service themselves, we quantify the spillover effects of this practice on patients who are actually placed on service. Using an estimation strategy that combines the Heckman correction procedure and a heteroskedasticity-based identification approach, we find that off-service placement has substantial negative spillover effects on the efficiency of care delivered to on-service patients. In particular, we find that a 10 percentage point increase in the level of off-service placement during a patient's hospitalization is associated with a 10.9% increase in length of stay. Through a series of counterfactual analyses, we propose alternate routing and capacity-planning policies that could meaningfully improve the efficiency of care in the inpatient setting.

History: Accepted by Stefan Scholtes, healthcare management.

Funding: This work was supported by Claude Marion Endowed Faculty Scholar Award; Fishman-Davidson Center for Service and Operations Management.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.02202.

Keywords: healthcare • hospitals • off-service placement • capacity pooling • admission control

1. Introduction

Hospitals often face significant variability in demand in terms of both the number of patients needing care and the type of care needed by each patient. Whereas such variability exists on the demand side, hospitals operate with a fixed number of beds not only across the entire hospital, but also within each specialty service (e.g., cardiology, general surgery, etc.), which is allocated a fixed number of beds. The mismatch between the number and type of patient arrivals and the capacity of hospital beds presents significant challenges in matching the supply with the demand.

One of the strategies employed by hospitals and many other industries that face similar problems is capacity pooling. This practice allows the hospital to use underutilized capacity in a less busy service when other services are at or near full capacity. Pooling the capacity of hospital beds results in the placement of patients of a focal service in a bed that is located in a unit that is designated for another service; this is called "off-service placement" (Stylianou et al. 2017, Dong et al. 2020a, Song et al. 2020). This practice is common across many hospitals. Dong et al. (2020a) report that 22% of patients are placed off service in a large teaching hospital in Southeast Asia, Stylianou et al. (2017) find this percentage to be around 10% in a hospital in England, and we find that around 20% of patients are placed off service in our study hospital in the United States. Whereas off-service placement allows for a more efficient use of beds, recent empirical work finds that off-service placement has negative consequences when it comes to the care provided for the patients who are placed off service. Stylianou et al. (2017) find that these patients have longer lengths of stay on average. Song et al. (2020) obtain similar results using an instrumental variable approach to estimate the causal effect of off-service placement, finding that off-service patients experience longer lengths of stay and a higher likelihood of readmission within 30 days.

An important question that remains is whether the practice of off-service placement has any downstream effects on the rest of the patient population that is not impacted by the practice firsthand. In other words, are there any spillover effects of off-service placement of

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

which hospital administrators should be aware? The existing literature has not yet explored this possibility of off-service placement having a broader spillover effect on patients who have not been placed off service themselves but do belong to the same service that has some of its patients placed off service. A priori, the answer to this question is not obvious. Whereas it is possible that the consequences of off-service placement are limited to the negative first order effects previously documented, the broader impact on the workflow of the physicians who are caring for all patients on their service—regardless of their placement location—could lead to substantial negative spillover effects whereby patients who are placed on service are also negatively impacted. Understanding these spillover effects is important to hospital administrators because they have significant implications for managing hospital capacity, especially given that many hospitals operate at very high levels of utilization.

Using an estimation approach that combines the Heckman correction procedure and a heteroskedasticity-based identification approach, we find that off-service placement has substantial negative spillover effects on the efficiency of care delivered to on-service patients. In other words, patients placed on service tend to experience longer lengths of stay when the average level of off-service placement for the service is high. Specifically, a 10 percentage point increase in the level of off-service placement during a patient's hospitalization is associated with a 10.9% increase in length of stay. Using several alternate methodological approaches offers additional support for these findings. Our analyses also suggest that an important mechanism driving the negative spillover effects is the challenge of coordinating care between physicians and nurses.

Using the point estimates from our empirical analyses, we conduct a series of counterfactual analyses to show the expected performance of several alternate routing and capacity planning related policies that may be able to reduce the negative impact of off-service placement and retain the benefits of capacity pooling. We find that limiting the practice of reserving on-service beds in anticipation of future demand can lead to significant reductions in the overall level of off-service placement, which, in turn, is expected to result in improvements in the efficiency of care. Policies such as boarding patients for an extra hour when an on-service bed is expected to become available soon or prioritizing early discharges may also be effective in reducing the level of off-service placement; both of these policies are expected to lead to shorter lengths of stay. Whereas adding capacity to the most congested service is a relatively straightforward approach to alleviating off-service placement, we find that simply reassigning a unit from the least utilized service to the most congested service can also lead to substantial improvements.

A key contribution of this work lies in the quantification of the spillover effects of off-service placement. In doing so, we add to the recent stream of work analyzing the effects of off-service placement in healthcare delivery settings, especially with regards to challenges and unintended consequences when implementing capacity pooling strategies (Stylianou et al. 2017, Dong et al. 2020a, Song et al. 2020, Kim et al. 2023). This work also contributes to the literature on capacity management in healthcare delivery settings (Shi et al. 2016, Dong and Perry 2020, Dai and Shi 2021, Xie et al. 2021). The practical implications of our work are substantial given that many hospitals utilize off-service placement as a capacity-pooling strategy. For hospital managers, our work further highlights the importance of better managing off-service placement because its impact is not only limited to those patients who are placed off service themselves, but also extends to the patients who are placed in on-service beds, effectively impacting the entire population of hospitalized patients.

The rest of the paper is organized as follows. Section 2 discusses off-service placement in more detail, paying particular attention to the impact of off-service placement on patients placed on service. The research setting and data are introduced in Section 3. Section 4 motivates our empirical strategy and presents the main results, in which we recover the causal effect of off-service placement on on-service patients. We discuss several alternate estimation approaches and present the estimation results in Section 5. Section 6 considers mechanisms that may be driving the spillover effects of off-service placement. The procedures and results of several counterfactual simulation studies are discussed in Section 7. Section 8 concludes.

2. Off-Service Placement: First Order and Spillover Effects

In recent years, there has been a growing body of work that seeks to understand the effects of off-service placement on patient outcomes and system performance. In this section, we provide a brief overview of the existing literature and motivate why there is a need to better understand the potential spillover effects of this practice.

2.1. First Order Effects of Off-Service Placement

To date, the research that studies this widespread practice of off-service placement can be characterized as focusing on its first order effects. Some of this work seeks to address the question of how off-service placement impacts the efficiency and quality of care received by patients who themselves have been placed off service. Using an instrumental variables approach, Song et al. (2020) estimate that being placed off service is associated with a 23% increase in length of stay and a

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

13% increase in the likelihood of 30-day hospital readmission for hospitalized medical/surgical inpatients. Similar effects are documented in studies that focus on specific specialty services, such as general (Bai et al. 2018, Kohn et al. 2021), pulmonary (Kohn et al. 2020), and cardiac (Alameda and Suárez 2009) medicine.

Meanwhile, others focus on the decision-making process underlying the bed placements and its impact on system-wide performance. Shi et al. (2016) develop a stochastic network model that allows for off-service placement to show the effects of this practice on average waiting time performance. Dong et al. (2020a) examine some of the factors that drive the on- versus off-service placement decision-making process. Using structural estimation methods, they find that bed managers are more likely to place patients off service when the service is busy and the admission occurs overnight. They also illustrate that a uniform routing policy could reduce the overall levels of off-service placement and improve system performance. In contrast, Dai and Shi (2019) treat the bed manager's decision making as a Markov decision process and solve for the optimal routing policy using approximate dynamic programming. Izady and Mohamed (2021) propose a routing policy by which a cluster of services has a designated flex unit for admitting patients. They find that an optimal configuration of these clusters can lead to reductions in the cost of denied admissions.

2.2. Spillover Effects of Off-Service Placement

Beyond the first order effect, placing patients off service may have an important second order effect as well, which we refer to as the spillover effect. There are two ways in which a spillover effect of off-service placement could impact on-service patients. First, there may be a spillover effect at the service level, wherein on-service patients belonging to a particular service may be impacted by the extent to which there are off-service patients who also belong to the same service. Second, there may be a spillover effect at the unit level, wherein on-service patients located in a particular unit may be impacted by the extent to which there are off-service patients who are located in the same unit.[1] In this paper, we focus specifically on the service-level spillover effects for two interrelated reasons.[2] First, we are interested in understanding how the practice of off-service placement impacts the work being carried out by both physicians and nurses. Whereas unit-level spillover effects are expected to be impacted by the work being done by nurses only, service-level spillover effects should be impacted by the work of both physicians and nurses. Second, prior work by Song et al. (2020) shows that the work of physicians may be more impacted by off-service placement than that of nurses given their examination of potential mechanisms underlying the first order effects of off-service placement. For brevity, we refer to service-level spillover effects simply as the spillover effects in the remainder of this paper.

A large body of prior research shows that the efficiency and quality of care for patients is impacted by operational factors, such as the system's load (Kc and Terwiesch 2009, Kuntz et al. 2015, Berry Jaeker and Tucker 2017), service-level mismatch (Kim et al. 2015, Chan et al. 2019), and facility layout (Meng et al. 2021). The spillover effect of off-service placement, separately accounting for the previously documented operational factors, such as system load, may be another important aspect that has been overlooked. To operationalize the extent to which a given service is engaged in off-service placement, we first count the number of off-service patients belonging to a service at a given point in time and divide it by the total number of on-service beds assigned to the service. Then, we evaluate the overall impact of off-service placement for a given on-service patient by taking the average level of off-service placement experienced by the patient over the course of the patient's hospitalization. This approach allows us to measure the level of off-service placement so that it is comparable across services of different sizes.

