Management Science

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

Mixing It Up: Operational Impact of Hospitalist Caseload and Case-Mix

Masoud Kamalahmadi, Kurt M. Bretthauer, Jonathan E. Helm, Alex F. Mills, Edwin C. Coe, Alisa Judy-Malcolm, Areeba Kara, Julian Pan

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

MANAGEMENT SCIENCE
Vol. 69, No. 1, January 2023, pp. 283–307
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Mixing It Up: Operational Impact of Hospitalist Caseload and Case-Mix

Masoud Kamalahmadi,[a,*] Kurt M. Bretthauer,[b] Jonathan E. Helm,[b] Alex F. Mills,[c] Edwin C. Coe,[d] Alisa Judy-Malcolm,[e] Areeba Kara,[e] Julian Pan[f]

[a] Miami Herbert Business School, University of Miami, Coral Gables, Florida 33146; [b] Kelley School of Business, Indiana University Bloomington, Bloomington, Indiana 47405; [c] Zicklin School of Business, Baruch College, City University of New York, New York, New York 10010; [d] St. Anthony Hospital, Gig Harbor, Washington 98332; [e] Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana 46202; [f] Lean Care Solutions Corporation Pte. Ltd, Singapore 139959, Singapore
*Corresponding author

Contact: mkamalahmadi@miami.edu, https://orcid.org/0000-0002-4132-9780 (MK); kbrettha@indiana.edu, https://orcid.org/0000-0002-7785-750X (KMB); helmj@indiana.edu, https://orcid.org/0000-0001-5577-5530 (JEH); alex.mills@baruch.cuny.edu, https://orcid.org/0000-0003-4087-0441 (AFM); edwincoe@chifranciscan.org (ECC); ajudymalcolm@iuhealth.org (AJ-M); akara@iuhealth.org, https://orcid.org/0000-0001-5107-3857 (AK); jp@leancaresolutions.com (JP)

Received: September 24, 2019
Revised: February 28, 2021; November 2, 2021
Accepted: January 4, 2022
Published Online in Articles in Advance: March 15, 2022

https://doi.org/10.1287/mnsc.2022.4342

Copyright: © 2022 INFORMS

Abstract. Hospitalists are medical doctors that specialize in the care of hospitalized patients, a role that until recently belonged to primary care physicians. We develop an operational model of hospitalist-patient interactions with rounding and responding service modes, optimizing hospitalist caseload and case-mix to achieve the maximal reduction in patient length of stay (LOS). We show that hospitalists are effective at reducing LOS for patients with complex conditions, corroborating intuitive reasoning. However, the optimal hospitalist case-mix also includes "simple" patients with few interventions and short LOS, as they can effectively reduce discharge delays. This actionable insight is particularly salient for small community hospitals with simple, short-stay patients, where hospitalists may be undervalued due to the prevailing belief that they are primarily effective for complex patients. We conduct a comparative case study of a small community hospital and a large academic hospital, drawing a stark contrast between the two in terms of ideal caseload and patient coverage. Despite the fact that the academic hospital treats higher complexity patients, hospitalists at the community hospital should actually have a lower caseload than hospitalists at the academic hospital due to shorter stays in the community hospital. We find that both hospitals are understaffed but for different reasons: the academic hospital needs to staff more hospitalists to reduce the current caseload of its hospitalists, whereas the community hospital needs to staff more hospitalists to expand its hospitalist coverage to more patients. We estimate that these hospitals can save on average $1.5 million annually by implementing the optimal staffing policies.

History: Accepted by Stefan Scholtes, healthcare management.
Funding: This work was supported by a PSC-CUNY Award, jointly funded by The Professional Staff Congress and The City University of New York.
Supplemental Material: The e-companion and data files are available at https://doi.org/10.1287/mnsc.2022.4342.

Keywords: healthcare operations management • hospitalist caseload • case-mix • length of stay

1. Introduction

Until the late 1990s, primary care physicians (PCPs) managed all aspects of care for patients in the U.S. healthcare system, ranging from preventive care to the care of critically ill hospitalized patients (Wachter and Goldman 1996). In the last few decades, medical care has shifted to ambulatory settings, and many illnesses that were once occasions for hospital admission are now treated in an outpatient setting, while only the most critical patients are hospitalized (Sox 1999). As a result, PCPs now spend a larger fraction of their time on outpatients, while inpatients require more intensive care. These two trends have motivated a substantial number of physician groups and hospitals to introduce a new specialty called "hospital medicine" that is dedicated to inpatient care (Meltzer 2001). Those who specialize in hospital medicine are called *hospitalists*.

Wachter and Goldman (1996) first defined a hospitalist as a physician who dedicates at least 25% of his or her practice to inpatient care. Today, most hospitalists work full-time with hospitalized patients. They practice in many specialties, most commonly internal medicine and family medicine (Glasheen et al. 2011). According to Wachter and Goldman (2016), the number of hospitalists is more than 50,000 and hospital medicine is the fastest-growing specialty in U.S. healthcare history. As of 2016, about 75% of U.S. hospitals have at least one hospitalist (Wachter and Goldman 2016).

In the hospitalist model of care, when a patient is admitted to a hospital, the PCP transfers responsibility for the patient's care to a hospitalist. As the attending physician, hospitalists admit patients, coordinate specialists' consultations, oversee diagnostic needs, schedule procedures, and generally accelerate patients' progress toward discharge (Molinari and Short 2001). Upon discharge, responsibility for the patient's care returns to the PCP.

A hospitalist provides service in two ways (Tipping et al. 2010). Direct patient care services, which account for about 20% of each shift, involve interactions with patients, for example, visiting patients for diagnosis, examination, and treatment, and providing discharge instructions. Indirect patient care services, which account for about 70% of each shift, include working with electronic medical records (EMRs), reviewing test results, documenting the treatment and discharge notes, and communicating with other care providers. The rest of each shift is spent on education, professional development, traveling within the hospital, and other activities.

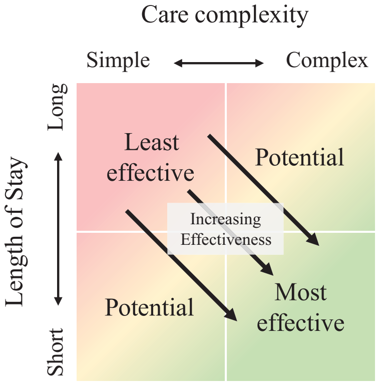
Studies in the clinical literature have demonstrated the pros and cons of the hospitalist model. The hospitalist model can reduce costs and length of stay (LOS) without adversely affecting quality outcomes (see Section 3 for a full review of the relevant literature). This, however, comes at an increased number of patient hand-offs between PCPs and hospitalists, which could potentially harm patients' experience and outcomes. Electronic medical records facilitate communication and coordination between hospitalists and PCPs, which has alleviated this issue (Kripalani et al. 2007). Overall, the benefits of the hospitalist model seem to outweigh its costs, as is evident in the rapid growth of hospitalists in U.S. hospitals. As such, we focus on the benefits of the hospitalist model, which have been studied primarily from a clinical perspective. To the best of our knowledge, this recent major change in hospital care delivery has not been studied in the operations management literature. Fundamental questions remain regarding how to implement the hospitalist model most effectively, especially surrounding hospitalists' caseload. In the 2018 Hospitalist Career and Compensation Survey, 45% of hospitalists reported a caseload of more than 17 patients (n = 522 hospitalists), which they believed to be excessive for operational efficiency (Today's Hospitalist 2018). In a recent commentary, Wachter (2014, p. 794), who is known as the "father of hospitalist medicine," asks why there have been virtually no studies of hospitalist caseload.

To address these concerns, we aim to answer the research question: *What is the ideal hospitalist caseload?* We measure caseload as the volume of patients attended by a hospitalist. Michtalik et al. (2013b) suggest that the characteristics of different patient groups such as complexity of care need to be included in caseload analyses as well. Medical studies have shown that patients with different conditions may benefit differently from a hospitalist's care (Kuo and Goodwin 2010). It is important to understand which patients benefit most from a hospitalist's care. Therefore, our second research question is: *What is the ideal hospitalist case-mix?* We define *case-mix* as the set of patient types that are attended by a hospitalist.[1]

We seek a caseload and case-mix that will be most effective in reducing average LOS by taking a systemwide view of their impact on hospital operations. We capture the impact of hospitalists on a patient's LOS via a model of non-value-added delay time that occurs when patients need a physician's intervention to progress to their next stage of care. For example, a patient may need a change of medication, an analysis of a diagnostic image, or the development of discharge instructions. Because a hospitalist is on-site, he or she can reduce these delays by responding to patients' needs more quickly than a nonhospitalist. We integrate this patient-level model into a system-model to characterize the caseload and case-mix of a hospitalist that achieve the greatest reduction in hospital census or, equivalently in average LOS.

Figure 1 summarizes the key conceptual findings of our analysis. Using the number of hospitalist interventions as a proxy for clinical complexity, we find that hospitalists are more operationally effective for patients with shorter lengths of stay and/or more complex conditions. In our numerical study of hospital inpatients with 45 common diagnosis-related groups (DRGs), we find that hospitalists are most effective on the most complex DRGs. However, the optimal case-mix also includes DRGs with short LOS and low complexity. We also show that the optimal caseload (between 9 and

Figure 1. (Color online) Hospitalist Effectiveness in Terms of LOS Reduction as a Function of Complexity and Length of Stay



11 patients in our case studies) is sensitive to the complexity of patients' conditions through the *frequency* by which patients need hospitalist interventions. For example, complex patients need more hospitalist interventions than simple patients, but they also have longer LOS. For operational efficiency, a hospitalist who cares for complex and long-stay patients may need to have a larger caseload than a hospitalist who cares for simple and short-stay patients. The optimal caseload is also a function of the hospitalist's visit rate (i.e., the number of patient visits per hour).

We further solve the model using data from two partner hospitals in the United States that differ in size, mission (academic or community), and patient characteristics (complexity and LOS). We characterize the optimal staffing for each hospital and estimate the operational and financial improvement of adopting the optimal staffing. We find that the optimal hospitalist case-mix includes patients with long-complex and short-simple conditions at both hospitals; however, despite the fact that the academic hospital treats higher complexity patients, hospitalists at the community hospital should actually have a lower caseload than hospitalists at the academic hospital due to shorter stays in the community hospital. Although this matches the current practice, we find that hospitalists are underloaded at the community hospital and substantially overloaded at the academic hospital. We find that both hospitals should increase their staffing levels. This would benefit the community hospital through speeding up the patient discharges, and benefit the academic hospital through lowering the caseload of their current hospitalists, allowing them to be more effective in managing each treatment and reducing delays. We estimate that these hospitals can save on average $1.5 million annually by implementing the optimal staffing policies recommended by our model.

1.1. For Healthcare Industry Readers

Practice-oriented methods, results, and insights for the hospitalist model of care can be found in Sections 2–4 and 7–8. In Section 2, we provide a high-level discussion of our methods of analysis. Specifically, we present our conceptual model of hospitalist work routines and their modes of interaction with patients and how it relates to practice. We further provide narrative and quantitative descriptions of the specific hospitals and hospital settings studied in our data-driven analysis: Level 1 (large academic hospital) and Level 3 (small community hospital). Section 3 relates our hypotheses and results to the clinical literature on the impact of hospitalist care on hospital cost and quality (see Table 1), as well as the impact of caseload and case-mix on hospitalists. Section 4 presents the key findings of the paper as hypotheses (summarized here) that are supported by the analyses in Sections 5–8.

Results: Our findings suggest that:

Case-mix: Hospitalists can most effectively reduce length of stay by attending a mix of complex and simple patients.

Caseload: To most effectively reduce length of stay, hospitalists' caseloads should be adjusted to account for case-mix, where a lower caseload is recommended if case-mix includes more complex patients and/or more patients with short lengths of stay.

Sections 5–6 present mathematical analysis to support our hypotheses, which can be skipped if the primary interest lies in implication for management practice. Toward this end, Section 7 uses the State Inpatient Database (2015) to study the question of optimal hospitalist case-mix selection in generality by analyzing a large and heterogeneous set of DRGs. Section 8 demonstrates the impact of applying our insights on caseload through a detailed counterfactual analysis of two specific hospitals, highlighting differences in how hospitalists should be deployed in a large academic hospital versus a small community hospital. A key finding is that hospitalists may be more effective in the large hospital by reducing caseload by reducing coverage (not all patients attended by a hospitalist). In contrast, the community hospital may extract significant cost savings by increasing hospitalist coverage through hiring more hospitalists and increasing each hospitalist's caseload.

2. Case Study Context

As part of this research, we conducted multiple site visits at and obtained data from a large academic hospital and a small community hospital in the U.S. Midwest. "Academic Hospital" has about 600 staffed beds and serves over 30,000 patients annually. Inpatients are attended by either a hospitalist or a family physician affiliated with their primary care practice. Each hospitalist is localized to a medical or a surgical floor; that is, he or she primarily sees medical or surgical patients. In spite of this difference, the workflow of the hospitalists is similar in both floors. We focus on hospitalists who primarily see patients with medical conditions. During the day shift from 7 a.m. to 7 p.m., nine hospitalists are on duty in the medical floors. Each hospitalist manages the care of 15 patients on average. Our practicing hospitalist coauthors helped develop the workflow map of a hospitalist shown in Figure 2. Hospitalists start their day by rounding on their patients, meaning that they visit their patients one after another. Each patient visit takes on average 30 minutes, including on average 10 minutes that are spent inside the patient's room for medical assessment, and on average 20 minutes that are spent outside of

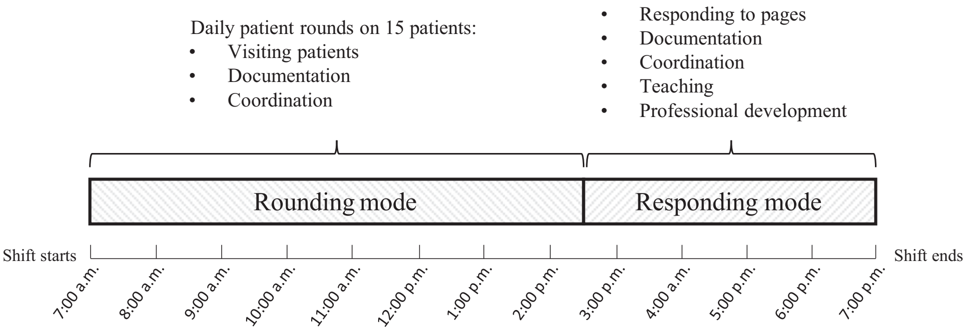
Table 1. Literature Review of Medical Studies

		Outcomes for hospitalists' patients			
Study	Clinical setting	Hospitalization cost	LOS	Mortality	Readmission
Palmer et al. (2001)	Patients cared for by hospitalists, generalists, and specialists at West Virginia University Hospitals	↓	↓	—	—
Lindenauer et al. (2002)	Heart failure cases cared for by hospitalists and general internists at a hospital in Massachusetts	—	↓	N/A	N/A
Phy et al. (2005)	Hip fracture patients cared for by hospitalists and standard groups	N/A	↓	—	—
Roy et al. (2006)	Hip fracture admissions cared for by hospitalists or nonhospitalists at a community medical center	↓	↓	N/A	N/A
Lindenauer et al. (2007)	76,926 patients cared for by hospitalists, general internists, and family physicians, at 45 hospitals	↓	↓	—	—
Southern et al. (2007)	Patients cared for by hospitalist and nonhospitalist teams at Weiler Hospital	N/A	↓	—	—
Roytman et al. (2008)	Heart failure patients cared for by hospitalists, cardiologists, general internists, and family practitioners at a community hospital	↓	↓	—	—
Kuo and Goodwin (2010)	58,125 admissions by hospitalists and nonhospitalists at 454 hospitals	↓	↓	—	N/A
Howrey et al. (2011)	10,884 stroke patients attended by hospitalists or nonhospitalists	N/A	↓	—	↑
Yousefi and Chong (2013)	34,524 admissions to LH Oshawa, Lakeridge Health, Canada, attended by family physicians, internal medicine subspecialists, or hospitalists	N/A	↓	↓	↓
Stevens et al. (2017)	560,561 admissions cared for by hospitalists, PCPs, or generalists	N/A	↓	—	—

Note. ↓ or ↑ indicates a statistically significant difference for hospitalists' patients in the specified direction; — indicates no difference, and N/A indicates outcome was not studied.

the patient's room documenting the patient's progress in the EMR system, and coordinating his or her care with nurses and other providers. Hospitalists at this hospital typically finish visiting all their patients by about 2:30 p.m., after which they respond to nurses' and patients' requests as needed while spending most of their time on computer interactions, communication, teaching, and professional development. Hospitalists are also responsible for admissions and discharges. During the night shift from 7 p.m. to 7 a.m., two hospitalists provide emergency coverage and admit new patients who arrive overnight (these admissions are distributed among the day hospitalists the next morning). Except in emergency situations, the night hospitalists do not examine existing patients, most of whom are asleep. Because of this difference, we focus

Figure 2. Day-Shift Hospitalists' Workflow in Academic Hospital



only on the day shift. In the academic hospital, management's goal is for most, if not all, general medical patients to be attended by a hospitalist.

"Community Hospital" has fewer than 100 staffed beds and serves around 7,000 patients annually. Inpatients are attended by either a hospitalist, a specialist (e.g., surgeon, cardiologist), or their own primary care physician. During the day shift, three hospitalists are on duty. Each hospitalist manages the care of seven patients on average. Their main care responsibilities are similar to those of the academic hospital (except that they do not have teaching responsibilities), but hospitalists only attend a small fraction of the patients. Moreover, hospitalists may attend a mix of medical and surgical patients at the same time due to the lower patient volume.

Although neither hospital has specific rules for assigning patients to hospitalists or nonhospitalists, hospitalists tend to attend more complex patients in both hospitals. Some of the most common conditions seen by hospitalists at these hospitals are septicemia, intracranial hemorrhage, heart failure, renal failure, gastrointestinal hemorrhage, diabetes, cellulitis, and simple pneumonia.

The difference in the total volume of patients attended by hospitalists and the caseload of each hospitalist between these two hospitals highlights the importance of our research questions regarding the ideal hospitalist caseload and case-mix, and adds a third research question: *What portion of the overall patient population should be attended by hospitalists?* We use the information that we obtained during our site visits and interviews to inform our model of hospitalist work modes in Sections 5–6. In Section 8, we use the data from these hospitals to parameterize our model and evaluate the third research question.

3. Literature Review

In this section, we review the clinical literature on hospitalist outcomes and caseload, and then we review the relevant operations management (OM) literature on staffing in healthcare systems.

3.1. Hospitalist Outcomes in the Clinical Literature

Many empirical studies have compared the outcomes of patients who are attended by hospitalists with patients who are attended by a nonhospitalist (e.g., a PCP). Table 1 summarizes the findings of these studies, showing that the hospitalist model is broadly associated with lower cost and LOS, and insignificant differences in mortality and readmissions. In a systematic review, White and Glazier (2011) report that hospitalists' patients had shorter LOS and lower costs with comparable quality outcomes in 70% of the studies (n = 65 studies). Using meta-analysis, Rachoin et al. (2012) estimate that hospitalists' patients have a mean LOS that is 0.44 days shorter than nonhospitalists' patients. The extent of LOS reduction depends on patients' conditions. Southern et al. (2007) find that reduction in LOS was greatest for patients requiring close clinical monitoring and for those requiring complex discharge planning. Kuo and Goodwin (2010) find that reduction in LOS was greatest in older, complicated, nonsurgical patients.

These findings motivate us to ask why a hospitalist may reduce LOS for some types of patients and not for others. The answer is not straightforward, and because the impact of the hospitalist model is largely in operational outcomes, it is important to study it from an OM perspective.

3.2. Hospitalist Caseload

In a recent survey, 40% of hospitalists report unsafe caseload at least monthly (n = 506 hospitalists; Michtalik et al. 2013a). Hospitalists report that excess caseload prevents them from fully discussing treatment options, leads to unnecessary laboratory tests, delays admissions or discharges, and worsens patient satisfaction, though there are few studies in the literature quantifying the effects of caseload on hospitalists. Lurie and Wachter (1999) propose a very simple model to determine the number of hospitalists needed to provide care at a hospital by fixing the hospitalist caseload at 10 patients per hospitalist and then using the average annual census to calculate the number of needed hospitalists. Elliott et al. (2014) study the effects of caseload on the quality and efficiency of care provided by hospitalists. The authors conclude that increases in hospitalist caseload are associated with clinically meaningful increases in LOS and cost, and they report that LOS increases exponentially as caseload increases above approximately 15 patients. Responding to Elliott et al. (2014), Wachter (2014) questions why there have been virtually no other studies about hospitalist caseload. Michtalik et al. (2013b) propose that patient complexity should also be considered in determining hospitalist caseload. Our paper contributes to this body of literature on hospitalist caseload and efficacy by providing an analytical model of their workflow process and using it to study the impacts of hospitalist caseload and case-mix on operational outcomes.

3.3. Healthcare Staffing in Operations Management

Staffing decisions in medical facilities have been studied extensively in operations management. In the outpatient setting, research focuses on panel size and design, and appointment scheduling (Green and Savin

2008, Bavafa et al. 2013, Ozen and Balasubramanian 2013, Liu and Ziya 2014, Zacharias and Armony 2017). The key trade-off in characterizing the optimal panel size is as follows: if the panel is too large, then patients have to wait for a long time to get an appointment. If the panel is too small, then waiting times would be short, but the overall coverage would be low. This trade-off is common in many other settings such as call-centers (see Gans et al. 2003). We apply a similar trade-off in the inpatient setting, where service times are much longer. In contrast to outpatient and call-center models, where the server serves one customer at a time, we study the case where multiple patients receive care from the same provider simultaneously. In the inpatient setting, many have studied staffing of nurses and other providers (Wright et al. 2006, de Véricourt and Jennings 2011, He et al. 2012, Green et al. 2013, Wang and Gupta 2014). To the best of our knowledge, no research in operations management has analyzed staffing or panel design for hospitalists, an area that has some unique operational characteristics relative to other service and healthcare providers.

A key feature of hospitalists' jobs is repeated interactions with patients. Recent studies have modeled this kind of repeated interactions in service systems. De Véricourt and Jennings (2011) formulate a closed queueing model to determine efficient staffing policies in a nursing home where patients alternate between needing assistance and not. Yankovic and Green (2011) develop a finite-source queueing model to optimize nursing levels where the demand for nursing care comes from patient arrival and departure as well as patients' requests during their stays. Yom-Tov and Mandelbaum (2014) use an Erlang-R model to optimize staffing when patients alternate between a service and a delay phase. Dobson et al. (2013) study prioritization of customers that visit multiple stations during service. Chan et al. (2014) model provider-controlled service rate where changing the service rate impacts the likelihood of a customer requiring rework. Campello et al. (2016) study workload of case managers who are assigned to multiple homogeneous customers and interact with them repeatedly. Of these studies, the three that are most conceptually similar to our work are de Véricourt and Jennings (2011), Yankovic and Green (2011), and Campello et al. (2016). Our work differs in two crucial ways from these studies. First, hospitalists typically use two different modes of interactions with patients: rounds and responding. These studies do not model rounds, but they do bear some similarities to our responding model. A main difference in de Véricourt and Jennings (2011) and Yankovic and Green (2011) is that any available nurse can intervene when patients require assistance. In contrast, we model a situation in which hospitalists attend specific patients. In this way, we are more similar to Campello et al. (2016); however, they specify a predefined maximum load and assignment rule. In contrast, we endogenize the caseload as a decision. We further integrate rounding and responding in a rounding-responding model that captures the true hospitalist practice and differs from other models in the literature. Another key feature of our study is heterogeneity of patient care complexity, which is not considered in any of these studies.

To the best of our knowledge, the research with the closest similarity to our model of rounds is by Chan et al. (2016) and Dong and Perry (2020). They propose queuing models for patient flow dynamics where patients must be inspected before discharge. Chan et al. (2016) analyze the effects of number and timing of inspections on the queue length. Dong and Perry (2020) assume that inspections take place at a specific time during the day and discharge delays are constant. They quantify the effect of the discharge policy on bed utilization and throughput. Shi et al. (2015) also model inpatient LOS while assuming that discharges happen at predetermined times during the day, and they investigate the effect of different discharge policies on patients' waiting time. Our study has two main differences from these studies: first, we study the case where patients require repeated interventions and delays are not limited to discharge; and second, we endogenize the timing of interventions as a function of provider caseload. We also assume that delays are a function of the inspections, which are a function of caseload.

4. Hypothesis Development

Based on the clinical and operations management (OM) literature, we develop hypotheses regarding the ideal hospitalist case-mix and caseload that will be most effective in reducing average length of stay (LOS).