There are several potential mechanisms through which increases in the level of off-service placement might negatively impact on-service patients. First, physicians provide care for all patients who belong to the service regardless of whether the patient is placed in an on- or off-service bed. In contrast, nurses are only responsible for the care of patients who are located in their unit regardless of the service to which the patient belongs. As a result, having patients placed in an off-service unit creates an ad hoc team comprising the service's physicians and the unit's nurses, who do not have an established working relationship. Higher levels of off-service placement may increase the number of such ad hoc teams being formed, and these are likely to result in increased coordination costs (Reagans et al. 2005, Dobson et al. 2009). Having more off-service patients may also create disruptions to the provider workflow as the care coordination meetings and rounds for off-service patients require physicians to allocate additional time outside of their normal routine (Gesensway 2010). Typically, this involves physically traveling to the unit where the off-service patient is located, which consumes additional time and further disrupts the physician's workflow. Meng et al. (2021) find that the distance between patient beds and the nurses' station significantly affects care patterns such that nurses are more likely to batch tasks for patients who are located in rooms that are farther away. Similarly, we expect the provision of care to be different for patients who are placed on versus off service, and the care provided for off-service patients may, in turn, affect the way in which care is provided for on-service patients. Taken together, we hypothesize that increases in the

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

level of off-service placement lead to decreases in the efficiency of care delivered to patients who are placed on service.

3. Research Setting and Data

3.1. Research Setting

We collaborated with a large academic medical center located in the northeastern region of the United States. As of 2016, this hospital had 473 medical/surgical inpatient beds, which were located across 17 units and allocated to eight services. Here, a unit refers to a physical location where there are a certain number of beds. In turn, each unit is designated to a particular service, which is a department comprising a single clinical specialty or a related group of specialties that tend to have smaller volumes. Figure A.1 in Online Appendix A shows which units are designated for which services in the study hospital. Because each unit belongs to a specific service, we are able to determine whether the patient occupying a bed in a particular unit has been placed on service (if the patient belongs to the service for which the unit has been designated) or off service (if the patient belongs to a service other than the one for which the unit has been designated). Thus, at the service level, the number of patients who are placed off service can vary over time, and the number of distinct off-service units across which these patients are placed can also vary over time.

In our study hospital, there are eight services: cardiac medicine, cardiac surgery, east surgery, general medicine, neurology, oncology medicine, transplant, and west surgery.[3] Note that physicians belong to a particular service, whereas nurses belong to a particular unit. Such organization is ubiquitous among hospitals as physicians often specialize in particular clinical specialties, whereas nurses are generally trained to provide nursing care for all types of medical/surgical inpatients.[4] In effect, this means that physicians are responsible for the care of all patients who belong to their service regardless of whether the patient is placed on or off service. In contrast, nurses are responsible for the care of all patients who are located in their unit regardless of the service to which the patient belongs. This has implications for the familiarity that accumulates between physicians and nurses; the most frequent teamwork and coordination happens between physicians of a given service and the nurses working on the unit that is designated to that particular service.

Whereas physicians and nurses are responsible for patient care, the decision to place the patient in a particular on- or off-service bed is made by the hospital's bed managers. Bed managers are centralized at the hospital level and serve the role of admission controllers. They are experienced nurses with clinical knowledge, and they coordinate with the admitting service and the various units across the hospital to assign the patient to a specific inpatient bed. The bed manager does not have any discretion over to which service to admit the patient; the service makes this determination based on the patient's medical conditions and clinical needs. The bed manager, however, does have the discretion to determine where to physically place the patient given the availability of beds in different units across the hospital. For example, if there is an incoming cardiac medicine patient but all on-service beds for cardiac medicine are occupied, the bed manager can assign this patient to an open bed in a unit that belongs to general medicine; this creates an off-service placement for the cardiac medicine service. Note, when the bed manager decides to place a patient off service, only other medical/surgical beds within the same level of care are considered. In other words, this cardiac medicine patient would not be placed in a critical-care bed (e.g., in a cardiac intensive care unit) or a nonmedical/surgical bed (e.g., in an obstetrics unit).

3.2. Data and Analysis Sample

Our data consist of detailed patient and operational data from November 1, 2015, to September 30, 2016. By combining multiple proprietary data sources, we can accurately track the location of all patients at the bed–hour level over the study period. In addition, we have detailed information about each patient encounter, including demographic characteristics, primary diagnosis, complications, and comorbidities. Comparing the patient's service and bed assignment allows us to determine whether the patient was placed on or off service. The granularity of our data also allows us to track, at the hour level, the number of on- and off-service patients belonging to a given service.

During the study period, there were 48,767 complete patient encounters at this hospital. To define our analysis sample, we first exclude 24,698 patient encounters without any time in a medical/surgical bed (e.g., observation stays and same-day discharges). Then, we exclude 5,112 patient encounters in which patients were transferred multiple times to different medical or surgical units, including from an on-service bed to an off-service bed or vice versa, in order to isolate the spillover effect of off-service placement from the direct effect of being placed off service. We also exclude 746 patient encounters in which the patient's service designation changed during the hospitalization. In other words, we restrict our analysis sample to those who were placed in an on-service bed during the entire duration of their hospital stay. This leaves us with 18,211 total medical/surgical patient encounters. Because our analysis focuses on how the on-service patient population is impacted by the spillover effects of off-service placement, we further restrict our sample to patients who were placed in an on-service bed, which leaves us with

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

a total of 14,787 on-service patient encounters across 13,379 unique patients.

3.3. Outcome Measures

We focus on the efficiency of care delivered to on-service patients, using length of stay as a proxy. We define length of stay as the time from each patient's first admission into a medical/surgical bed until the patient's time of discharge. For example, if a patient's first entry point to the hospital is the emergency department, we do not consider the patient to have begun the hospital stay until the patient leaves the emergency department and is transferred into a medical/surgical bed. Our definition of length of stay is purposefully aligned with the goal of analyzing the impact of off-service placement and is designed to represent the efficiency of care a patient receives once the patient begins a stay in a medical/surgical bed. Because this measure is right skewed, we log transform it to calculate the logged length of stay. We report summary statistics of each of these measures in Table 1.

3.4. Key Explanatory Variable

Our key explanatory variable of interest is the average level of off-service placement as experienced by the focal on-service patient during the hospitalization. We construct this measure by first calculating hourly snapshots of the number of patients who are placed off service divided by the number of on-service beds assigned to the focal service. This measures the proportion of a service's patients who are placed off service during each hour of a patient's hospitalization. Then, using these hourly snapshots, we define the average level of off-service placement by calculating the mean of the hourly snapshots over the duration of each on-service patient's hospital stay.

3.5. Patient and Operational Characteristics

In our estimations, we account for several patient and operational characteristics that may also impact our outcome measures of interest. We control for patient demographic characteristics, including age and gender. We also account for three proxies of patient severity: diagnosis-related group (DRG) cost weight, an indicator for the presence of complications or comorbidities, and an indicator for whether the patient spent time in the intensive care unit (ICU). The DRG cost weight represents the relative level of resources needed to treat a patient with a certain diagnosis. The average DRG cost weight of all patients admitted to a given hospital is sometimes referred to as the case mix index, and the average DRG cost weight in our analysis of 1.82 indicates that patients admitted to our study hospital have diagnoses that are, on average, 82% more costly to treat than the average Medicare patient. We also observe and account for the presence of complications or comorbidities as classified in the DRG. Another proxy of severity is whether the patient spent time in the ICU, which suggests the need for higher intensity care.

We also account for several characteristics of each patient's hospitalization, including the number of intrahospital transfers, the time of day of admission, whether the patient was admitted on a weekday, and the month of admission. The average patient in our analysis sample had 2.82 transfers during hospitalization, which includes the admission and discharge events as these constitute transfer events into or out of a bed. We observe and account for heterogeneity in admission time and day by

Table 1. Summary Statistics of Analysis Sample

	Mean	Standard deviation	Minimum	Maximum
Length of stay, days	4.15	3.33	0.70	23.8
Logged length of stay, days	1.48	0.54	0.53	3.21
Average level of off-service placement, %	17.7	16.1	0	97.1
Unit-level utilization, %	91.4	6.91	26.2	100
Service-level utilization, %	100.1	36.0	12.4	233.0
Age, years	62.2	16.7	16.8	107.4
Female, %	50.5	50.0	0	100
DRG cost weight	1.82	1.37	0.49	17.7
Complications or comorbidities, %	24.6	43.1	0	100
ICU encountered, %	14.9	35.6	0	100
Number of transfers	2.82	0.89	2	9
Admitted on weekday, %	85.0	35.7	0	100
Admission shift, %				
AM shift	12.0	32.5	0	100
PM shift	27.6	44.7	0	100
Overnight shift	60.4	48.9	0	100

Note. N = 14,787.

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

controlling for the shift during which the patient was admitted, whether the patient was admitted on a weekday, and the average DRG cost weights of all patients admitted to the same service in the five-hour window around the focal patient's admission.[5] We also include service and unit fixed effects to control for any underlying differences in patient heterogeneity and productivity. To further control for unobserved factors, we also include two interactions: (1) service fixed effects interacted with the time of day the patient was admitted and (2) service fixed effects interacted with whether the patient was admitted on a weekday.

Finally, we account for operational factors that may impact a patient's stay.[6] In particular, we account for the overall busyness of the unit and of the service so that we can isolate these effects from the spillover effect that we are interested in identifying. Specifically, we control for the average hourly utilization level of the unit in which each patient was placed and the service to which the patient belonged. We construct the hourly unit-level utilization by dividing the number of beds that are unavailable for an incoming patient (i.e., occupied, reserved, or closed beds) by the number of total beds in the unit (i.e., all unavailable beds plus open beds). We construct the hourly service-level utilization by dividing the total number of patients currently belonging to the service by the total number of beds in all units that have been designated to the service. Given the prior research that shows that high levels of workload impose an inverted U-shaped effect on patient outcomes (Kuntz et al. 2015, Berry Jaeker and Tucker 2017), we include the squared terms of both unit- and service-level utilization measures in our analyses as well. To further account for the workload of providers, we also include (a) the average DRG cost weights of all patients in the same service during the focal patient's hospitalization and (b) the average count of movements (i.e., admissions, discharges, and transfers) of all patients in the same service. This extensive list of control variables allows us to measure the effect of off-service placement separately from the overall workload effect.

4. Main Empirical Strategy and Results

To identify the spillover effects of off-service placement, our goal is to estimate the following:

$$y_i = \alpha + \beta \cdot \textit{Average level of off-service placement}_i + \gamma \cdot \mathbf{X}_i + \epsilon_i. \quad (1)$$

Here, y_i is the logged length of stay for patient i, $\mathbf{X}_i$ is a vector of control variables described in Section 3.5, and ϵ_i captures the error term. The spillover effect of off-service placement is captured by β. Simply estimating this equation, for instance, using an ordinary least squares (OLS) approach, would not provide evidence regarding the causal relationship between the extent of off-service placement and the efficiency of care for on-service patients because of the endogeneity of bed placement decisions and several factors that may affect both the level of off-service placement and the efficiency of care for patients placed on service. In what follows, we first describe two potential alternate explanations and the direction of potential bias in each case. Then, we provide details on the specific empirical approach we employ to estimate the spillover effects of off-service placement.