4.1. Case-Mix

Medical studies have reported that hospitalist care is associated with a reduction in patients' LOS, and the largest reduction is achieved for patients with complex conditions (see Section 3.1). For example, Southern et al. (2007) find that sepsis patients attended by a hospitalist had on average 3.7 days shorter LOS than similar patients attended by nonhospitalists. Sepsis is a complex condition that requires long hospitalization, close monitoring, and careful discharge planning. Other conditions that experienced large LOS reductions were congestive heart failure, stroke, asthma, and pneumonia (Southern et al. 2007). It is conjectured that the more complex patients require closer monitoring and regular intervention as the doctor works to manage evolving health condition(s).

This may require ordering multiple tests, coordinating procedures, and developing and refining treatment plans. Intuitively, on-site hospitalists, due to physical proximity and the singular focus on hospitalized patients, can make these adjustments more rapidly, reducing delays to recovery and treatment progress due to faster response times.

On the other hand, patients with few complications tend to have short LOS and are ready for discharge quickly. Therefore, discharge delay has a proportionally larger impact on LOS. Since these patients have a higher throughput rate by turning over more quickly, small reductions over a larger volume of patients may have a greater aggregate benefit than larger reductions over fewer (longer staying) patients. A hospitalist attending these patients would ensure timely access to care for more patients and/or reduce congestion in inpatient units (Powell et al. 2012), creating a less stressful work environment for the healthcare personnel (Kuntz et al. 2015) and a safer environment for patients (Trzeciak and Rivers 2003, Sprivulis et al. 2006). Hospitalists are in the hospital and can substantially reduce discharge delays (Palmer et al. 2001), as well as facilitate the aforementioned benefits.

In summary, hospitalists may reduce LOS by more rapidly moving the care of a patient forward, with more opportunities presented in more complex patients, or by minimizing the time between discharge readiness and discharge. Hence, we propose the following two hypotheses for the hospitalists' ideal case-mix:

Hypothesis 1a. *Holding caseload constant, hospitalists are more effective in reducing average LOS when attending patients with more complex conditions.*

Hypothesis 1b. *Holding caseload constant, hospitalists are more effective in reducing average LOS when attending patients with shorter LOS.*

Note that complexity of care and LOS are highly correlated. Therefore, Hypotheses 1a and 1b can be considered to be competing hypotheses.

4.2. Caseload

When a worker serves multiple customers simultaneously, his or her *caseload*, or the average number of customers in simultaneous service, affects the outcomes of the service for each customer (Luo and Zhang 2013, Tan and Netessine 2014). A hospitalist cares for multiple patients simultaneously. Each additional patient consumes a portion of the hospitalist's limited time and capacity, making him or her less responsive to the needs of the other patients, which increases delays. Elliott et al. (2014) find that LOS increases exponentially as hospitalist caseload increases. A smaller caseload allows the hospitalist to be more responsive, but it limits the system-level impact of hospitalists because fewer patients can be attended by the same number of hospitalists. Therefore, the optimal caseload should balance patient waiting time and coverage.

Empirical evidence from our observations indicates the hospitalists' caseload can vary substantially across hospitals. We saw this in our case study hospitals. At Academic Hospital, where most patients have complex conditions with long LOS, each hospitalist cares for 15 patients, whereas at Community Hospital, where most patients have simple conditions and short LOS, each hospitalist cares for seven patients. This suggests that there may be an interaction between hospitalist case-mix and caseload. On the one hand, patients with more complex conditions tend to need more hospitalist interventions. If their hospitalist has a large caseload, then they will experience prolonged delays at multiple occasions during hospitalization. On the other hand, patients with simpler conditions often recover more quickly and turn over faster. If their hospitalist has a large caseload, then they will experience unnecessarily long discharge delays. Since complexity and LOS are correlated, we propose the following two complementary hypotheses.

Hypothesis 2. *Holding patients' average LOS constant, as complexity of conditions in the case-mix increases, hospitalist ideal caseload decreases.*

Hypothesis 3. *Holding patients' average complexity constant, as LOS of conditions in the case-mix decreases, hospitalist ideal caseload decreases.*

5. The Model

In this section, we formulate a model to determine the steady-state caseload and case-mix per hospitalist that minimizes average hospital census. When arrivals are exogenous, minimizing the average census is equivalent to minimizing the average length of stay (LOS). Unnecessarily long LOS is costly for the hospital and undesirable for the patient. First, we propose a general model without imposing any structure on patients' LOS and patient-provider interactions. We then develop a multiphase model of patients' LOS, as well as a model of patient-provider interactions with rounding and responding service modes to endogenize the effect of hospitalist caseload on patients' LOS.

5.1. Caseload and Case-Mix Optimization Model

We consider a hospital scheduling n hospitalists per shift. There are I types of patients. Type $i \in \{1, 2, \ldots, I\}$ patients arrive at an average rate of λ^i per unit time. A patient's LOS has two components: service requirement and delay. Service requirement is the value-added time that the patient must stay in the hospital for clinical care and recovery before discharge. The service requirement of each type i patient is independent and identically distributed according to a random variable Ψ^i with expected value $S^i = E[\Psi^i]$. At certain points during the treatment and recovery process, the patient requires an intervention by the attending physician. Since the attending physician has more than one patient, the patient may experience a random non-value-added delay waiting for the physician. The delay time depends on whether the patient is attended by a hospitalist or not. If a patient is attended by a hospitalist, then his or her delay time depends on the hospitalist's caseload. Otherwise, the delay time depends on the time between two consecutive visits by the nonhospitalist. The nonhospitalist can be the patient's PCP or a specialist (e.g., cardiologist, neurologist). We assume that there are sufficiently many nonhospitalists or their visits are sufficiently infrequent that the nonhospitalist visit interval is independent of the total nonhospitalist workload. This is reasonable since many nonhospitalists such as PCPs or specialists spend most of their time outside the hospital.

Let r be the hospitalist caseload, defined as the average number of patients per hospitalist. The decision variables are r^i, $i \in \{1, 2, \ldots, I\}$, which denote the average number of type i patients per hospitalist and describe the hospitalist case-mix. The caseload of each hospitalist (the patient-to-hospitalist "ratio") is then given by $r = \sum_{i=1}^{I} r^i$. We assume identical caseload and case-mix for hospitalists, but relax this assumption in Appendix B.1 in the e-companion. We use an average-case model, so the decision variables represent the steady-state caseload and case-mix per hospitalist. Hospitalists' caseload and case-mix do change throughout the day, so our results should be interpreted as the "ideal" caseload and case-mix. Let $D_H^i(r)$ be the expected delay for a type i patient attended by a hospitalist having caseload r, which is an increasing function of r. Let $S_H^i(r)$ be the expected LOS of such a patient. Finally, let D_N^i and S_N^i be the expected delay and the expected LOS of a type i patient attended by a nonhospitalist. Then we have $S_H^i(r) = S^i + D_H^i(r)$ and $S_N^i = S^i + D_N^i$.

Arrivals of type i patients occur with rate of λ^i and are divided between hospitalists and nonhospitalists with rates of λ_H^i and λ_N^i, respectively, such that $\lambda_H^i + \lambda_N^i = \lambda^i$. Using Little's law, the average census of type i patients in the hospital is $\lambda_H^i S_H^i(r) + \lambda_N^i S_N^i$, and consequently the average hospital census is $\sum_{i=1}^{I}(\lambda_H^i S_H^i(r) + \lambda_N^i S_N^i)$. We formulate the following optimization model to determine hospitalist caseload and case-mix that minimizes the average hospital census:

$$\min_{\lambda_H^i, \lambda_N^i, r^i, i \in \{1,2,\ldots,I\}} \sum_{i=1}^{I} \left(\lambda_H^i S_H^i(r) + \lambda_N^i S_N^i\right) \tag{1}$$

$$\text{subject to} \quad \lambda_H^i + \lambda_N^i = \lambda^i, \quad \forall i \in \{1,2,\ldots,I\} \tag{2}$$

$$\begin{aligned} nr^i &= \lambda_H^i S_H^i(r), \quad \forall i \in \{1,2,\ldots,I\} \\ r^i, \lambda_H^i, \lambda_N^i &\geq 0, \quad \forall i \in \{1,2,\ldots,I\}. \end{aligned} \tag{3}$$

Objective function (1) minimizes the average hospital census, which is equivalent to minimizing the average LOS. Constraints (2) divide arrivals between hospitalists and nonhospitalists. Constraints (3) enforce Little's law: the average number of type i patients assigned to n hospitalists equals their average arrival rate times their average LOS. If all of the patients of a given type i are attended by hospitalists, then $\lambda_H^i = \lambda^i$ in (3), and each hospitalist gets $\lambda^i S_H^i(r)/n$ patients from that type. Let $\omega_i(r) = \lambda^i S_H^i(r)/n$. A feasible solution to our defined problem satisfies $r^i \leq \omega_i(r)$, $\forall i \in \{1, 2, \ldots, I\}$. We call $r^i \leq \omega_i(r)$ the *demand constraint* for type i.

See Table EC.1 in Appendix A of the e-companion for a summary of notations used in this paper.

Next, we propose a model of patients' LOS to characterize service requirements. We then develop a model of patient-provider interactions to characterize delays.

5.2. Modeling Length of Stay

Following Faddy et al. (2009), we model the service requirement of a patient as a continuous-time Markov chain (CTMC) with states (known as *phases*) such that after spending time in each phase, a patient either progresses to the next phase or moves to an absorbing state that corresponds to discharge. Given this structure, the service requirement follows a *Coxian phase-type distribution* (Marshall and McClean 2004, Payne et al. 2011). Figure 3 depicts the general form of the CTMC, where the time spent in phase j is exponentially distributed with parameter μ_j and each patient progresses from phase j to phase $j+1$ with probability p_j or to the absorbing state with probability $q_j = 1 - p_j$. In Appendix C.1 in the e-companion, we use real patient data from a partner hospital to validate this model. The expected total service requirement for a patient with J phases is $S = E[\Psi] = 1/\mu_1 + p_1(1/\mu_2 + p_2(1/\mu_3 + \ldots p_{J-1}(1/\mu_J))) = \sum_{j=1}^{J} P_j/\mu_j$, where $P_1 = 1$ and $P_j = p_1 p_2 \ldots p_{j-1}$ is the probability of visiting phase j for $j > 1$. Observe that $\sum_{j=1}^{J} P_j$ is the expected number of phases. In this study, a higher number of phases indicates a more complex plan of care. In later sections, we will add the superscript i to the parameters to describe the LOS of type i patients. For patient type i, J^i describes the number of phases, μ_j^i describes the service rate in phase j, p_j^i describes the transition probability from phase j to $j+1$, and P_j^i describes the visit probability of phase j.

We model non-value-added delay by assuming that the patient requires intervention from the attending physician at the end of each phase (prior to the transition). This assumption is based on interviews with several hospitalists who stated that, as the coordinator of patient care, their interventions are required to drive patient treatment and recovery forward. Examples of interventions are when the patient needs a medication change to move to the next state of inpatient recovery, or when the physician needs to process discharge paperwork for the patient to move to the absorbing state. Since the length of each phase is stochastic, and the physician attends multiple patients (hospitalists) or is not always on-site (nonhospitalists), patients experience delays at the end of each phase. These delays depend on the way the physician interacts with patients, and the number of patients he or she is treating (hospitalists) or the frequency of his or her hospital visits (nonhospitalists).

5.3. Modeling Patient-Provider Interactions

We interviewed and shadowed several hospitalists at work to develop a better contextual grounding about their workflow. Hospitalists typically start their day by visiting all of their patients one after another in a round-robin fashion, which we call the *rounding* mode or "rounds" (Figure 2). During rounds, hospitalists assess and document patients' progress, and coordinate necessary interventions with other providers. Hospitalists with a heavy caseload may spend nearly their entire shift on rounds. When rounds are finished, hospitalists switch to a different work pattern that we call *responding* (Figure 2). In this mode, hospitalists spend their time on indirect care services (such as working on EMRs and communicating with other providers) and provide direct care to patients as needs arise. We propose a model of patient-provider interactions with rounding and responding, where the hospitalist caseload determines not only the delays for direct patient care within each mode, but also the amount of time spent on rounding versus responding. In contrast to hospitalists, nonhospitalists are not always on-site at the hospital, and most commonly function in rounding mode only. Therefore, we model the interactions between nonhospitalists and patients only in rounding mode.

First, consider the case for hospitalists. Define $\pi(r) \in [0,1]$ as the fraction of time a hospitalist spends rounding when his or her caseload is r. The hospitalist then spends $1-\pi(r)$ responding. The expected total LOS of a patient can then be expressed as the weighted combination of the expected LOS based on the fraction of time spent in each mode (since each phase is exponentially distributed, we can apply PASTA and observe that the probability a phase ends during the rounding mode is $\pi(r)$). Let $S_H^{\text{RND}}(r)$ and $S_H^{\text{RSP}}(r)$ denote the expected LOS of a patient attended by a hospitalist with caseload r when $\pi(r)=1$ and $\pi(r)=0$, respectively. Then,

$$S_H(r) = \pi(r)S_H^{\text{RND}}(r) + (1-\pi(r))S_H^{\text{RSP}}(r). \tag{4}$$

The expected LOS for a patient attended by a nonhospitalist S_N is a special case of $S_H^{\text{RND}}(r)$, where delays depend on the visit frequency of nonhospitalists, not their caseload. We model the expected LOS in the pure rounding mode $S_H^{\text{RND}}(r)$ and S_N in Section 5.3.1. We then model the expected LOS in the pure responding mode $S_H^{\text{RSP}}(r)$ in Section 5.3.2. We characterize $\pi(r)$ and $S_H(r)$ in Section 5.3.3.

5.3.1. Rounding Service Mode.

In this section, we model the interactions of patients and providers in the pure rounding mode to characterize $S_H^{\text{RND}}(r)$ and S_N. Rounding is the practice of visiting patients in a sequential manner. First, consider the case for hospitalists. Let γ be the hospitalist visit rate, that is, the number of patients visited per time unit. The average time between visits to successive patients is therefore γ^{-1}. This interval includes the time spent on direct care during the encounter and the time spent on some of the tasks related to indirect care (e.g., documentation, consultation) that happen immediately after each encounter. For analytical tractability, we assume that γ is deterministic. In a time-motion study, Kara et al. (2019) report CV (coefficient of variation) = 0.3 for the duration of hospitalist-patient encounters,

Figure 3. CTMC Representing a J-phase Coxian Phase-Type Distribution

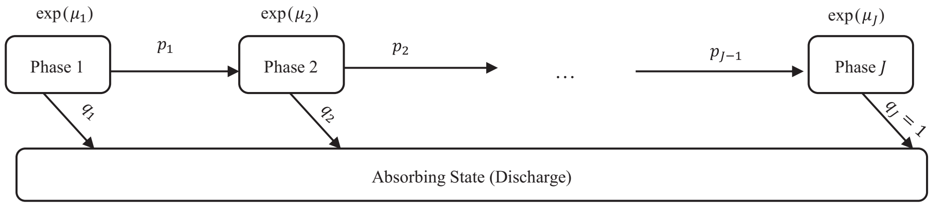
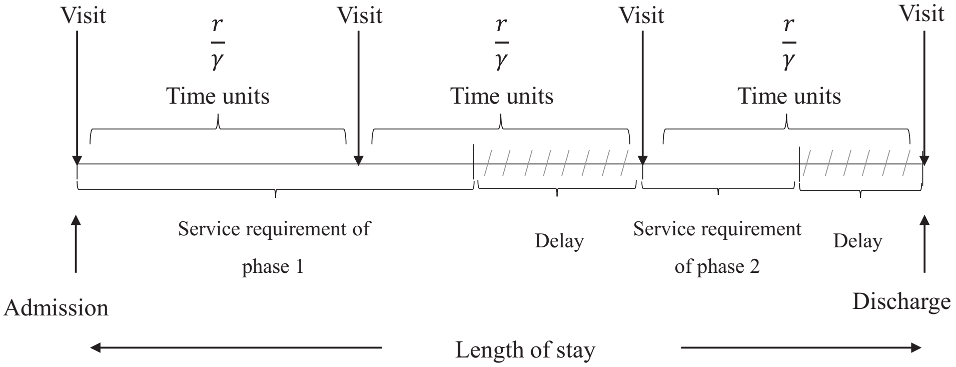


Figure 4. Hospitalist-Patient Interactions via Rounds for Example Patient with Two Phases



suggesting that visit lengths have low variability. Hence, treating them as deterministic is a reasonable approximation to attain tractability. A hospitalist with caseload r who works in pure rounding mode revisits each patient once every r/γ time units, as depicted in Figure 4. A large caseload increases the time between visits of each patient and prolongs delays. Although a smaller caseload results in more frequent visits, it also means that the hospitalist attends fewer patients, limiting his or her system-level impact.

Lemma 1 (Expected LOS: Rounding Mode). *The expected LOS in the rounding mode for a type i patient with J^i phases of care, who is attended by a hospitalist with caseload r is:*

$$S_H^{\mathrm{RND},i}(r) = \sum_{j=1}^{J^i} P_j^i \left(\frac{r/\gamma}{1 - e^{-\mu_j^i r/\gamma}} \right). \tag{5}$$

The proof of Lemma 1 and all subsequent analytical results is given in Appendix D of the e-companion.

We remark that $S_H^{\mathrm{RND},i}(r)$ is convex and increasing in r.

Now, consider the case for nonhospitalists. Let ξ be the nonhospitalist visit rate; that is, a nonhospitalist visits each patient once every $1/\xi$ time units. The expected LOS for a nonhospitalist's patient would be

$$S_N^i = S_H^{\mathrm{RND},i}\left(\frac{\gamma}{\xi}\right) = \sum_{j=1}^{J^i} P_j^i \left(\frac{1/\xi}{1 - e^{-\mu_j^i/\xi}} \right).$$

This also implies that $r \leq \gamma/\xi$.

5.3.2. Responding Service Mode.

In this section, we develop a queueing model for the responding mode of hospitalist-patient interactions to characterize $S_H^{\mathrm{RSP}}(r)$. We model the responding mode for a single hospitalist using a closed queueing network with two nodes and r patients as depicted in Figure 5. A patient with J phases spends a random amount of time distributed exponentially with parameter μ_j for phase j in Node 1 recovering without the hospitalist's presence, and then moves to Node 2, where he or she requires a visit from the hospitalist. Consistent with the assumptions necessary to analyze a closed Jackson network, we assume that the hospitalist visit time is independent and identically distributed according to an exponential distribution with rate γ. After each intervention, the patient returns to Node 1 to begin the next phase of care until he or she is eventually discharged. At discharge, we assume the bed is immediately filled by another patient of the same type; this captures the steady state behavior of case-mix similar to the rounding model. Since each patient is assigned to a single hospitalist, Node 2 functions as a single server station. One modeling challenge in the responding mode is that a patient who enters the queue waits until the hospitalist visits those who are in the queue ahead of him or her. Therefore, the delay experienced by each patient here depends on the hospitalist's entire case-mix, not just total caseload.

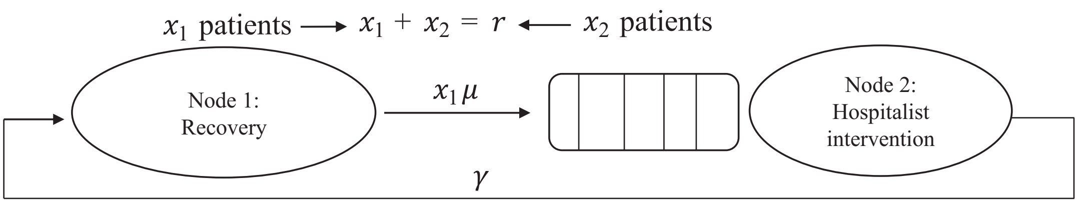
We begin by analyzing patients who have only one exponentially distributed phase and only experience discharge delay. We first analyze the model with only a single patient type. The state of the system is (x_1, x_2), where x_k is the number of patients at Node k, and $x_1 + x_2 = r$ as depicted in Figure 5. Since all patients have identically distributed service requirements, the queueing network is a Jackson network, which has a product-form solution.

Lemma 2 (Expected LOS: Responding Mode, Single Phase, Single Patient Type). *The expected LOS of a patient with a single phase in the responding mode with a single patient type is*

$$S_H^{\mathrm{RSP}}(r) = \frac{1}{\mu} + (r-1)\left(\frac{1}{\gamma} - \left(\frac{1}{\gamma} + \frac{1}{\mu} \right) \frac{\Gamma(r-1, \gamma/\mu)}{\Gamma(r, \gamma/\mu)} \right). \tag{6}$$

Although $S_H^{\mathrm{RSP}}(r)$ in (6) is derived from a model where r is discrete, it happens to be a continuous function of

Figure 5. Closed Queueing Network with Two Nodes and r Patients



r, so we will use (6) to interpolate for average values of r that fall between two integers.

The closed queueing network with multiple patient types or patients with multiple phases does not permit an analytical solution, so we present an approximation.

Approximation 1 (Average Service-Rate Approximation). For the type i patients, replace the service rate of each phase, μ_j^i, by the weighted average service rate $\bar{\mu}^i$, defined as

$$\bar{\mu}^i = \frac{\text{expected number of phases}}{\text{expected total service requirement}} = \frac{\sum_{j=1}^{J^i} P_j^i}{\sum_{j=1}^{J^i} \frac{P_j^i}{\mu_j^i}}.$$

Then approximate the expected LOS as follows:

$$S_H^{\text{RSP},i}(r) = \left(\frac{1}{\bar{\mu}^i} + (r-1)\left(\frac{1}{\gamma} - \left(\frac{1}{\gamma} + \frac{1}{\bar{\mu}^i}\right)\frac{\Gamma(r-1, \gamma/\bar{\mu}^i)}{\Gamma(r, \gamma/\bar{\mu}^i)}\right)\right)\sum_{j=1}^{J^i} P_j^i. \tag{7}$$

Note that $\bar{\mu}^i$ is the average of the service rate of different phases, weighted by the probability of visiting each phase. In (7), we multiply the expected length of a single phase and delay with rate $\bar{\mu}^i$ with the expected number of phases $\sum_{j=1}^{J^i} P_j^i$. Assuming a single patient type, if service rates are equal in every phase, then (7) gives the exact expected LOS. Otherwise, the accuracy of this approximation depends on the variability of service rates around $\bar{\mu}^i$. We use the same approximation for the case with multiple patient types. In other words, for each patient type, we assume that all of the other patients cared for by the same hospitalist are of their type. For example, suppose a hospitalist is attending 10 patients of type 1 and five patients of type 2. In calculating the expected LOS of type 1 patients, we assume that all 15 patients are type 1, and in calculating the expected LOS of type 2 patients, we assume that all 15 patients are type 2. In Appendix C.2 in the e-companion, we use a discrete-event simulator to show that (7) closely approximates the expected LOS in the pure responding mode.

5.3.3. Rounding-Responding Service Mode. Combining the rounding and responding models, we capture the reality of hospitalist-patient interactions, where hospitalists typically spend the first part of their shifts rounding and the remainder responding. Let $\pi(r) = r/\gamma$, which corresponds to the case where hospitalists switch to responding mode after rounding on all their patients once. We can now use (4) to calculate the expected total LOS of a patient. For example, with a single type of patient with one phase, the expected LOS will be

$$S_H(r) = \frac{r}{\gamma}\left(\frac{\frac{r}{\gamma}}{1 - e^{-\mu r/\gamma}}\right) + \left(1 - \frac{r}{\gamma}\right)\left(\frac{1}{\mu} + (r-1)\left(\frac{1}{\gamma} - \left(\frac{1}{\gamma} + \frac{1}{\mu}\right)\frac{\Gamma(r-1, \gamma/\mu)}{\Gamma(r, \gamma/\mu)}\right)\right). \tag{8}$$

In Appendix C.2 in the e-companion, we use a discrete-event simulator to show that this is a good approximation for the hybrid rounding-responding mode.

In (8), the hospitalist's caseload impacts the expected LOS in two different ways: (i) higher caseload prolongs the time between visits in rounding mode and increases waiting time in responding mode, leading to additional delays, and (ii) higher caseload increases the amount of time spent rounding, leading to indirect delays because responding is a more efficient service mode. Because the proportion of time spent in rounding mode increases as there are more patients, for reasonable values of caseload (e.g., $r \in [10, 20]$), the behavior of the rounding-responding model resembles that of the rounding mode more closely. Therefore, many of the analytical results that we will present using only the rounding model in Section 6.2 are closely approximated in the rounding-responding model.

6. Caseload and Case-Mix Analysis

In this section, we first analyze the caseload and case-mix model in the general form without imposing any structure on patients' LOS and patient-provider interactions. We then extend these analyses to the case where patient-provider interactions are modeled in the rounding service mode. The full rounding-responding model is not analytically tractable. In later sections, we test the rounding-responding model numerically and show that the insights derived from the analytical results in this section hold numerically in the full rounding-responding model.