4.1. Potential Alternate Explanations and Biases

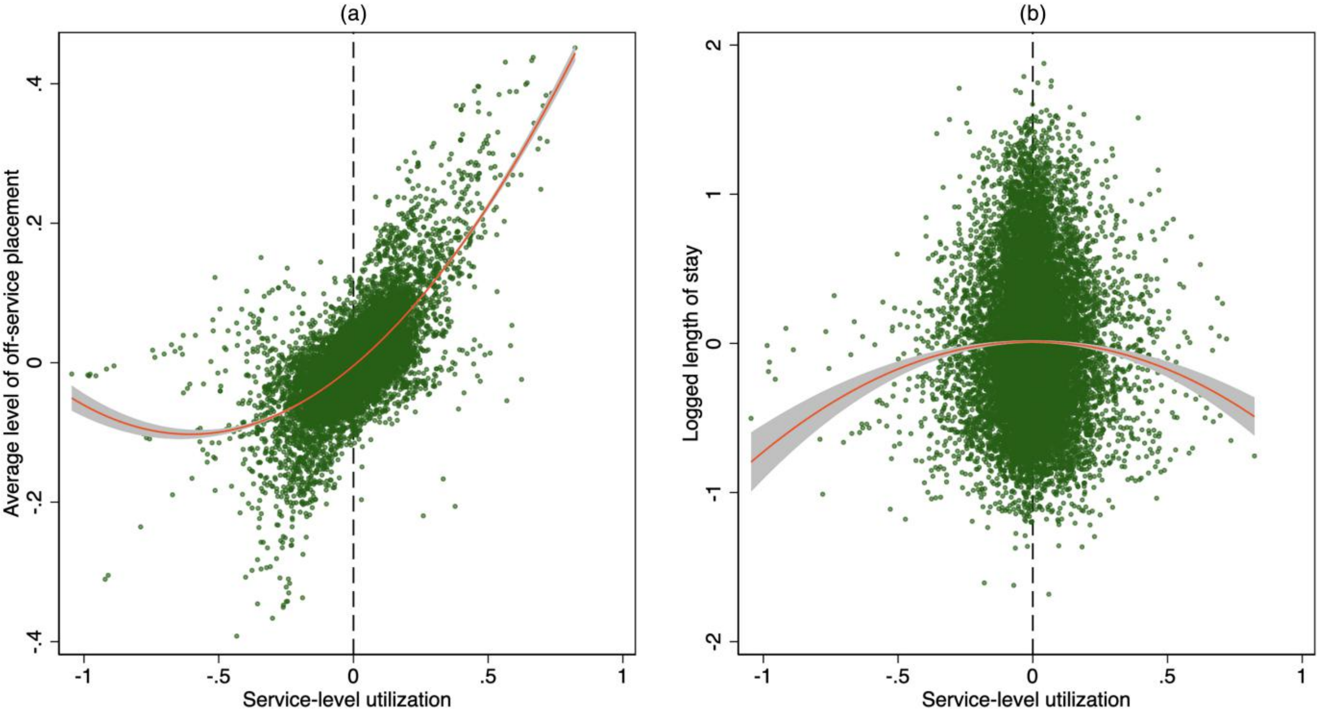
Other than evidence of a spillover effect, there are two potential alternate explanations that could also lead to a positive correlation between the length of stay and the level of off-service placement experienced by the patient, captured by β in Model (1). One alternate explanation relates to supply-side factors, with which periods of high workload can result in high levels of off-service placement as well as delays in care, manifesting as longer lengths of stay. A second alternate explanation relates to demand-side factors, with which patients with higher severity might be selectively placed in on-service beds during periods of high congestion and high levels of off-service placement. We consider each of these alternate explanations and use our data to show that each of the factors would, in fact, downwardly bias the estimation of the true spillover effect such that the true impact would be greater than what a naive OLS estimation of Model (1) would suggest.

4.1.1. Supply-Side Factors. The first alternate explanation considers supply-side factors, with which we would observe a positive correlation between the length of stay and the level of off-service placement if (a) the level of off-service placement increases during periods of high workload and (b) such congestion causes delays in care. Given that the level of off-service placement is closely related to the overall workload of the service's physicians, we focus on understanding the potential bias caused by physician workload. In Figure 1, we plot the residuals of the average level of off-service placement and length of stay against service-level utilization, after partialing out the control variables described in Section 3.5.

Panel (a) shows that off-service placement is utilized during periods of higher congestion (i.e., higher service-level utilization), which is quite intuitive. Panel (b) shows that, during these times of higher utilization, there is a negative correlation between workload and length of stay. In other words, we find the inverted U-shaped relationship identified in Kuntz et al. (2015) and Berry Jaeker and Tucker (2017) such that the speed-up effect

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

Figure 1. (Color online) Average Level of Off-Service Placement and Length of Stay at Different Levels of Service-Level Utilization



Notes. This figure plots the average level of off-service placement in panel (a) and the logged length of stay in panel (b) against service-level utilization. The residuals from partialing out the control variables described in Section 3.5 are plotted as dots. Quadratic models are used for the prediction lines and the 95% confidence intervals.

is dominant during times when off-service placements are most utilized. Hence, the direction of these relationships suggests that an OLS estimation would be biased downward; that is, it would provide an underestimate of the true effect.

4.1.2. Demand-Side Factors. The second alternate explanation considers demand-side factors, with which we would observe a positive correlation between the length of stay and the level of off-service placement if (a) patients with higher severity require longer lengths of stay and (b) patients with increasingly higher levels of severity are placed on service as the level of off-service placement increases. The former is reasonable and can be confirmed empirically; patients who are older, diagnosed with a DRG code with higher cost weights, and have complications or comorbidities have longer lengths of stay (see column (1) of Table A.1 in Online Appendix A). For the latter, our data show that the average severity (measured by DRG cost weights) of on-service patients decreases as the level of off-service placement increases (see column (2) of Table A.1 in Online Appendix A). Intuitively, this is because the extent to which admission controllers can utilize off-service placement becomes more and more limited as the entire system becomes more congested. In other words, as more and more patients are placed off service from a particular service, there is a decrease in the admission controller's ability to exercise discretion in selectively placing sicker patients on service. Therefore, the direction of these relationships also suggests that an OLS result would, in fact, be biased downward and the true spillover effect is larger.

4.2. Heckman Correction Procedure to Address Demand-Side Factors

To rule out the potential alternate explanations and isolate the spillover effect, we develop an estimation approach that allows us to address the biases arising from the factors described. First, we consider the demand-side factors. From previous studies, we know that the decision to place patients in an on- versus off-service bed is a function of patient severity, which is not fully observable to the econometrician (Song et al. 2020). Although we focus in this paper on the population of on-service patients in order to estimate the spillover effects of off-service placement, the selection of patients into the on-service population still poses endogeneity concerns. For example, if the focal service is near full capacity, a relatively sicker patient arriving to the service is more likely to be placed on service than a relatively healthier patient. On the other hand, as we

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

observe in Section 4.1, the average severity of the on-service patients tends to decrease as system-wide congestion increases. In both cases, the severity of the patients admitted to on-service beds is not random, and therefore, the resulting bias must be accounted for.

To correct this bias resulting from the nonrandom selection of patients into on-service beds, we use the two-step Heckman correction procedure (Heckman 1979). Specifically, we first set up a probit selection equation that predicts a given patient's likelihood of being placed on service. Then, we use the inverse Mills ratio (IMR) from the probit selection equation as an additional control variable in the outcome equation. The underlying idea is that the bias from sample selection can be treated as a form of omitted variable bias, and thus, including the IMR as a control variable resolves the selection problem.

The Heckman correction procedure performs best when there are factors that primarily affect the selection but are excluded from the outcome equation (Wooldridge 2010). As such, for our analyses, we utilize the congestion level of units on the receiving end of off-service placements as an instrument. Specifically, we define the primary off-service unit for each service as the unit that was used most frequently by the service for off-service patients. Then, for each on-service patient, we measure the utilization level of that primary off-service unit in the hour prior to the time of admission and use this variable as the instrument. The exclusion restriction is satisfied because the workload levels of the physicians and nurses in other services are unlikely to affect the care delivery for the focal patient, especially conditional on the workload of the physicians in the focal service and nurses in the focal unit. The relevance condition is satisfied because patients are more likely to be placed on service when the primary off-service unit is busier (see Section 4.1, in which we show that the use of off-service placement is limited by the capacity of the units on the receiving end). Using this excluded variable, we estimate the following selection equation using a probit model:

$$\begin{aligned} &P(onservice_i = 1) \\ &\quad = \Phi(\alpha + \delta_1 \cdot Preadmission\ utilization\ of\ primary \\ &\qquad off\text{-}service\ unit_i + \zeta \cdot \mathbf{X}_i). \end{aligned} \tag{2}$$

To estimate this selection equation, we use a sample that includes both on- and off-service patients. Thus, in addition to the 14,787 on-service patient encounters, we include all applicable off-service patient encounters, which yields a sample of 18,211 patient encounters. Further discussions on the distributional assumption and the robustness of the selection model can be found in Online Appendix C.