6.1. Analysis of the General Model

Substituting (2) and (3) in (1) yields

$$\min_{r^i,\, i\in\{1,2,\ldots,I\}} \sum_{i=1}^{I} \lambda^i S_N^i + \sum_{i=1}^{I}\left(nr^i\left(1-\frac{S_N^i}{S_H^i(r)}\right)\right), \tag{9}$$

which is a function of r^i only. In (9), $\sum_{i=1}^{I}\lambda^i S_N^i$ is the average hospital census in the absence of hospitalists, which does not depend on the decision variables r^i. The remainder of the expression is the reduction in hospital census due to the use of n hospitalists. Define $R_i(r_i, r) = nr^i(1 - S_N^i/S_H^i(r))$ to be the change in hospital census when each hospitalist has r patients, of which r^i are type i. $R_i(r_i, r)$ captures the trade-off between a high and a low hospitalist caseload: when hospitalists attend fewer patients, each patient's LOS is reduced by a larger amount (since $S_H^i(r)$ is increasing in r), but the effect is applied to those few patients. When hospitalists attend many patients, each patient's benefit is less, but it is aggregated over a greater number of patients.

We will show that under certain conditions, the patient types can be ordered according to the extent to which hospitalists reduce an individual patient's LOS. When this happens, addition of patient types into the hospitalists' case-mix should follow this ranking. First, we prove that if there is sufficient demand from the type of patient for whom hospitalists bring the greatest individual reduction in LOS, then hospitalists should only attend patients of that type.

Lemma 3 (Optimal Caseload and Case-Mix Without Demand Constraints). *Define the optimal caseload for a hospitalist that only serves type I patients as* $r_i^* = \arg\min_r R_i(r, r)$, $\forall i \in \{1, 2, \ldots, I\}$. *Further, define patient type* [1], *the type that achieves the greatest census reduction at its optimal caseload, as* $[1] = \operatorname{argmin}_{i\in\{1,2,\ldots,I\}} R_i(r_i^*, r_i^*)$. *If* $\omega_{[1]}(r_{[1]}^*) \geq r_{[1]}^*$, *then the optimal hospitalist caseload is* $r_{[1]}^*$, *of which all patients are type* [1].

In Lemma 3, type [1] is the patient type that satisfies $S_N^{[1]}/S_H^{[1]}(r_{[1]}^*) \geq S_N^i/S_H^i(r_i^*)$, $\forall i \neq [1]$ (in case of a tie, any of the tied types can be chosen). The ratio of nonhospitalist LOS to hospitalist LOS is largest for type [1] when the LOS reduction for each type i is evaluated at its own optimal caseload r_i^*.

When there is enough demand from this patient type, it is the only type included in the case-mix. When there is not sufficient demand from the highest-ranked patient type to reach the optimal hospitalist caseload, that is, $\omega_{[1]}(r_{[1]}^*) < r_{[1]}^*$, the solution is not as straightforward in general. However, if we fix the total caseload for each hospitalist, then we can characterize the optimal case-mix as follows.

Lemma 4 (Optimal Case-Mix for a Fixed Caseload with Demand Constraints). *Given total caseload* r, *let* $[i]_r = \arg\max_{i\notin\{[1]_r,[2]_r,\ldots,[i-1]_r\}} S_N^i/S_H^i(r)$, *where* $[i]_r$ *is the label of the type that has the ith largest ratio of nonhospitalist to hospitalist expected LOS given total caseload r. The optimal case-mix is obtained by adding patients in label order (i.e.,* $[1]_r$, $[2]_r, \ldots$*) until caseload r is reached.*

For a fixed caseload, patient types should be added to the hospitalist case-mix in decreasing ranked order of nonhospitalist to hospitalist LOS ratio until caseload r is achieved.

The results in this section were derived from the most general form of the model without imposing any structure on patients' LOS and providers' service modes. We showed that the hospitalist case-mix should include types with higher ratio of nonhospitalist to hospitalist LOS. But which factors lead this ratio to be high in practice? To answer this question, we need to impose more structure on the patients' LOS and on the interactions between patients and physicians.

6.2. Analysis of the Model with Coxian LOS in Rounding Service Mode

In this section, we analyze the model with the Coxian LOS and in the pure rounding mode as described in Section 5.3.1. We begin our analysis by considering a special case of the main problem, where all patient types have only one phase of care. We then extend this analysis to the case with multiple phases of care.

6.2.1. Single Phase with Discharge Delay. In this section, we consider a special case where all patient types are "simple" in the sense of having only one phase, which results in only one delay at the time of discharge. The Coxian phase-type distribution with only one phase is the exponential distribution. The expected LOS of a type i patient attended by a hospitalist would be

$$S_H^i(r) = \frac{r/\gamma}{1 - e^{-\mu^i(r/\gamma)}}.$$

Inserting the exponential LOS into the optimization model in (9), we obtain

$$\min_{r^i,\, i\in\{1,2,\ldots,I\}} \left\{ \sum_{i=1}^{I} \frac{\lambda^i/\xi}{1-e^{-\mu^i/\xi}} + \sum_{i=1}^{I}\left(nr^i\left(1-\left(\frac{\gamma/\xi}{r}\right)\left(\frac{1-e^{-\mu^i r/\gamma}}{1-e^{-\mu^i/\xi}}\right)\right)\right)\right\}. \tag{10}$$

First, we expand on Lemma 4 and prove that the optimal order by which patient types enter the hospitalist case-mix is independent of the caseload and depends solely on service rates.

Proposition 1 (Optimal Order of Types: Rounding Mode, Single Phase). *If all patient types have a service requirement with a single phase, then a patient type with a larger service rate* μ^i *will have a larger ratio of expected nonhospitalist to hospitalist LOS for all caseloads* $r \in (0, \gamma/\xi)$.

Proposition 1 suggests that for simple patients, the "best" patient type for hospitalists' operational effectiveness is the type with the highest service rate (shortest service requirement), which supports Hypothesis 1b. Applying Lemma 3, we further obtain the following.

Proposition 2 (Best Patient Type and Caseload: Rounding Mode, Single Phase). *If there is ample demand for patient type* [1], *that is,* $r^*_{[1]} \le \omega_{[1]}(r^*_{[1]})$, *where* $[1] = \mathrm{argmax}_{i \in \{1,2,\ldots,I\}} \mu^i$, *then the optimal case-mix of a hospitalist includes only type* [1] *patients, and the optimal caseload is*

$$r^*_{[1]} = \frac{\gamma}{\mu^{[1]}} \ln\left(\frac{\mu^{[1]}/\xi}{1 - e^{-\mu^{[1]}/\xi}}\right). \tag{11}$$

The optimal caseload for the best patient type is increasing linearly in the hospitalist visit rate, decreasing in the nonhospitalist visit rate, and decreasing in the patient service rate. This last observation supports Hypothesis 3. Complexity is held constant (at one phase): as the patient's service requirement decreases, the ideal caseload decreases.

If the demand constraint for the best type becomes binding, that is, $\omega_{[1]}(r^*_{[1]}) < r^*_{[1]}$, we can follow Proposition 1 and assign all of type [1] patients to hospitalists, and then move on to the second best type, that is, the type with the second largest service rate, [2]. In other words, we set $r_{[1]} = \omega_{[1]}(r)$ in (10), and solve it assuming only the caseload of type [2] can vary. Once the optimal caseload r^* is derived, we check the demand constraint for type [2]; if $r^*_{[2]} < \omega_{[2]}(r^*)$, then r^* is the optimal caseload. Otherwise, we set $r_{[2]} = \omega_{[2]}(r)$ in (10), repeat the same procedure for type [3], and so on. Proposition 3 generalizes this approach to all patient types and proves optimality.

Proposition 3 (Optimal Caseload and Case-Mix: Rounding Mode, Single Phase). *The optimization problem in* (10) *can be divided into I single-variable subproblems, one of which produces the optimal solution to the main problem. In each subproblem i, we only include the best i types of patients, that is, type* [1], [2], . . . , [i]. *We set* $r_j = \omega_j(r)$, $\forall j \in \{[1],[2],\ldots,[i-1]\}$, *and* $r_{[i]} = r - \sum_{j=[1]}^{[i-1]} \omega_j(r)$. *The ith subproblem becomes*

$$\min_{r \ge \sum_{j=[1]}^{[i-1]} \omega_j(r)} \; n \sum_{j=[1]}^{[i-1]} \omega_j(r) + \lambda^{[i]} S_N^{[i]} + n\left(r - \sum_{j=[1]}^{[i-1]} \omega_j(r)\right)\left(1 - \frac{S_N^{[i]}}{S_H^{[i]}(r)}\right). \tag{12}$$

The smallest i for which the demand constraint for type [i] *is not binding in the optimal caseload, that is,* $r^*_{[i]} < \omega_{[i]}(r^*)$, *produces the optimal caseload to the main problem. The optimal case-mix is to add patient types sequentially in decreasing order of service rates until the optimal caseload is reached. If such i does not exist, then assign all patients to hospitalists.*

Our analysis in this section provides support for two of our hypotheses in a special case where hospitalists only round and patients have a "simple" plan of care. We found that, for a fixed caseload, hospitalists are more effective in reducing average LOS when attending patients with shorter LOS, which supports Hypothesis 1b, and as LOS becomes shorter, the optimal caseload becomes smaller, which supports Hypothesis 3. We further showed that a straightforward stacking procedure that adds patients sequentially in decreasing order of service rates computes the optimal caseload and case-mix. Next, we analyze patients with multiple phases and interventions.

6.2.2. Multiple Phases and Interventions. Consider a patient with multiple treatment phases captured by a Coxian phase-type distribution. The expected LOS for a patient attended by a hospitalist would be

$$S_H^i(r) = \sum_{j=1}^{J} P_j^i \left(\frac{r/\gamma}{1 - e^{-\mu_j^i r/\gamma}}\right).$$

We first apply Lemma 3 to determine which patient type should be highest ranked when there is no demand constraint, as follows.

Lemma 5 (Optimal Caseload: Rounding Mode, Multiple Phases, Single Patient Type). *The objective function* (9) *for a single patient type with a J-phase Coxian LOS is strictly convex in r, and* r^* *is the solution to*

$$\left(\sum_{j=1}^{J} \frac{P_j}{1 - e^{-\mu_j r/\gamma}}\right)^2 = \left(\sum_{j=1}^{J} \frac{P_j/\xi}{1 - e^{-\mu_j/\xi}}\right)\left(\sum_{j=1}^{J} \frac{P_j \mu_j e^{-\mu_j r/\gamma}}{(1 - e^{-\mu_j r/\gamma})^2}\right). \tag{13}$$

Although the solution to (13) cannot generally be written in closed form, in the special case where phases are identically distributed, $r^* = (\gamma/\mu)\ln((\mu/\xi)/(1 - e^{-\mu/\xi}))$, which matches the optimal caseload for the single-phase patients considered in Section 6.2.1. This special case is illustrative in two ways. First, it provides the intuition that *the hospitalist's optimal caseload may be sensitive to complexity (as measured by number of phases) primarily through the average phase length.* Later, we show that this intuition is generally supported in numerical experiments with real data. Second, this special case suggests that Approximation 1 can be used to provide additional insights into the hospitalist ideal caseload and case-mix. Recall that in Approximation 1 we replaced the service rate of each phase μ_j by the average service rate $\bar{\mu}$, weighted according to the probability of visiting each phase P_j. We can apply the same approximation here, which would then lead to a heuristic caseload $\bar{r}$ as

$$\bar{r} = \frac{\gamma}{\bar{\mu}} \ln\left(\frac{\bar{\mu}/\xi}{1 - e^{-\bar{\mu}/\xi}}\right).$$

Although this heuristic does not provide the exact optimal solution, it simplifies the problem in a way that allows us to develop additional intuition about hospitalists' operational effectiveness. Recall that in Section 6.2.1 we found support for Hypotheses 1b and 3: among patients with a single phase, for a fixed caseload, hospitalists are more effective in reducing average LOS when attending patients with shorter LOS, and as LOS becomes shorter, the optimal caseload becomes smaller. Approximation 1 suggests similarly that *among patients with multiple phases, for a fixed caseload, hospitalists are more effective in reducing average LOS when attending patients with shorter weighted average phase length (larger $\bar{\mu}$), and as the average phase length becomes shorter, the optimal caseload becomes smaller.*

Two factors can lead to large $\bar{\mu}$ and thus a higher ranking according to the average service rate heuristic: short expected total service requirement and large expected number of phases. A high number of phases indicates a complex plan of care that requires many interventions and introduces more potential delays. Therefore, holding caseload constant, among patients with similar total expected service requirement, a hospitalist is more operationally effective when attending patients with more complex condition, supporting Hypothesis 1a, and as complexity of conditions increases, the ideal caseload decreases, supporting Hypothesis 2. On the other hand, among conditions with similar complexity, a hospitalist is more operationally effective when attending patients who have shorter service requirement, supporting Hypothesis 1b, and as service requirement decreases, the ideal caseload decreases, supporting Hypothesis 3. Both cases imply a shorter weighted average phase length, indicating that a patient requires more frequent attention. In Section 7, we show that patients with gastrointestinal (G.I.) hemorrhage (mean LOS = 4.5 days, mean procedures = 2.1) and patients with chest pain (mean LOS = 2.1 days, mean procedures = 0.5), are *both* prioritized over patients with diabetes (mean LOS = 4.3 days, mean procedures = 0.5). G.I. hemorrhage patients have a similar LOS as diabetes patients but have on average four times the number of procedures. Conversely, chest pain patients have a similar average number of procedures as diabetes patients but spend half as long in the hospital. The hospitalist has more frequent opportunities to reduce delays when attending either G.I. hemorrhage or chest pain patients than when attending diabetes patients.