4.3. Lewbel's (2012) Heteroskedasticity-Based Identification Approach to Address Supply-Side Factors

Whereas the Heckman correction resolves the endogeneity problem caused by nonrandom placement into on-service beds, there is still a lingering identification challenge coming from unobserved factors that may affect both the level of off-service placement and the efficiency of care for a given on-service patient. For example, provider workload and unobserved strategies employed by bed managers may change during particular days of the week or times of the day, affecting both the level of off-service placement as well as the efficiency of care (see Section 4.1). To resolve those potential biases coming from supply-side factors and rule out the second alternate explanation described, we use a heteroskedasticity-based identification approach developed by Lewbel (2012), which has been adopted across multiple disciplines, including some recent work in operations management (e.g., Dong et al. 2020b, Gnanlet et al. 2021, Li et al. 2023). At a high level, Lewbel's (2012) approach exploits the heteroskedasticity present in the first stage to construct valid instrumental variables (IVs) for the endogenous variable. Formally, we can write the system of equations as follows:

$$\begin{aligned} y_i = \alpha + \beta \cdot &Average\ level\ of\ off\text{-}service\ placement_i \\ &+ \gamma \cdot \mathbf{X}_i + \epsilon_{1i}. \end{aligned} \tag{3}$$

$$Average\ level\ of\ off\text{-}service\ placement_i = \alpha + \eta \cdot \mathbf{X}_i + \epsilon_{2i}. \tag{4}$$

In order to construct valid IVs, Lewbel's (2012) approach requires finding a set of $\mathbf{Z} \subseteq \mathbf{X}$ that satisfies two key assumptions: (A1) $cov(Z, \epsilon_1\epsilon_2) = 0$ and (A2) $cov(Z, \epsilon_2^2) \neq 0$. The first assumption (A1) ensures that the information used to construct the heteroskedasticity-based IVs does not come from common elements in ϵ_1 and ϵ_2; that is, Z is uncorrelated with the covariance between the error terms in Equations (3) and (4). The interpretation of the second assumption (A2) is that there is heteroskedasticity in Equation (4) that is correlated with elements in Z. This allows us to construct IVs that are correlated with the endogenous variable—in our case, the average level of off-service placement. Drawing an analogy to the standard IV framework, A1 corresponds to the exclusion restriction, whereas A2 corresponds to the relevance condition. Once the set of valid Z that satisfies the conditions is identified, IVs are constructed by $(\mathbf{Z} - \bar{\mathbf{Z}})\hat{\epsilon}_{2i}$, where $\bar{\mathbf{Z}}$ is the sample mean of $\mathbf{Z}$ and $\hat{\epsilon}_{2i}$ is the estimated residuals from estimating Equation (4). We refer to these as Lewbel IVs.

Conceptually, A1, which requires homoskedasticity with respect to unobserved determinants of length of stay and the level of off-service placement, likely rules

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

out patient characteristics (e.g., age, diagnosis) as valid candidates for Lewbel IVs because these factors are likely related to the amount of care needed by each patient and, therefore, are likely heteroskedastic with respect to length of stay. Similarly, variables that measure staffing levels or workload are also likely to have large and complex effects on the efficiency of care, ruling them out as good candidates. Variables related to the admission characteristics, on the other hand, can potentially be good candidate variables.

For instance, consider a variable that captures whether a patient was admitted during the week versus over the weekend. Because of the operational differences between weekdays and weekend days (e.g., staffing levels), the overall level of off-service placement is also likely different, thus satisfying A2; more precisely, there is a difference in the variance of the residuals conditional on all observables. On the other hand, it is unlikely that the distribution of the unobserved determinants of length of stay would be different across patients who were admitted during the week compared with those who were admitted over the weekend, especially given the extensive set of control variables we include in our models. For instance, for A1 to be invalid, there must be a difference in the variance of unobserved determinants of length of stay between patients who were admitted during the week versus over the weekend but with the same diagnosis, complications, age, gender, levels of workload for physicians and nurses, et cetera. Therefore, we use four variables that capture characteristics related to the time of admission as candidates for Lewbel IVs: whether the patient was admitted during a weekday, the shift during which the patient was admitted, the month in which the patient was admitted, and the average severity of concurrently admitted patients.

Each of the two assumptions described have a testable implication that can be verified using the data. A1 can be verified by applying the Pagan and Hall (1983) test for heteroskedasticity to the system of equations. Failing to reject homoskedasticity with respect to **Z** provides evidence supporting the first assumption (Baum and Lewbel 2019). A2 can be verified by applying any standard heteroskedasticity test, such as the Breusch and Pagan (1979) test, to Equation (4). To satisfy A2, we want to reject homoskedasticity with respect to the selected **Z**. Whereas failing the statistical tests does not necessarily mean that the assumptions are violated (e.g., rejecting homoskedasticity in the test for A1 could be caused by harmless heteroskedasticity of the error term; see Baum and Lewbel 2019), these testable implications have great value in providing support for the overall appropriateness of the Lewbel (2012) approach. Therefore, we select variables that satisfy the statistical tests—namely, weekday admission, the month of admission, and the average severity of concurrently admitted patients—as our set of Lewbel IVs and report the p-values of the two tests for all relevant tables reporting our main estimation results in Section 4.5.[7]

4.4. Combining the Two Approaches

We estimate the spillover effect of off-service placement by combining the two approaches described: the Heckman correction procedure and Lewbel's (2012) heteroskedasticity-based identification approach. We do this by first estimating the selection Equation (2) on the full patient sample that includes both on- and off-service patients. Then, using these estimates, we compute the IMR. Finally, we incorporate the IMR into Lewbel's (2012) approach by including it as an additional control variable in estimating the outcome equation.

In essence, we estimate the following equation to construct Lewbel IVs once the IMR is obtained:

$$\begin{aligned} &\mathit{Average\ level\ of\ off\text{-}service\ placement}_i \\ &\quad = \alpha + \xi \cdot IMR_i + \eta \cdot \mathbf{X}_i + \epsilon_{2i}. \end{aligned} \tag{5}$$

Using the estimated residuals, $\hat{\epsilon}_{2i}$, Lewbel IVs are constructed by $(\mathbf{Z} - \bar{\mathbf{Z}})\hat{\epsilon}_{2i}$. Then, the following system of equations is estimated on the on-service patient sample using two-stage least squares (2SLS):

$$\begin{aligned} &\mathit{Average\ level\ of\ off\text{-}service\ placement}_i \\ &\quad = \alpha + \phi \cdot \mathit{Lewbel\ IVs}_i + \xi \cdot IMR_i + \eta \cdot \mathbf{X}_i + \epsilon_{3i}, \end{aligned} \tag{6}$$

$$\begin{aligned} y_i = {} & \alpha + \beta \cdot \mathit{Average\ level\ of\ \widehat{off\text{-}}service\ placement}_i \\ & + \lambda \cdot IMR_i + \gamma \cdot \mathbf{X}_i + \epsilon_{1i}. \end{aligned} \tag{7}$$

In other words, the average level of off-service placement is instrumented by the Lewbel IVs that are constructed using the heteroskedasticity-based identification approach. The two approaches, along with an extensive list of control variables, resolve the endogeneity problems arising from patient selection as well as other unobserved factors and, thus, allow us to isolate and recover the causal relationship of interest (Das et al. 2003). Because of the complex "first" stage setup, we use the bootstrap method to compute standard errors for the two-stage least squares procedure.

4.5. Main Estimation Results

We begin by estimating the selection Equation (2) using the full sample of on- and off-service patients, for which the results are shown in column (1) of Table 2. The results confirm that the excluded variable used in the Heckman correction procedure strongly predicts the on- versus off-service placement decision in the direction we would expect. We find that patients are less likely to be admitted to an on-service bed when the units on the receiving end of off-service placement are more congested ($\delta = 0.531$, $p < 0.001$). Next, we estimate the first

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

Table 2. Spillover Effects of Off-Service Placement

	(1)	(2)	(3)
	On-service placement	Average level of off-service placement	Logged length of stay
Selection equation			
Preadmission utilization of primary off-service unit	0.531*** (0.114)		
First stage			
Weekday admission[a]		−0.148*** (0.0245)	
Month of admission[a]		Included***	
Severity of concurrently admitted patients[a]		0.000340 (0.00332)	
Second stage			
Average level of off-service placement			1.089* (0.458)
Inverse Mills ratio		0.0143 (0.0129)	0.508*** (0.153)
Model	Probit	2SLS first stage	2SLS second stage
Controls	Yes	Yes	Yes
Observations	18,211	14,787	14,787
F-statistic			62.24
A1 Pagan–Hall test *p*-value			0.377
A2 Breusch–Pagan test *p*-value			0.000

Notes. Heteroskedasticity-robust standard errors for model (1) and bootstrapped standard errors with 10,000 replications for models (2) and (3) in parentheses. Controls for model (1) include age, sex, DRG cost weight, complications and comorbidities, ICU encountered, admit shift, weekday admission, average DRG cost weights of all patients admitted to the same service in five-hour window, admit month, service fixed effects, (service × admit shift) fixed effects, and (service × admit month) fixed effects. In addition to these controls, models (2) and (3) further include service-level utilization, (service-level utilization)^2, average count of service-level patient movements, average DRG cost weights of all patients admitted to the same service, unit-level utilization, (unit-level utilization)^2, number of transfers, and unit fixed effects.

[a] Lewbel instrumental variables constructed by $(\mathbf{Z}-\bar{\mathbf{Z}})\hat{\epsilon}_{2i}$. Month of admission consists of binary indicators for each month of our study period. We omit the coefficients for brevity. Joint χ^2 test indicates high statistical significance ($p < 0.001$).

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

stage Equation (6) of Lewbel's (2012) heteroskedasticity-based two-stage least squares regression using the sample of patients placed on service. We report the results in column (2) of Table 2. We can confirm that the Lewbel IVs jointly are strong predictors of the endogenous variable, which is later further supported by the large *F*-statistic of 62.24 (Table 2, column (3)).

Finally, we estimate the second stage Equation (7) of Lewbel's (2012) approach to obtain the estimates of the spillover effects of off-service placement. We report these results as well as additional statistics that support our use of this approach in column (3) of Table 2. We find that a higher level of off-service placement that a service has during a patient's hospitalization is associated with a longer length of stay for the focal patient. Specifically, the coefficient on the average level of off-service placement indicates that a 10 percentage point increase in the mean of the proportion of patients placed off service leads to a 10.9% increase in length of stay. For the median patient in our sample with 3.04 days of hospitalization, this equates to an eight-hour delay in the overall care process. Given that the standard deviation of the average level of off-service placement among our patient samples is 16 percentage points, we find that the magnitude of the spillover effect is economically and operationally significant.

The magnitude of the estimated spillover effects also confirms the expected direction of bias discussed in Section 4.1. Compared with the OLS results, presented in Table A.3 of Online Appendix A, the magnitude of the spillover effects estimated using the Heckman correction method and the Lewbel IVs is slightly more than double. Column (3) of Table 2 also shows that the coefficient on the IMR is positive and significant. This indicates that there is indeed selection bias arising from the on- versus off-service placement decision, which the Heckman correction is able to address. For the IVs to be valid, we also need to satisfy assumptions A1 and A2 described in Section 4.3. The last two rows of Table 2 report the test statistics for the testable implications of the assumptions. We find that the Pagan–Hall test for A1 returns a *p*-value of 0.377 and the Breusch–Pagan

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

test for A2 returns a p-value less than 0.001. Because we fail to reject the null hypothesis for A1 and reject the null hypothesis for A2, these test results provide evidence that the Lewbel IVs are valid.

5. Alternate Estimation Approaches and Results

In this section, we provide additional evidence supporting our findings by using alternate methods that do not rely on a selection model or IV approach. First, we utilize econometric tools developed to estimate discrete treatment effects: propensity score matching and minimum-biased estimation. Then, we provide more general findings by using a dose-response function estimation approach and a semiparametric estimation approach. We find that our results are highly robust to the alternative estimation methods and specifications.

5.1. Propensity Score Matching

We start by conducting a propensity score matching analysis. In order to facilitate the matching procedure, we discretize the level of off-service placement experienced by a given on-service patient by defining a binary variable indicating whether the focal patient experienced a level of off-service placement throughout hospitalization that is above or below the median compared with the level experienced by all patients in the patient's service. By doing so, we define 7,396 "treated" patients who experienced high levels of off-service placement and 7,391 "control" patients who experienced low levels of off-service placement. To proxy for the information to which admission controllers have access when making placement decisions, we use a number of variables related to physician workload, nurse workload, patient characteristics, admission characteristics, and the patient's service; this allows us to predict each patient's propensity score using a logistic regression model. We report the estimated average treatment effect in Table 3.

The results indicate that the length of stay increases by 10.9%, obtained by exponentiating the estimated coefficient, if a patient experiences a high level of off-service placement. We note that the average level of off-service placement experienced by patients in the treatment (high level) group is 21.8%, and the average level experienced by patients in the control (low level) group is 13.7%. Therefore, the average treatment effect can be interpreted as an 10.9% increase in length of stay when the level of off-service placement increases by 8.1 percentage points. Given our main results presented in Table 2 that estimate a 10.9% increase in length of stay per 10 percentage point increase in the level of off-service placement, which translates to an 8.8% increase in length of stay per 8.1 percentage point increase in the level of off-service placement, we find our main results to be highly robust to the alternative method of propensity score matching.

5.2. Minimum-Biased Estimation

The key assumption underlying propensity score matching is often referred to as the conditional independence assumption. The minimum-biased estimation method (Millimet and Tchernis 2013) aims to minimize the bias that is introduced when the conditional independence assumption does not hold by trading off bias minimization for greater statistical power. This is done by first identifying the propensity score that theoretically minimizes the bias and then estimating the treatment effect using samples within a certain interval, in which a smaller interval leads to lower bias at the expense of lower statistical power. This method has an additional benefit of providing a test for the direction of bias, if present, by starting with a

Table 3. Spillover Effects of Off-Service Placement Using Propensity Score Matching

	(1) Logged length of stay
Experienced high levels of off-service placement	0.104***
	(0.0119)
Model	Propensity score matching
Observations	14,787

Notes. Coefficient reported as an average treatment effect. Robust Abadie–Imbens standard errors in parentheses. Propensity scores computed using a logistic regression model. The variables used in the matching procedure include service-level utilization, (service-level utilization)^2, average count of service-level patient movements, average DRG cost weights of all patients admitted to the same service, unit-level utilization, (unit-level utilization)^2, age, sex, DRG cost weight, complications and comorbidities, ICU encountered, number of transfers, admit shift, weekday admission, average DRG cost weights of all patients admitted to the same service in five-hour window, admit month, and service fixed effects.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

Table 4. Spillover Effects of Off-Service Placement Using the Minimum-Biased Estimator

	(1) Logged length of stay (MB)	(2) Logged length of stay (MB-EE)
25% sample of treatment and control groups	0.111 (0.085, 0.136)	0.111 (0.084, 0.137)
20% sample of treatment and control groups	0.111 (0.081, 0.139)	0.110 (0.081, 0.138)
15% sample of treatment and control groups	0.124 (0.083, 0.155)	0.125 (0.081, 0.155)
10% sample of treatment and control groups	0.145 (0.090, 0.178)	0.143 (0.089, 0.179)
5% sample of treatment and control groups	0.179 (0.090, 0.225)	0.178 (0.090, 0.226)

Notes. MB, minimum-biased estimator; MB-EE, minimum-biased estimator with Edgeworth expansion. Ninety-five percent confidence intervals in parentheses are obtained using 1,000 bootstrap repetitions; x% sample corresponds to θ defined in Millimet and Tchernis (2013). As θ decreases, a smaller sample is chosen in favor of reducing bias at the expense of statistical power. MB-EE relaxes the joint normality assumption. Covariates include service-level utilization, (service-level utilization)^2, average count of service-level patient movements, average DRG cost weights of all patients admitted to the same service, unit-level utilization, (unit-level utilization)^2, age, sex, DRG cost weight, complications and comorbidities, ICU encountered, number of transfers, admit shift, weekday admission, average DRG cost weights of all patients admitted to the same service in five-hour window, admit month, and service fixed effects.

wider interval and observing how the estimated treatment effect changes as the interval narrows.

As seen in Table 4, we find that the estimates recovered through the propensity score matching estimator are robust to the possibility of the conditional independence assumption not being met. In fact, the results from the propensity score matching analysis serve as the lower bound of the treatment effect as indicated by the direction in which the minimum-biased estimation results change as the sample interval becomes more and more restrictive. This provides additional evidence supporting our expectation of the direction of the potential bias (see Section 4.1) and the main results (see Section 4.5). Furthermore, results from the extension of the minimum-biased estimator with Edgeworth expansion suggest that our findings are also robust to the distributional assumption made by the minimum-biased estimator (Millimet and Tchernis 2013).

5.3. Dose-Response Function Estimation

The propensity score matching and minimum-biased estimation methods both require a discretized form of the level of off-service placement, which results in some degree of information loss. Similar to our main model, we can use a generalized propensity score method that estimates a dose-response function in order to estimate the effect of a continuous treatment variable (Hirano and Imbens 2004). This method, however, relies on distributional assumptions; Hirano and Imbens (2004) assume normality, and Guardabascio and Ventura (2014) later propose a method that relaxes this assumption to consider a family of exponential distributions, neither of which fit our data well. With this caveat in mind, we nevertheless find a positive and significant relationship between the length of stay and the level of off-service placement experienced by on-service patients. We present the results in Figure A.2 of Online Appendix A.

5.4. Semiparametric Estimation

Building on the dose-response function estimation, we also examine whether the spillover effects of off-service placement vary in the average level of off-service placement using a nonlinear model. To allow for a high degree of flexibility in our model, we follow Robinson (1988) to perform a semiparametric analysis in which the main explanatory variable is assumed to have a nonparametric functional form. To address the endogeneity concern, we use a control function approach in which the residuals from the first stage equation are included as a control variable in the second stage regression. In effect, we estimate the following second stage equation:

$$\begin{aligned} y_i = \alpha + G(&\mathit{Average\ level\ of\ off\text{-}service\ placement}_i) \\ &+ \lambda \cdot IMR_i + \gamma \cdot \mathbf{X}_i + \mu \cdot \hat{\epsilon}_{3i} + \epsilon_{1i}. \end{aligned} \tag{8}$$

The nonparametric function, G, is estimated using a kernel regression, and $\hat{\epsilon}_{3i}$ is estimated by the residuals from the first stage Equation (6) using Lewbel IVs. Estimation results are presented in Figure A.3 of Online Appendix A. The results are consistent in finding significant spillover effects of off-service placement. Furthermore, the semiparametric analysis suggests that the spillover effects are generally linear.

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

6. Potential Mechanisms for the Spillover Effects of Off-Service Placement

Our discussions with several practitioners and hospital administrators suggest there may be two potential mechanisms that may underlie the negative spillover effects of off-service placement: (a) challenges in coordination and (b) the physical distance. Our analyses yield strong support for the former (presented subsequently) and no statistical evidence of the latter (see Online Appendix D).

When more patients are placed off service, it becomes increasingly challenging to coordinate care between physicians and nurses, even for those patients who are on service, because routines become interrupted. For example, physicians and nurses typically hold daily morning meetings to discuss the care plan for each patient, and the routine process is to hold this discussion for patients who are on service. The interruptions that arise when trying to replicate this process for off-service patients (e.g., nurses trying to reach the physician and the physician leaving the main unit to reach the off-service unit) leads to a disruption in the provider workflow, which may cause delays in care for all patients. Furthermore, whereas providers who are jointly responsible for a given patient's care can engage in fluid and continuous communication throughout the day when they are colocated (as they are when caring for an on-service patient), when they are jointly responsible for an off-service patient's care, it takes much more time and effort to identify who else is involved in the patient's care, travel to the other unit to communicate with the other provider in person, or incur delays arising from asynchronous forms of communication (e.g., pages, text messages, emails). As such, the ad hoc nature of off-service placement can lead to delays not only for off-service patients themselves, but can also permeate to the care delivered to all other patients who belong to the service.

To examine this possibility, we employ a fuzzy regression discontinuity design (RDD). The intuition behind the design is as follows. Consider a variable x that captures the average hourly number of off-service units that are being utilized by the focal on-service patient's service during hospitalization. In other words, we track the number of off-service units that are used by the focal patient's service during each hour of the patient's hospitalization and then take the average of these hourly snapshots across the patient's entire hospitalization. Suppose that the value of this variable, x, is 0.95 for a particular patient. This suggests that it is possible that only one off-service unit was used by the service during the patient's hospitalization. Compare this patient to another patient for whom this value is $x = 1.05$. For this latter patient, at least two units must have been engaged for off-service placement by that service at some point during the patient's hospitalization. Leveraging the integer threshold—1.00 in this example—allows us to employ a fuzzy RDD design because (a) it is possible that only one unit was engaged for off-service placement during the former patient's hospitalization, whereas (b) at least two units must have been engaged during the latter's hospitalization. By examining samples near the integer thresholds, we can then estimate the effect of engaging additional off-service units during a focal patient's hospitalization. Using our data, we illustrate this graphically in Figure A.4 of Online Appendix A.