We evaluate the performance of the average service rate heuristic in Appendix C.3 in the e-companion. We prove a special case where ranking patient types by $\bar{\mu}$ is optimal. For other cases, we run a number of numerical experiments using real patient data that show that ranking by $\bar{\mu}$ is suboptimal only when two types have very similar $\bar{\mu}$'s, in which case the suboptimal ranking has minimal impact on the solution performance because the two types are so similar. We also show that the heuristic caseload is very close to the actual optimal caseload. Approximation 1 is used for tractability in our numerical studies in Sections 7 and 8.

In this section, we highlighted several insights for the hospitalist's rounding service mode when patients' service requirements are Coxian phase-type. First, hospitalists are more effective in reducing delays for patients with many procedures, which dovetails with the generally held clinical belief that hospitalists are more effective at managing the care of complex patients and supports Hypothesis 1a. Second, among patients of similar complexity, hospitalists are more effective when their case-mix includes patients with short service requirements, which supports Hypothesis 1b, and to our knowledge has no clinical counterpart. Thus, to reduce hospital census and/or increase throughput, a hospitalist's case-mix should include patients with complex plan of care and patients with short LOS. Third, holding service requirements fixed, as complexity of conditions increases, hospitalist ideal caseload decreases, supporting Hypothesis 2; and holding complexity fixed, as LOS decreases, the ideal caseload decreases, supporting Hypothesis 3.

7. Mixing It Up: The Evidence for a Balanced Hospitalist Case-Mix

In the previous section, we derived analytical support for our hypotheses in a special case where hospitalists only round. In this section, we parameterize our model with data from the 2005 California State Inpatient Database (SID) and investigate the optimal ranking of patients in the ideal hospitalist case-mix numerically using the rounding and responding model. In a numerical study using actual patient data from a wide variety of patient types in a large data set, we confirm the insights from our previous analysis in the rounding-responding setting, and we expand upon them to answer other questions that are difficult to address in the analytical model. We previously analyzed patient heterogeneity in one dimension (complexity or service requirement), but what is the ideal case-mix when patient types are heterogeneous in both dimensions simultaneously? Are hospitalists most operationally effective when they focus on patient types that are similar in both dimensions, or should they mix it up and attend patients with a variety of complexities and service requirements?

7.1. Data Description

The SID data include inpatient discharges from 457 hospitals in California in 2005. Each record contains demographic information and data about the patient's plan of care, including diagnosis-related group (DRG), LOS, number of procedures (NPR), and day of each procedure relative to admission day. We use data for adult patients with medical conditions. Medical and surgical patients are often physically separated in the hospital and attended by different physician groups; as a robustness check, we repeat an analysis on adult surgical patients in Appendix B.2 in the e-companion. We remove obstetrics patients, and patients who died, left against medical advice, or had LOS longer than 60 days (these are not representative of the general adult patient population), resulting in 1,557,372 patient records. To ensure we have enough data to fit our LOS model for each DRG, we limit our analysis to the 45 most common DRGs. These common DRGs represent 20% of the 223 DRGs but account for 70% of the patient visits. The final sample includes 1,088,847 records. Table EC.10 in Appendix E in the e-companion provides summary statistics of the data by DRG.

7.2. Model Fitting and Validation

We consider each procedure as a touch-point when the physician intervenes to begin the next phase of care. Examples of procedures include biopsy and x-ray. Hospitalists coordinate these procedures with other care providers and take actions to respond to the results. The time between procedures allows us to apply the model of repeated physician-patient interactions using our phase-type approach. Because the SID records LOS and procedure times in terms of days, we add random noise, uniformly distributed on [0, 1] to LOS and procedure times so that the resulting time data are continuous.[2] We fit a Coxian phase-type distribution for each DRG by computing the expected time spent in each phase and transition probabilities. Since the SID data do not include the physician information, we cannot impute what portion of the time between two procedures are delays. We consider the entire time to be the service requirement. Thus, the estimated service rates are smaller than the actual ones. This bias may affect the ranking of individual DRGs in the optimal case-mix. However, since we are interested in the characteristics that determine the rankings, and not the ranking of individual DRGs, this bias does not affect our main insights.

Table 2 reports the negative log likelihood (–*LL*) (smaller is better) and the Akaike Information Criterion (AIC) statistic (smaller is better) for the top 10 DRGs that enter the hospitalist case-mix (see Section 7.3) for our Coxian approach using the procedure time and LOS data, and some other distributions that are commonly used in the literature: Exponential, Weibull, and Lognormal, using only the LOS data. The Coxian approach provides the best fit for DRG 125 and 124, the second best fit for DRG 122, 174, and 395, and third best fit for the rest of them. Clearly, no one distribution is best for every DRG, and neither is our goal to determine the *best* fit for the data. The Coxian phase-type distribution is the only model that allows us to capture the dynamics of patients' progress and their repeated interactions with their providers. Hence, our goal is only to show that the Coxian model is a *reasonable* fit for the LOS data, and the overall results support this contention.

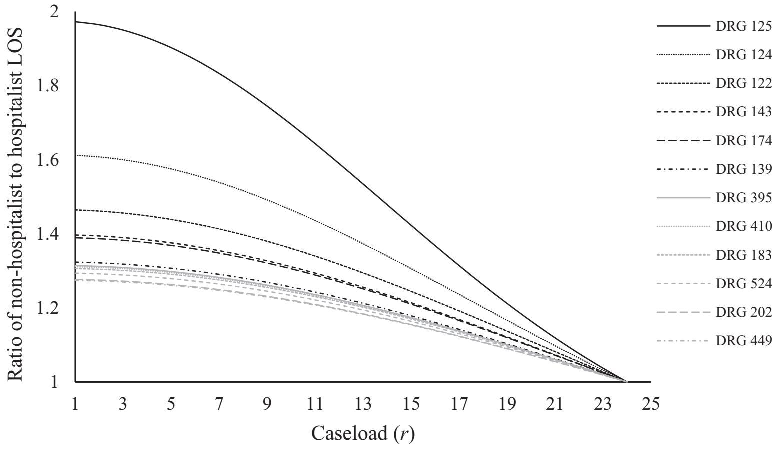
7.3. Mixing It Up: Optimal Hospitalist Case-Mix

In this section, we test Hypotheses 1a and 1b using the rounding-responding model of hospitalist-patient interactions, and the 45 most common DRGs from the SID data. Our goal is to understand the characteristics that determine the ranking of different patient types in the optimal case-mix. Recall that Hypotheses 1a and 1b describe the hospitalist's ideal case-mix while holding the caseload constant. In Lemma 3, we proved that, while holding the caseload constant, patient types should be ranked in decreasing order of nonhospitalist to hospitalist expected LOS in that

Table 2. Goodness of Fit Statistics for the Top 10 DRGs in Section 7

	–Log-likelihood				AIC			
DRG	Coxian	Exponential	Weibull	Lognormal	Coxian	Exponential	Weibull	Lognormal
125	27,971	29,854	28,330	28,868	55,972	59,710	56,663	57,738
124	37,149	39,074	37,940	37,621	74,335	78,150	75,882	75,245
122	24,689	25,637	24,028	24,883	49,405	51,275	48,057	49,768
143	115,030	118,240	106,970	113,086	230,079	236,482	213,943	226,173
174	88,750	95,020	89,158	85,360	177,530	190,042	178,317	170,722
139	24,482	24,957	23,045	23,826	48,981	49,916	46,092	47,654
395	47,495	48,964	48,044	46,751	95,016	97,930	96,090	93,504
410	32,507	34,280	32,363	31,629	65,037	68,562	64,727	63,259
183	46,580	47,623	44,242	45,347	93,179	95,249	88,486	90,696
524	30,021	30,355	28,076	27,626	60,068	60,711	56,153	55,254

Figure 6. Ratio of Expected Nonhospitalist to Hospitalist LOS at Different Caseloads



caseload. To calculate the expected nonhospitalist and hospitalist LOSs, we set $\xi = 1$ and $\gamma = 24$, which correspond to a nonhospitalist visiting the hospital once a day and a hospitalist spending 1/24th of the shift on each patient encounter. Since visit rates are identical across patient types, the choice of these parameters does not affect the ranking of the patients in the case-mix. For computational tractability, we approximate each patient's expected LOS using Approximation 1 and (8); that is, we replace μ in (8) by $\bar{\mu}$ for each patient type, and we multiply (8) by the average number of phases. Throughout this section, the term "optimal case-mix" refers to the optimal case-mix when LOSs are approximated by Approximation 1.

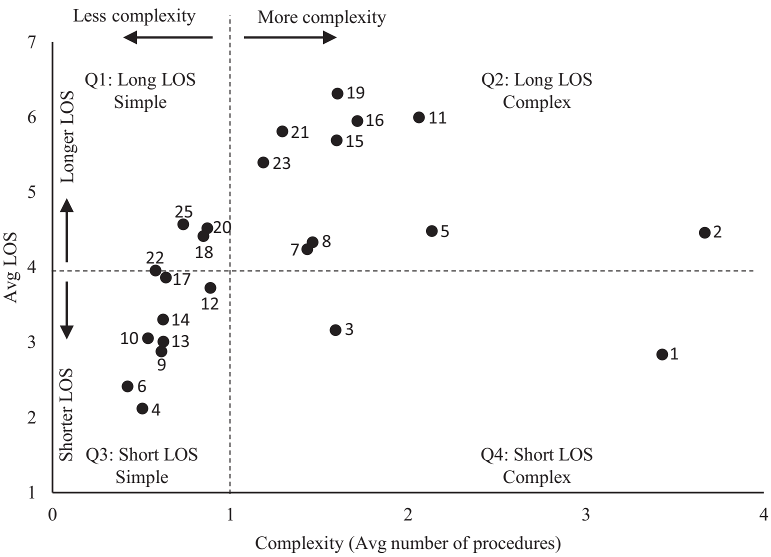
Figure 6 shows the ratio of expected nonhospitalist to hospitalist LOS at different caseloads between 1 and 24 patients for the top 12 DRGs in the optimal case-mix. At $r = 24$, the ratio is equal to 1 for all DRGs. In line with Proposition 1, we find that the optimal ranking is independent of the caseload (the ratios do not intersect when $r < 24$) and is in decreasing order of the average service rates $\bar{\mu}^i$. In Table 3, we report the top 12 DRGs. We classify the service requirement of each DRG as "short" if the average LOS is less than four days, and "long" otherwise. We classify the complexity of each DRG as "simple" if the average number of procedures (excluding discharge) is less than one, and "complex" otherwise.