For the fuzzy RDD analysis, we define the running variable, $\tilde{x}$, by taking the average number of off-service units, x, and subtracting its nearest integer, $\hat{x}$. When $\tilde{x}$ is strictly greater than zero, at least $\hat{x} + 1$ units must have been involved for off-service placement during the focal patient's hospitalization. When $\tilde{x}$ is less than or equal to zero, there is some probability that only $\hat{x}$ units were involved during the focal patient's hospitalization. Because this is a fuzzy RDD setting, the running variable serves as an instrument for the true treatment. We identify the treatment status by using historical data to identify whether the maximum number of off-service units used during each patient's hospitalization is strictly greater than $\hat{x}$, the nearest integer of the average number of off-service units used. We utilize the fuzzy RDD setup by running a 2SLS regression in which treatment status is instrumented by the binary variable indicating whether $\tilde{x}$ is strictly greater than zero. We drop observations for which $\tilde{x}$ is exactly zero because encounters with exactly the same number of off-service units used throughout the hospitalization can be different in unobservable ways. We present the results of this estimation in Table 5.

In column (1), we see that the running variable is a strong predictor of the treatment. In column (2), we find that the treatment effect is 0.149; that is, we find that the length of stay increases by 16% (obtained by exponentiating the coefficient) if one or more additional units are used for off-service placement during the patient's hospitalization ($p = 0.080$). To assess whether this effect is the result of an increase in coordination costs as opposed to a greater number of off-service patients, we check for any discontinuity in the average level of off-service placement across the integer thresholds. As we see in Figure A.5 of Online Appendix A, we do not observe any such discontinuity in the level of off-service placement. Furthermore, we conduct an additional analysis in which we directly account for the average level of off-service placement as an additional control variable. Our findings remain consistent, and we report the results in Table A.4 of Online Appendix A. We note that the results presented here use the optimal bandwidth computed using the approach described in Calonico et al. (2014). For robustness, we present the results using different bandwidth choices

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

Table 5. Spillover Effects of an Additional Off-Service Unit

	(1) Treatment	(2) Logged length of stay
$\tilde{x}$ is greater than zero	0.392***	
	(0.0229)	
Treatment: additional off-service unit used during hospitalization		0.149^{+}
		(0.0852)
Observations	12,346	3,828

Notes. Standard errors in parentheses. Local linear regression used to construct the point estimator. Optimal bandwidth = 0.155 selected using approach described in Calonico et al. (2014). Covariates include service-level utilization, (service-level utilization)^2, average count of service-level patient movements, average DRG cost weights of all patients admitted to the same service, unit-level utilization, (unit-level utilization)^2, age, sex, DRG cost weight, complications and comorbidities, ICU encountered, number of transfers, admit shift, weekday admission, average DRG cost weights of all patients admitted to the same service in five-hour window, admit month, service fixed effects, unit fixed effects, (service × admit shift) fixed effects, and (service × admit month) fixed effects.

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

and degrees of polynomials in Table A.5 of Online Appendix A, in which we find rather consistent results across different bandwidth choices.

7. Counterfactual Analyses

So far, we focus on identifying, quantifying, and understanding the spillover effects of off-service placement. Our analyses suggest that, not only does off-service placement have a negative first order effect on those patients who are placed off service, it also has a negative spillover effect on patients who are placed on service. In what follows, we conduct a series of counterfactual analyses to, first, identify alternate routing policies that may be successful in mitigating the spillover effects of off-service placement and to estimate the magnitude of the potential gains from adopting each of the policies. Next, we assess the degree of under-capacity and misallocation of capacity by conducting another set of counterfactual analyses in which we consider alternate capacity-planning policies. Specifically, we evaluate the impact of adding capacity to the hospital as well as reassigning a unit from the most underutilized service to the most overutilized service. These analyses may help hospital administrators determine ways in which they can continue to realize the benefits of capacity pooling and minimize their negative spillover effects.

7.1. Setup

We leverage the granularity of our data and perform simulations using the actual admissions, transfers, and discharges of patients in the observed data.[8] Similar to the approach used in Bertsimas and Pauphilet (2023), we assume our data provide an accurate representation of the underlying data-generating process, which includes the patient flow, the severity of different types of patients, and the service rates for patients as a function of both observable and unobservable factors.

To begin the simulation, we take a snapshot of the hospital on the first day, and then, we simulate patient movements under alternate routing and capacity-planning policies whenever patients move into and out of medical/surgical units. Throughout the simulation, we carefully track the utilization levels of each unit and each service at the hour level. Because nonmedical/surgical units (e.g., intensive care units, observation beds) are beyond the scope of this study, we route movements into and out of those units to reflect the observed movements. Once all patients are rerouted using an alternate routing policy, we record all on- versus off-service placement decisions for each patient and the location of all patients for each hour. These hourly snapshots are then used to calculate the extent of the (counterfactual) off-service placement for each service for each hour. Next, we compute the average levels of off-service placement experienced by each on-service patient in exactly the same way that they were calculated using the observed data for the empirical analyses. Finally, we calculate the predicted counterfactual lengths of stay for all patients in our sample. For off-service patients experiencing the first order effects of off-service placement, we compute the first order effects of off-service placement using the same methodology used in Song et al. (2020) and include the estimates in Table F.1 of Online Appendix F. For on-service patients experiencing the spillover effects, we use the estimates presented in column (3) of Table 2. For robustness, we additionally consider service type-specific spillover effects and present the estimates as well as the resulting counterfactual results in Online Appendix E.

7.2. Alternate Routing Policies

7.2.1. On Until Zero Beds. Bed managers in the hospital often keep some number of on-service beds reserved in anticipation of future admissions or transfers, especially when the number of open beds remaining on service is small and the arriving patient seems to have a relatively low level of severity. This first policy does not allow for on-service beds to be left unoccupied in

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

anticipation of future demand. Instead, all arriving patients are placed in on-service beds until there are no more beds available. Once all on-service beds are exhausted, patients are placed in the service's primary off-service unit; once that is full, they are placed in the service's secondary off-service unit. For each service, we define a primary and secondary off-service unit by using the observed data and identifying the two units where the greatest number of off-service patients were placed; these are shown in Figure F.1 of Online Appendix F. Restricting the number of units across which off-service patients are placed could help reduce coordination costs between physicians and nursing teams. In cases when a service has multiple on-service units, we prioritize placements into units that do not serve as a designated off-service unit for another service to ensure that as many patients as possible are placed on service. Once both primary and secondary off-service units have been exhausted, patients are placed in any medical/surgical unit with the greater number of available beds.

7.2.2. On Until Zero Beds + Boarding. One way to reduce the incidence of off-service placement is to allow for some additional boarding time prior to admission into a medical/surgical bed. Whereas excessive time spent boarding is associated with undesirable outcomes (Chalfin et al. 2007, Rabin et al. 2012, Mathews and Long 2015), if it helps avoid an off-service placement, the combined benefits may outweigh the losses. Because bed managers can observe which beds are expected to become available in the next few hours (e.g., because of an expected discharge or expected transfer), if they are given the discretion to delay the bed placement of an incoming patient when an on-service bed is expected to become available soon, it may be possible to place the patient on service and avoid an off-service placement. Based on discussions with clinical leaders at our study hospital and to err on the side of being too conservative, we allow for only one hour of additional boarding time.

7.2.3. On Until Zero Beds + Earlier Discharge. Rather than allowing for additional boarding, which aims to reduce off-service placements by delaying admissions, another option is to expedite discharges. In other words, hospitals could prioritize discharging patients earlier in the day (i.e., morning versus afternoon or evening) to facilitate admissions into on-service beds (Shi et al. 2016). Benson et al. (2006) find that 12% of all surgical patients in a UK hospital experienced delays in discharge despite being medically fit to leave. When simulating this policy, we assume that patients who were discharged between 11 a.m. and 5 p.m. could have been discharged at 11 a.m. if physicians and nurses had reorganized their days to prioritize discharges; these times were also determined based on our discussions with clinical leaders at our study hospital. When discharges occur earlier in the day, patients arriving in the afternoon who would otherwise be placed in an off-service bed could be placed on service.

7.2.4. On Until Zero Beds + Hospital-Wide Flex Units. Rather than designating primary and secondary off-service units for each service, this policy designates two units as hospital-wide flex units. From the observed data, we identify the two units that received the highest number of off-service patients and designate these as the two hospital-wide flex units. These two units no longer serve as on-service units for a specific service; instead, they provide off-service care for off-service patients across all services. Under this policy, each incoming patient is placed on service as long as there is an available on-service bed. Once there are no more on-service beds available, patients are sent to the flex unit with the most available beds. Once both flex units become full, patients are then placed in any medical/surgical unit with the most available beds.

7.2.5. On Until a Few Beds. This policy seeks to mimic the behavior that was prohibited in the three preceding policies: reserving on-service beds in anticipation of future arrivals. Here, we allow patients to be placed off service when there are fewer than five on-service beds remaining. To simulate the bed managers' decisions, we use the following approach to calculate probabilities from the observed data, which allow us to determine the likelihood of a given patient being placed off service. First, for each service, we calculate the proportion of patients who were placed off service given x open beds at the time of their arrival ($x \leq 5$). For example, if there were 100 admissions to the general surgery service when one general surgery bed was open and 30 of these patients were placed off service, the probability of an incoming general surgery patient being placed off service when exactly one on-service bed is available is 0.3. Then, using these probabilities, if a patient is assigned to be placed off service, we first route them to the primary off-service unit and then to the secondary off-service unit. We show the set of computed probabilities for each service in Table F.2 of Online Appendix A.

7.2.6. On Until a Few Beds + Protected Services. Across the different services within the hospital, some tend to have higher levels of their patients placed off service than others. Specifically at our study hospital, the cardiac surgery, east surgery, oncology, and transplant services are specialized services that try to minimize the incidence of placing their patients in off-service beds; this can be seen in Figure F.2 of Online Appendix F. In some cases, this is because of licensing restrictions.[9] As another alternate routing policy, we designate these four services as "protected services" and

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

minimize the chances of their needing to place their own patients off service by restricting off-service patients from other services from flowing into units that are designated to them. In other words, on-service units for the protected services do not serve as designated off-service units for other services, and patients in the protected services are always placed on service as long as there is an open bed. For nonprotected services, we continue to implement the "on until a few beds" policy. This alternate routing policy is the strictest policy that we test in the sense that it explicitly reduces the level of capacity pooling in an attempt to reduce off-service placements and their associated spillover effects.

7.3. Alternate Capacity Planning Policies

7.3.1. Add One Unit to Cardiac Medicine or General Medicine. Ultimately, the need for off-service placement comes from the limited capacity of on-service beds. Therefore, one intuitive solution that can mitigate both the first order and spillover effects of off-service placement is to add capacity to the most heavily utilized service. As an alternate capacity-planning policy, we add a 15-bed unit to the cardiac medicine service, which was the service with the highest rate of off-service placement at the study hospital during the study period. We also consider another policy by which we instead add the 15-bed unit to the general medicine service, which was the service with the greatest number of patients placed off service. Each of these two policies aims to end the vicious cycle of off-service placement; placing patients off service leads to reduced capacity for the receiving service, which, in turn, leads to an increased need for the receiving service to place its own patients off service. By adding more capacity to the most heavily utilized services, we evaluate how much off-service placement can be eliminated by leveraging additional capacity and the extent to which this would yield gains in productivity.

7.3.2. Reassign One Unit to General Medicine. Whereas simply adding capacity could be the most straightforward solution, adding beds can be very costly. The hospital must invest not only in equipment (e.g., hospital beds, medical devices), but also providers and space. A more cost-effective solution is to reassign a unit from an underutilized service to the service with the highest levels of congestion. Using historic census data, hospital managers can identify the appropriate services and units for such reassignment. For our study hospital, the east surgery service was the most underutilized service. Therefore, for this alternate capacity-planning policy, we estimate the effect of reassigning a unit from the east surgery service to the general medicine service, which historically had the greatest number of off-service patients. For all three capacity-related policies, we use the "on until zero beds" routing policy to simulate patient movements.

7.4. Simulation Results

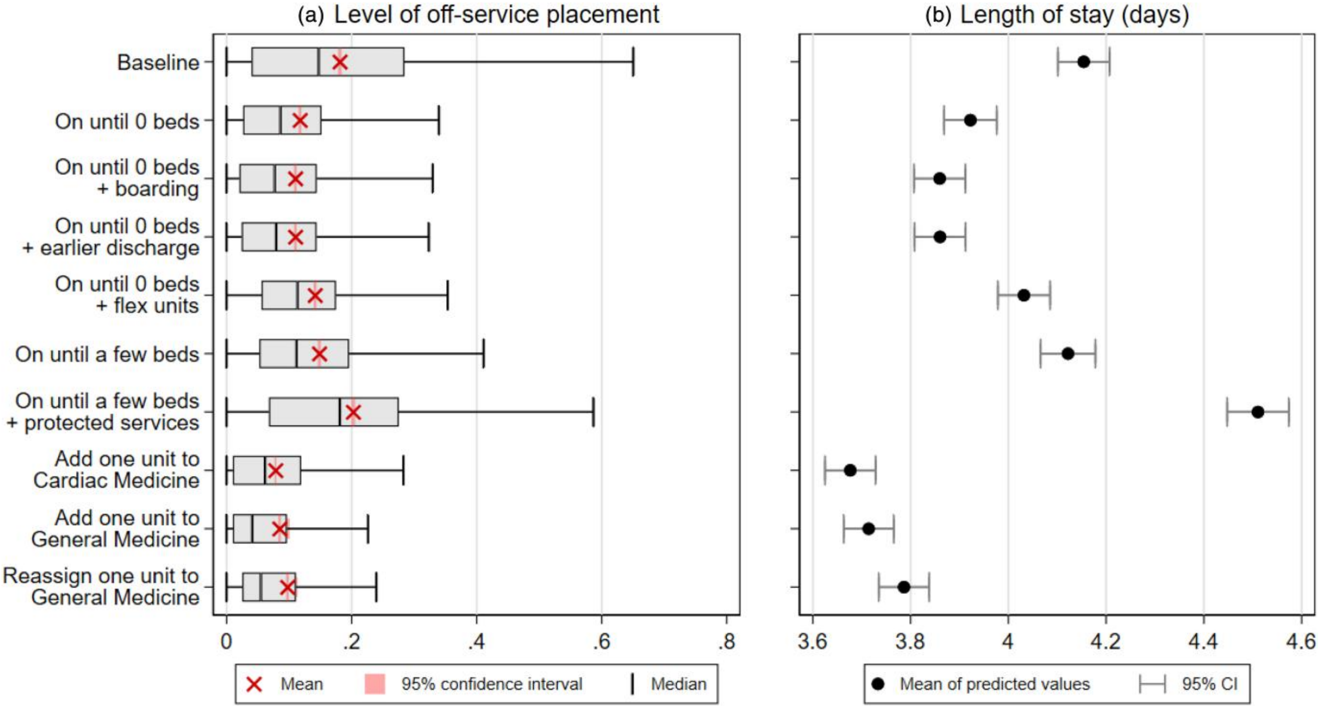
The counterfactual results derived from our simulations illustrate that both alternate routing and capacity-planning policies could indeed reduce the overall level of off-service placement and, in turn, result in reductions in the average patient length of stay. The boxplots in the first panel of Figure 2 show the counterfactual average levels of off-service placement experienced by on-service patients for each of the alternate routing and capacity policies we consider. The top row labeled "Baseline" reflects the observed data. The subsequent rows represent each of the alternate routing and capacity-planning policies described in Sections 7.2 and 7.3. The second panel of Figure 2 plots the sample mean and 95% confidence interval of the average length of stay for all patients, including both on- and off-service patients.

Based on the counterfactual outcomes from implementing the "on until zero beds" policy, we see that the practice of reserving on-service beds in anticipation of a future arrival is one of the drivers that increases the overall level of off-service placement. Significant reductions in off-service placement can be achieved by restricting bed managers from reserving on-service beds and allowing them to place patients off service only if there is no available bed on service. That said, the observed behavior of reserving beds could be a result of the bed manager having private information (unobservable to the researcher) about upcoming patient arrivals. Thus, the true gains from implementing a "on until zero beds" policy may be more muted. Nevertheless, our results suggest that the practice of reserving beds should be limited and only utilized when absolutely necessary.

We consider boarding and earlier discharge as an add-on to the "on until zero beds" policy. We find that boarding a patient for an extra hour when an on-service bed is anticipated to become available may be an effective policy that could reduce both the overall level and the volatility of off-service placement. In practice, boarding patients who are transferred from other areas of the hospital (e.g., the emergency department) demands additional resources from those areas; thus, the overall impact on the entire system must be considered. For patients who are being admitted directly, boarding the patient for a little more time until an on-service bed becomes available as opposed to admitting the patient as soon as possible into an off-service bed can be an effective solution. The benefits of additional boarding, however, must be weighed against potential clinical concerns in delaying care provision (Chalfin et al. 2007, Rabin et al. 2012, Mathews and Long 2015). Earlier discharge also seems to reduce the

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

Figure 2. (Color online) Simulation Results



Notes. This figure reports the counterfactual level of off-service placement and the resulting average length of stay. Panel (a) presents the box-plots of the average levels of off-service placement for all on-service patients. Each box plot shows, for the corresponding measure and routing policy, the sample mean and its 95% confidence interval, the median, the first and third quartile, and the minimum and maximum (excluding outliers). Panel (b) presents the average length of stay for all patients including both on- and off-service patients.

overall level of off-service placement and yields improvements with respect to the efficiency of care.

Next, the outcomes from implementing the "on until zero beds + hospital-wide flex units" and the "on until a few beds + protected services" policies provide insight into why careful planning of capacity pooling is crucial. Whereas the two policies differ in their approach, neither policy successfully reduces the overall level of off-service placement compared with the corresponding policies that do not restrict the flow of off-service placement because each fails to sufficiently account for the capacity constraints of the units that are on the receiving end of off-service placements. Our simulations illustrate that cordoning off a set of protected services leads to situations in which patients arriving to the other services experience high rates of off-service placement for sustained periods of time because of limited on-service capacity for prolonged periods. Similarly, with two hospital-wide flex units, patients experienced increases in the level of off-service placement compared with the "on until zero beds" policy because of the capacity constraints imposed on each of the two services that, in effect, lost one on-service unit. Furthermore, there were frequent periods of time when the two flex units did not provide enough capacity to serve off-service patients from all eight services.

We find that all three capacity planning–related policies are very successful in reducing the average level of off-service placement and, therefore, in improving the efficiency of care. Adding a 15-bed unit to the cardiac or general medicine service resulted in significant reductions in the average level of off-service placement compared with both the "on until zero beds" policy as well as the baseline (observed) data. Interestingly, the policy that reassigned one unit from east surgery, the most underutilized service, to general medicine, the service with the greatest number of off-service patients, also resulted in significant improvements. This suggests that careful capacity planning and continuous management and oversight may significantly reduce the negative consequences of off-service placement.

8. Discussion and Conclusions

In this paper, we investigate whether and to what extent there is a spillover effect of off-service placement that impacts the efficiency of care for patients who are placed on service. Using multiple empirical methods and methodological approaches, we find that on-service patients experience substantial negative spillover effects from off-service placement. Our analyses suggest that challenges in coordination between physicians and nursing

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

teams in off-service units may be an important driver of this effect. Our counterfactual results from the simulation studies provide insights into which other routing and capacity planning–related policies may be effective in reducing the incidence of off-service placement and improving outcomes for all patients regardless of their placement location. Our findings suggest that hospitals should limit the practice of reserving on-service beds in anticipation of future arrivals of sicker patients and they should place off-service patients across fewer units in order to minimize coordination costs between physicians and nurses. Policies such as earlier discharge initiatives, which do not target off-service placement specifically, but are designed to improve overall efficiency, can also lead to meaningful improvements by allowing more patients to be placed on service. In cases when the bed manager has visibility into upcoming discharges, another possibility would be to board a patient a little longer when an on-service bed is expected to open up soon. Of course, each of these policies must be carefully considered by weighing the benefits of reducing off-service placement against the potential costs incurred by doing so (e.g., increased boarding time). In terms of policies around capacity planning and management, we find that shifting hospital beds from underutilized services to overutilized services can be an effective strategy when strictly adding capacity (i.e., new beds) is not feasible.