In Figure 7, we plot the top 25 DRGs that enter the hospitalist case-mix by their average number of procedures (complexity) and their average LOS. The label next to each point indicates the DRG's rank (with 1 being the best rank). The fourth quadrant contains DRGs with both short LOS and a complex plan of care. Hospitalist care is most operationally effective for these DRGs, as it ensures that they are closely monitored and that the transition between phases is timely. However, there are not many DRGs that require numerous procedures in a short amount of time. The DRG that enters the hospitalist case-mix after the fourth quadrant DRGs (rank 4) is DRG 143: chest pain, which is a simple condition, having the fifth smallest number of procedures. DRG 143 is attractive because those patients have the shortest

Table 3. Summary Statistics of Top 12 Ranked DRGs in the Optimal Case-Mix

Rank	Diagnosis-related group (DRG)	Average LOS (days)	Average NPR	$\bar{\mu}$	Service	Complexity
1	125: Circ. disorders except ami, w card cath w/o complex diag.	2.84	3.45	1.56	Short	Complex
2	124: Circ. disorders except ami, w card cath & complex diag.	4.46	3.69	1.05	Long	Complex
3	122: Circ. disorders w ami w/o major comp, discharged alive	3.16	1.61	0.82	Short	Complex
4	143: Chest pain	2.12	0.52	0.71	Short	Simple
5	174: G.I. hemorrhage w cc	4.48	2.15	0.70	Long	Complex
6	139: Cardiac arrhythmia & conduction disorders w/o cc	2.41	0.43	0.59	Short	Simple
7	395: Red blood cell disorder	4.24	1.44	0.57	Long	Complex
8	410: Chemotherapy w/o acute leukemia as secondary diag.	4.33	1.48	0.57	Long	Complex
9	183: Esoph., gast. & misc digest disorders age > 17 w/o cc	2.88	0.63	0.56	Short	Simple
10	524: Transient ischemia	3.01	0.64	0.54	Short	Simple
11	202: Cirrhosis & alcoholic hepatitis	6.00	2.06	0.51	Long	Complex
12	449: Poisoning & toxic effects of drugs age > 17 w cc	3.72	0.89	0.51	Short	Simple

Figure 7. Average LOS vs. Complexity of Top 25 DRGs and Their Ranking in the Optimal Case-Mix



Note. The 24th rank has an average LOS of 12.1 days and is not shown for the sake of space.

average LOS, approximately two days. These patients experience long discharge delays if attended by a nonhospitalist who rounds once a day. A hospitalist is able to substantially reduce this delay. We observe that criteria for ranking have shifted from the most complex patients to the patients with the shortest service requirement. The next best DRGs are either in the second quadrant (long and complex) or the third quadrant (short and simple). The first quadrant contains DRGs with long average LOS and simple plan of care, for which hospitalists are the least operationally effective. The DRGs that are ranked 26th to 45th (not shown in the graph) mostly fall in this quadrant.

Figure 7 supports the suggestions that, among conditions with similar service requirement, hospitalists are most operationally effective for complex ones (Hypothesis 1a), and, among conditions with similar complexity, hospitalists are most operationally effective for those with short service requirement (Hypothesis 1b). When patients are heterogeneous in both dimensions, hospitalist case-mix should include a mix of patients with short-simple and long-complex conditions, rather than focusing solely on one quadrant. Attending some short-simple patients next to the long-complex ones helps reduce congestion in the system, which is an operational outcome that has not been considered in the clinical literature.

8. Case Study: Examining Case-Mix, Caseload, and Coverage at Two Partner Hospitals

We return to the two case study hospitals described in Section 2. By comparing the solution of our model with the real hospitalist staffing practice at these hospitals, we can study the cost-benefit trade-off of the hospitalist model. We also compare the use of hospitalists in two different types of hospitals.

8.1. Case 1: Community Hospital

Community Hospital provided us with data for calendar year 2017. Excluding obstetric and pediatric admissions, there are 4,626 records representing patients with 195 different DRGs. The data include admission and discharge dates, LOS, attending provider's ID and specialty, primary care provider's ID, and DRG. Unfortunately, the data do not include the number and timing of procedures. We use the average number of procedures (NPRs) from the SID data to approximate the average NPRs in this hospital. We adjust the NPRs proportional to the average LOS for each DRG to maintain the same number of procedures per unit time as the SID data. Table 4 summarizes the data.

Community Hospital employs five hospitalists, with three on duty on any given day. Each on-duty hospitalist requires 1.7 full-time equivalents (FTEs). Community Hospital had an average census of 59 patients. Each hospitalist managed seven patients on average, including both medical and surgical patients. Collectively, this accounts for 35% of patients. Of the remaining patients, 20% were attended by PCPs and 45% by specialists.

To calibrate our model, for each DRG i, we estimate $\bar{\mu}^i$ using the expression for the expected LOS when attended by a nonhospitalist,

$$S_N^i = \left(\frac{1/\xi}{1-e^{-\bar{\mu}^i/\xi}}\right)\sum_{j=1}^{J} P_j^i.$$

Table 4. Most Common DRGs at Community Hospital: Summary Statistics and Model Outputs

	All			Nonhospitalists			Hospitalists			Model output		
DRG number	Sample	Average LOS	Average census	Sample	Average LOS	Average census	Sample	Average LOS	Average census	Average census (All)	Average census (NH)	Average census (H)
416	515	6.2	8.8	299	6.9	5.6	216	5.4	3.2	9.3	5.6	3.7
88	393	3.9	4.2	320	3.9	3.4	73	4.1	0.8	4.1	3.4	0.7
127	336	4.7	4.3	246	4.6	3.1	90	4.9	1.2	4.1	3.1	1.0
316	148	4.3	1.7	67	4.8	0.9	81	3.9	0.9	1.8	0.9	0.9
89	135	4.0	1.5	90	4.2	1.0	45	3.5	0.4	1.5	1.0	0.5
294	118	3.8	1.2	65	3.7	0.7	53	4.0	0.6	1.1	0.7	0.5
138	105	3.6	1.0	70	3.6	0.7	35	3.6	0.3	1.0	0.7	0.3
183	103	3.5	1.0	47	3.5	0.5	56	3.4	0.5	0.9	0.5	0.5
278	101	3.8	1.0	49	3.8	0.5	52	3.7	0.5	1.0	0.5	0.5
14	100	4.3	1.2	61	4.3	0.7	39	4.2	0.5	1.1	0.7	0.4
174	85	3.7	0.9	40	4.2	0.5	45	3.3	0.4	0.9	0.5	0.4
321	67	3.5	0.6	37	3.9	0.4	30	3.0	0.2	0.7	0.4	0.3
415	66	10.9	2.0	33	11.3	1.0	33	10.6	1.0	1.9	1.0	0.9
297	63	3.2	0.6	32	3.6	0.3	31	2.8	0.2	0.6	0.3	0.3
Total	2,335	4.7	30.1	1,456	4.8	19.3	879	4.5	10.8	30.1	19.3	10.8

We use S_N^i from the data and solve this expression to obtain

$$\bar{\mu}^i = -\xi \ln\left(1 - \frac{\sum_{j=1}^{J} P_j^i}{\xi S_N^i}\right).$$

We then assess the quantitative predictive accuracy of our model using the hospital's census. In other words, we compare the average census of the hospital from the data with the one from our calibrated model. To do so, we set $n = 3$ hospitalists and $r = 7$ patients according to the current practice, with the case-mix following that of the empirical data. Moreover, we set $\gamma = 24$, which corresponds to a hospitalist spending 1/24th of the shift on each patient encounter. With a typical 12-hour hospitalist shift, this corresponds to visiting patients on half-hour intervals, which is consistent with the literature and our interviews with hospitalists. We conduct sensitivity analysis on the visit interval in Appendix B.3 in the e-companion. Recall that we only model the day shifts and assume that patients do not generate requests overnight (see Section 2). We tune the nonhospitalist visit rate ξ to maximize the predictive accuracy of the model, that is, to minimize the absolute difference between the actual and predicted average hospital census. This is achieved at $\xi = 1.7$ visits per day, which is reasonable because many of the nonhospitalist patients were attended by specialists, who often conduct rounds more than once per day. To assess the predictive accuracy of the model, we then compare the average hospitalist census of each DRG from the data with the same measure from our calibrated model. We find that, excluding DRGs with fewer than two hospitalist-attended observations, the hospitalist census predicted by the calibrated model (rightmost column in Table 4) falls within the 95% confidence interval for the mean hospitalist census for 96% of the DRGs, indicating that the model predicts the actual census well.

After calibrating the model, we analyze Community Hospital's census under the counterfactual setting where the caseload and case-mix are determined according to our model. We solve the model with objective function (9), decision variables (r^i), and the demand constraints in Mathematica 11, using the `NMinimize` function. Table 5 shows the top 10 ranked DRGs according to the model and their characteristics. The two rightmost columns report the proportion of patients from each DRG that were actually attended by hospitalists in the data and the proportion that should be attended by hospitalists in the optimal solution at the current staffing level. Although our model suggests that all the patients of these top 10 DRGs should be attended by hospitalists at the current staffing level, the majority of them were not actually attended by hospitalists according to the second-to-last column in Table 5. The DRGs that have particularly low hospitalist coverage have short LOS, suggesting that hospitalists at Community Hospital often attend complex patients (who tend to have long LOS). Eschewing these shorter LOS patients represents a missed opportunity to better impact LOS via a more varied case-mix.

Table 6 reports the current staffing policy, the model output for each n in $\{0, 1, \ldots, 7\}$, and the aggregate percent of patients covered by hospitalists in the model output. The optimal caseload typically ranges between eight and nine patients, indicating that hospitalists may currently be underloaded with a caseload of seven. With the current $n = 3$, a slight increase in the caseload of each hospitalist to 8.2 and using the optimal case-

Table 5. Summary Statistics of Top 10 Ranked DRGs in the Optimal Case-Mix at Community Hospital

							% attended by hospitalists	
Rank	Diagnosis-related group (DRG)	Average LOS (days)	Average NPR	$\bar{\mu}$	Service	Complexity	In the data	In the optimal solution[a]
1	310: Transurethral procedures w mcc	3.8	2.7	1.50	Short	Complex	30%	100%
2	217: Wnd debrid & skn grft exc hand …	11.2	5.1	1.46	Long	Complex	50%	100%
3	122: Circulatory disorders except ami, …	2.8	1.7	1.39	Short	Complex	16%	100%
4	175: G.I. hemorrhage w/o cc/mcc	2.7	1.7	1.38	Short	Complex	20%	100%
5	208: Disorders of the biliary tract …	2.8	1.0	1.27	Short	Simple	40%	100%
6	77: Periph/cranial nerve & other nerv …	5.4	3.4	1.24	Long	Complex	50%	100%
7	143: Chest pain	1.9	0.6	1.23	Short	Simple	26%	100%
8	209: Major joint replacement …	3.3	1.7	1.23	Short	Complex	9%	100%
9	477: Nonextensive o.r. proc …	5.2	3.0	1.18	Long	Complex	15%	100%
10	281: Trauma to the skin, subcut tiss …	3.1	1.6	1.18	Short	Complex	36%	100%

[a] At the current staffing level $n = 3$.

mix would result in a 3.1% reduction in hospital census. Whether $n = 3$ is the right number of hospitalists depends on how much money the hospital saves from a reduced census. These savings can occur through revenue increase by admitting new patients into the freed space or cost reduction by reducing the unit size (e.g., number of beds and/or staff). We conduct a cost-benefit analysis for each staffing level n based on an average total annual compensation of $300,000 per FTE (Today's Hospitalist 2018) and the observed 1.7 FTEs required per staffed hospitalist. To calculate the annual benefits, we first calculate the cost savings or revenue improvement associated with one day reduction in LOS for one patient. We call this quantity the *marginal benefit per patient day* (MBPPD). Based on the current staffing policy at Community Hospital, we impute that they save $1,766 per patient-day reduction. We then multiply the marginal benefit per patient day by the size of the census reduction associated with each staffing policy and the number of days per year (365 days) to get the annual benefits. In Figure 8, we compare the annual marginal cost and benefit of a hospitalist as a function of n for the imputed MBPPD of $1,766, along with two other estimated values for the MBPPD from the literature. Taheri et al. (2000) estimate the inpatient cost during the last day of hospitalization to be $660 (in 2019 dollars). An alternative value is estimated by the Kaiser Family Foundation (2016), which estimates the hospital adjusted expenses per inpatient day to be $2,435 (in 2019 dollars). At the low end of the range suggested by the literature, no hospitalists should be staffed; at the high end, six hospitalists should be staffed. At the imputed MBPPD, the hospital should staff five hospitalists each day instead of three if they use the optimal case-mix. At this staffing level, 79% of patients would be covered, an increase of 44 percentage points over the current policy.

In the last column of Table 6, we report the annual net savings compared with the current staffing policy at the imputed MBPPD (we calculate the net benefits of switching from the current policy to the policy represented in that row after subtracting the annual costs from the annual benefits). By increasing the staffing level to five hospitalists per day, and providing hospitalist care to the right number and mix of patients, we estimate that the hospital can save $1.31 million annually. At the current staffing level of three hospitalists per day, we estimate that the hospital can save $1.18 million by assigning the optimal case-mix and caseload to each hospitalist. This highlights the importance of hospitalists' caseload and case-mix design for ensuring efficient hospital care delivery.