Our findings have important managerial implications. For hospital administrators, this work further highlights the importance of better managing the practice of off-service placement, which is widespread among hospitals all around the world (e.g., Shi et al. 2016, Stylianou et al. 2017). Our findings illustrate that the effects of off-service placement reach well beyond those patients who are placed off service and, rather, impact all patients throughout the hospital; this highlights the need to reexamine the way in which hospitals leverage various capacity management strategies.

Our work opens up several avenues for further investigation of capacity-pooling strategies and their implications. Although our findings are robust to several alternate specifications, our data come from a single hospital. Given the widespread use of off-service placement, we invite other researchers to study these effects of off-service placement in different settings to provide external validation of our findings. Whereas we focus on the spillover effects on the efficiency of care in this work, studying the unintended consequences on the quality of care is also highly relevant for hospital managers. Furthermore, evidence on when exactly the delays occur—for example, the time between admission and the first diagnostic test, the time between when the result of the test becomes available and when the physician's follow-up consultation occurs—will help us better understand the intricate effects of off-service placement.

Methodologically, our study relies on the validity of the methods we use to address the endogeneity concern. Although we provide evidence supporting our findings by verifying the necessary assumptions for the IVs as well as by using multiple alternate methods, a setting that allows for random assignment of patients to on- versus off-service beds would be more robust. Future work could also extend our counterfactual analyses by moving beyond the first order effect of the alternate routing policies. When considering the potential implications for throughput, the reduction in the counterfactual length of stay represents a lower bound of the potential gains because policies that decrease patients' average length of stay would allow the hospital to admit additional patients and, thereby, increase throughput.

Hospitals continue to innovate to find ways to improve their efficiency and performance when operating with limited capacity. Understanding and addressing the various challenges surrounding the practice of placing patients off service will help hospital administrators implement better capacity-management practices and improve the efficiency and quality of care.

Acknowledgments

The authors thank Ryan Graue; Sarah Moravick; seminar participants at Baruch College, Dartmouth College, Johns Hopkins University, Michigan State University, University of Chicago, University of Miami, and Yale University; as well as participants of the INFORMS Annual Meeting, INFORMS Healthcare Conference, MSOM Conference, POMS Annual Conference, and Southeastern Health Economics Study Group. The authors gratefully acknowledge constructive feedback from the editors and reviewers throughout the review process.

Endnotes

[1] We provide a more detailed description of service- versus unit-level operations in Section 3.1.

[2] For interested readers, we report results from estimating unit-level spillover effects in Online Appendix B.

[3] East and west surgery are groups of relatively smaller surgical specialties.

[4] Exceptions include hospitalists, who specialize in providing inpatient care rather than a specific type of condition or a body part, and oncology nurses, who specialize in administering chemotherapy and other cancer treatments.

[5] For brevity, henceforth, we refer to this as the "average severity of concurrently admitted patients."

[6] These variables are computed using the full data set prior to applying the exclusions described in Section 3.2 to construct the analysis sample.

[7] For completeness, we also conduct our analyses using different combinations of the Lewbel IVs as a robustness check and present the results in Table A.2 of Online Appendix A.

[8] Another approach would be to construct a queueing model consisting of arrival and service rates that are functions of the spillover effects

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.

we estimate. Whereas such an approach provides flexibility to simulate virtually any counterfactual policies, its accuracy is limited by the many necessary assumptions underlying the model. We bypass these concerns regarding the extent to which such assumptions would represent the real world by using a data-based approach.

[9] For example, administering chemotherapy requires special nursing training and licensing, so patients who are admitted for cancer treatment are always placed on service in a unit designated for the oncology service.

References

Alameda C, Suárez C (2009) Clinical outcomes in medical outliers admitted to hospital with heart failure. *Eur. J. Internal Medicine* 20(8):764–767.

Bai AD, Srivastava S, Tomlinson GA, Smith CA, Bell CM, Gill SS (2018) Mortality of hospitalised internal medicine patients bed-spaced to non-internal medicine inpatient units: Retrospective cohort study. *BMJ Quality Safety* 27(1):11–20.

Baum CF, Lewbel A (2019) Advice on using heteroskedasticity-based identification. *Stata J.* 19(4):757–767.

Benson R, Drew J, Galland R (2006) A waiting list to go home: An analysis of delayed discharges from surgical beds. *Ann. Roy. College Surgeons England* 88(7):650–652.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Sci.* 63(4):1042–1062.

Bertsimas D, Pauphilet J (2023) Hospital-wide patient flow optimization. *Management Sci.*, ePub ahead of print September 25, https://doi.org/10.1287/mnsc.2023.4933.

Breusch TS, Pagan AR (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47(5):1287–1294.

Calonico S, Cattaneo MD, Titiunik R (2014) Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6):2295–2326.

Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP (2007) Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* 35(6):1477–1483.

Chan CW, Green LV, Lekwijit S, Lu L, Escobar G (2019) Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units. *Management Sci.* 65(2):751–775.

Dai JG, Shi P (2019) Inpatient overflow: An approximate dynamic programming approach. *Manufacturing Service Oper. Management* 21(4):894–911.

Dai JG, Shi P (2021) Recent modeling and analytical advances in hospital inpatient flow management. *Production Oper. Management* 30(6):1838–1862.

Das M, Newey WK, Vella F (2003) Nonparametric estimation of sample selection models. *Rev. Econom. Stud.* 70(1):33–58.

Dobson G, Pinker E, Van Horn RL (2009) Division of labor in medical office practices. *Manufacturing Service Oper. Management* 11(3): 525–537.

Dong J, Perry O (2020) Queueing models for patient-flow dynamics in inpatient wards. *Oper. Res.* 68(1):250–275.

Dong J, Shi P, Zheng F, Jin X (2020a) Structural estimation of load balancing behavior in inpatient ward network. Working paper, Columbia Business School, New York.

Dong Y, Skowronski K, Song S, Venkataraman S, Zou F (2020b) Supply base innovation and firm financial performance. *J. Oper. Management* 66(7–8):768–796.

Gesensway D (2010) Having problems finding your patients? Accessed December 26, 2023, https://www.todayshospitalist.com/having-problems-finding-your-patients/.

Gnanlet A, Sharma L, McDermott C, Yayla-Kullu M (2021) Impact of workforce flexibility on quality of care: Moderating effects of workload and severity of illness. *Internat. J. Oper. Production Management* 41(12):1785–1806.

Guardabascio B, Ventura M (2014) Estimating the dose-response function through a generalized linear model approach. *Stata J.* 14(1):141–158.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Hirano K, Imbens GW (2004) Applied Bayesian modeling and causal inference from incomplete-data perspectives. Gelman A, Meng X, eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, Wiley Series in Probability and Statistics, No. 2001 (John Wiley & Sons Ltd., Hoboken, NJ), 73–84.

Izady N, Mohamed I (2021) A clustered overflow configuration of inpatient beds in hospitals. *Manufacturing Service Oper. Management* 23(1):139–154.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

Kim SH, Zheng F, Brown J (2023) Identifying the bottleneck unit: Impact of congestion spillover in hospital inpatient unit network. *Management Sci.*, ePub ahead of print August 22, https://doi.org/10.1287/mnsc.2023.4887.

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.

Kohn R, Harhay MO, Bayes B, Song H, Halpern SD, Kerlin MP, Greysen SR (2021) Influence of bedspacing on outcomes of hospitalised medicine service patients: A retrospective cohort study. *BMJ Quality Safety* 30:116–122.

Kohn R, Harhay MO, Weissman GE, Anesi GL, Bayes B, Song H, Halpern SD, Greysen SR, Kerlin MP (2020) The association of geographic dispersion with outcomes among hospitalized pulmonary service patients. *Ann. Amer. Thoracic Soc.* 17(2):249–252.

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

Lewbel A (2012) Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *J. Bus. Econom. Statist.* 30(1):67–80.

Li Y, Wang X, Gong T, Wang H (2023) Breaking out of the pandemic: How can firms match internal competence with external resources to shape operational resilience? *J. Oper. Management* 69(3):384–403.

Mathews KS, Long EF (2015) A conceptual framework for improving critical care patient flow and bed use. *Ann. Amer. Thoracic Soc.* 12(6):886–894.

Meng L, Batt R, Terwiesch C (2021) The impact of facility layout on worker behavior: An empirical study of nurses in the emergency department. *Manufacturing Service Oper. Management* 23(4):819–834.

Millimet DL, Tchernis R (2013) Estimation of treatment effects without an exclusion restriction: With an application to the analysis of the school breakfast program. *J. Appl. Econometrics* 28(6):982–1017.

Pagan AR, Hall AD (1983) Diagnostic tests as residual analysis. *Econometric Rev.* 2(2):159–218.

Rabin E, Kocher K, McClelland M, Pines J, Hwang U, Rathlev N, Asplin B, Trueger NS, Weber E (2012) Solutions to emergency department "boarding" and crowding are underused and may need to be legislated. *Health Affairs* 31(8):1757–1766.

Reagans R, Argote L, Brooks D (2005) Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management Sci.* 51(6):869–881.

Robinson PM (1988) Root-*N*-consistent semiparametric regression. *Econometrica* 56(4):931–954.

Shi P, Chou MC, Dai JG, 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.

Stylianou N, Fackrell R, Vasilakis C (2017) Are medical outliers associated with worse patient outcomes? A retrospective study within a regional NHS hospital using routine data. *BMJ Open* 7:e015676.

Wooldridge J (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (MIT Press, Cambridge, MA).

Xie J, Zhuang W, Ang M, Chou MC, Luo L, Yao DD (2021) Analytics for hospital resource planning-two case studies. *Production Oper. Management* 30(6):1863–1885.

Downloaded from informs.org by [162.194.116.104] on 23 March 2025, at 12:48 . For personal use only, all rights reserved.