8.2. Case 2: Academic Hospital

Academic Hospital provided us with data containing all adult nonobstetric hospitalist admissions in 2017. In contrast to Community Hospital, this hospital has a large patient volume and has dedicated hospitalists for medical and surgical patients. Although hospitalists who are dedicated to one type of patients may occasionally end up attending patients from the other type due to overflow or patient transfers, the hospital's primary goal is to keep the hospitalists dedicated to medical or surgical floors. We analyze the group of hospitalists who primarily see medical patients. For our study, we only consider medical DRGs for these hospitalists, as the overflow patients are an exception and should not be considered as part of the "ideal" case-mix for medical hospitalists. This includes 7,865 admissions from 182 medical DRGs. The data include admission and discharge dates, LOS (average 6.4 days), attending provider's ID, DRG, and the total hospitalization cost. Similar to the community hospital's data, this data set does not include the number and timing of procedures. We use the SID data to approximate the average NPR of each DRG while adjusting these numbers proportional to the LOS. We analyze the 100 most common medical DRGs, which make up 7,474 admissions (95%). On average, $n = 9$ hospitalists attended 131 patients, with an average caseload of $r = 14.5$ patients per

Table 6. Current and Optimal Staffing Policies at Community Hospital

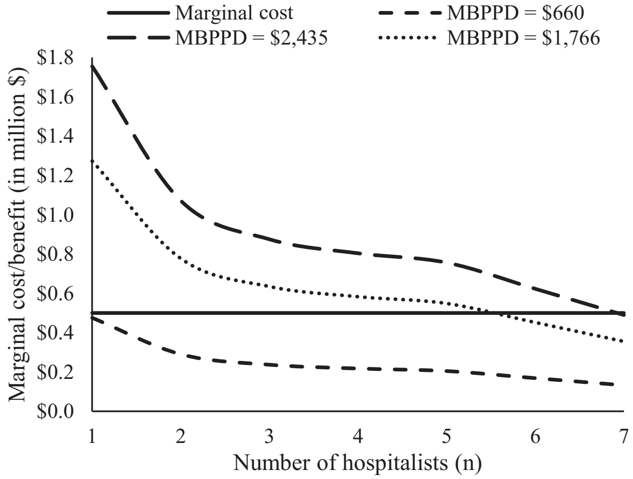
n	FTEs	Caseload	Hospital census	Improvement in census[a]	Hospitalist coverage[b]	Annual net saving[c]
Current policy						
3	5.0	7.0	59.0		35%	
Optimal policy						
0			61.3	−3.9%		
1	1.7	8.6	59.4	−0.7%	14%	$0.77 million
2	3.3	8.4	58.1	1.5%	29%	$1.05 million
3	5.0	8.2	57.2	3.1%	43%	$1.18 million
4	6.7	8.6	56.2	4.7%	61%	$1.27 million
5	8.3	8.8	55.4	6.1%	79%	$1.31 million
6	10.0	8.4	54.7	7.3%	92%	$1.27 million
7	11.7	7.3	54.2	8.1%	94%	$1.12 million

[a] Compared with the current hospital census.
[b] % of patients cared for by hospitalists.
[c] Compared with the current policy at the imputed MBPPD = $1,766.

hospitalist. The hospitalists worked seven-day-on-seven-day-off schedules, which requires 2.0 FTEs per staffed hospitalist.

For each DRG, we use the average LOS, the approximated average NPR, and $r = 14.5$, to numerically estimate $\bar{\mu}_i$ using (8), minimizing the mean squared error. Unfortunately, Academic Hospital's data set did not include nonhospitalists' admissions, so we could not assess the quantitative predictive accuracy of the model. However, Academic Hospital's data set included the actual cost of each patient, from which we can determine the cost per patient day. The actual average cost per patient day is $1,912 in the data. We impute the nonhospitalist visit rate from the hospitalization cost data as follows. According to our interviews with the hospital's practitioners, Academic Hospital tends to work near capacity most of the time, and an empty bed is immediately filled with a new admission. Therefore, a reduced LOS translates into additional revenues. The hospital has maintained a 10% operating margin in recent years. As such, we estimate the marginal benefit per patient day (MBPPD) to be $2,125. For the current staffing policy (18 FTE hospitalists) to be financially feasible, the annual benefits of 18 FTE hospitalists should be at least as large as the annual costs. Suppose that the average census in the absence of hospitalists would have been Υ patients. Then $18 \times \$300{,}000 \le (\Upsilon - 131) \times \$2{,}125 \times 365 \rightarrow 138 \le \Upsilon$. Therefore, the census in absence of hospitalists should be at least 138 patients. This corresponds to a nonhospitalist visit rate of at most 1.6. We use 1.6 as the imputed nonhospitalist visit rate and solve the model for $n = 0, 1, 2, \ldots, 13$ hospitalists. Table 7 reports the top 10 ranked DRGs in the optimal case-mix and their characteristics, which again follow the "mixing it up" principle.

Figure 8. Marginal Cost/Benefit Analysis for the Optimal Staffing Level at Community Hospital



The optimal per-hospitalist caseload is around 10 patients (Table 8, column 3), which is approximately 4.5 patients fewer than the observed caseload. This indicates that the hospitalists at the academic hospital may be overloaded. The optimal solution would substantially reduce the portion of the shift spent on rounds from 60% to 40%, because a lower caseload allows hospitalists to spend more time in the more efficient responding mode. Figure 9 shows the marginal cost and benefit of adding the last hospitalist in different staffing levels for the MBPPDs reported in the literature, and for MBPPD = $2,125, which we computed from the hospital's cost data and operating margin, and the fact that the hospital is mostly capacity-constrained. At the computed MBPPD of $2,125, the optimal staffing level is 11 hospitalists, an increase of two compared with the current policy. The addition of hospitalists allows for lower hospitalist caseloads and hence greater effectiveness in reducing delays, without reducing coverage too much. With 11 hospitalists with optimized case-mix and caseload, there will be an 84% hospitalist coverage in total

Table 7. Summary Statistics of Top 10 Ranked DRGs in the Optimal Case-Mix at Academic Hospital

Rank	Diagnosis-related group (DRG)	Avg LOS (days)	Avg NPR	$\bar{\mu}$	Service	Complexity
1	175: G.I. hemorrhage w/o cc/mcc	3.1	1.8	1.15	Short	Complex
2	122: Circulatory disorders except ami, …	5.7	3.4	0.91	Long	Complex
3	66: Epistaxis w mcc	4.9	2.7	0.90	Long	Complex
4	281: Trauma to the skin, subcut tiss …	4.9	2.5	0.86	Long	Complex
5	324: Urinary stones w/o esw lithotripsy …	5.3	2.6	0.78	Long	Complex
6	185: Dental & oral diseases w mcc	3.2	1.1	0.76	Short	Complex
7	251: Fx, sprn, strn & disl except femur …	6.2	3.0	0.74	Long	Complex
8	143: Chest pain	3.1	0.9	0.74	Short	Simple
9	524: Transient ischemia	2.7	0.7	0.73	Short	Simple
10	189: Other digestive system diagnoses …	4.2	1.7	0.72	Long	Complex

(Table 8, column 6). The LOS reduction for the remaining 16% of patients is simply not enough to offset either the negative effects of increased caseload or the cost of additional hospitalist(s). Increasing the staffing level to 11 hospitalists per day, and providing hospitalist care for the right number and mix of patients can lead to a 2.9% reduction in hospital census (Table 8, column 5), which we estimate to create $1.72 million saving annually (Table 8, column 7). Keeping the current number of hospitalists and simply optimizing the case-mix and volume results in a 1.7% reduction in hospital census or a savings of $1.69 million, mainly by reducing the current hospitalist caseload.

8.3. Comparison of the Optimal Policies at Community and Academic Hospitals

The current hospitalist staffing policies at Community and Academic Hospitals are quite different. This is partly driven by qualitative differences in the demand at the two hospitals. Community Hospital is a small hospital with a low patient volume. Patients who visit this hospital are less sick and stay for a shorter time (average 3.6 days), which is common in community and critical access hospitals, as sicker patients are often transferred up to a higher-level care. Community Hospital's three hospitalists cover 35% of the patients. On the other hand, Academic Hospital is large and tends to work near its capacity most of the time. Patients who visit Academic Hospital tend to be sicker and stay longer (average 6.4 days). Hospitalists cover almost all patients in the academic hospital.

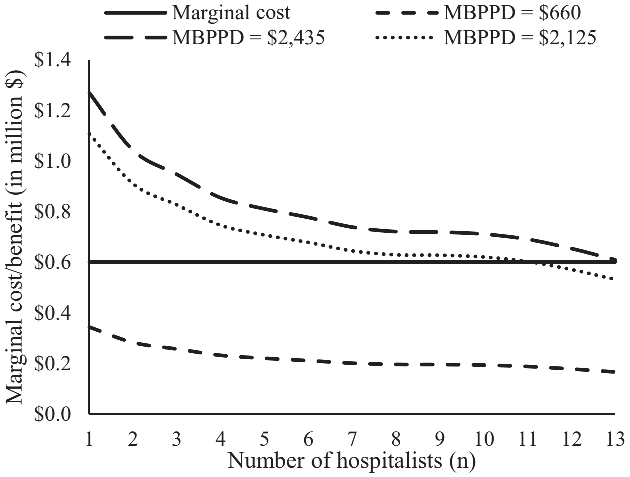
8.3.1. Optimal Case-Mix. Hospitalists at Community Hospital tend to attend a mix of medical and surgical patients due to low patient volume. Hospitalists at Academic Hospital are dedicated to medical or surgical patients. Regardless of this distinction, we find that in both hospitals, the hospitalists should attend a mix of patients with short-simple and long-complex conditions. This would ensure fast treatment and

Table 8. Current and Optimal Staffing Policies at Academic Hospital

n	FTEs	Caseload	Hospital census	Improvement in census[a]	Hospitalist coverage[b]	Annual net saving[c]
Current policy						
9	18	14.5	131		100%	
Optimal policy						
0			137.7			
1	2	9.9	136.3	−4.0%	7%	$0.73 million
2	4	9.9	135.1	−3.1%	15%	$1.04 million
3	6	9.8	134.0	−2.3%	22%	$1.26 million
4	8	9.6	133.1	−1.6%	29%	$1.41 million
5	10	9.8	132.1	−0.8%	37%	$1.52 million
6	12	9.6	131.3	−0.2%	45%	$1.59 million
7	14	9.8	130.4	0.5%	53%	$1.64 million
8	16	9.8	129.6	1.1%	61%	$1.67 million
9	18	9.8	128.8	1.7%	69%	$1.69 million
10	20	9.9	128.0	2.3%	77%	$1.71 million
11	22	9.7	127.2	2.9%	84%	$1.72 million
12	24	9.6	126.5	3.4%	91%	$1.69 million
13	26	9.3	125.8	4.0%	96%	$1.62 million

[a] Compared with the current hospital census.
[b] % of patients cared for by hospitalists.
[c] Compared with the current policy at the computed MBPPD = $2,125.

Figure 9. Marginal Cost/Benefit Analysis for the Optimal Staffing Level at Academic Hospital



recovery progress for complex patients (Hypothesis 1a) and fast discharge for simple patients (Hypothesis 1b), thereby maximizing hospital throughput. At Community Hospital, hospitalists would have a mixed medical-surgical case-mix. At Academic Hospital, hospitalists on the medical floor should mix it up within their dedicated patient groups.

8.3.2. Optimal Caseload. The optimal caseload is around 8.5 patients at Community Hospital and 10 patients for hospitalists dedicated to medical patients at Academic Hospital. This is actually quite different—a difference of 1.5 patients translates to roughly 45 more minutes spent on rounds. Given that patients at Academic Hospital tend to be sicker than patients at Community Hospitals, it could be counterintuitive that the hospitalists at Academic Hospital should have higher caseloads. However, we can understand this result by observing that sicker patients tend to have longer LOS. The average LOS of DRG 122 patients is 2.8 days in Community Hospital and 5.7 days in Academic Hospital. Longer LOS for the same DRG can mean that a patient requires more recovery time, during which they need a higher level of nursing care, but not necessarily more interventions per day. Relatively speaking, patients at Community Hospital recover quickly and therefore need interventions (including discharges) at greater frequency. A lower hospitalist caseload allows hospitalists to spend more time in responding mode and ensures that discharges are processed in a timely fashion.

We also observe an interaction between case-mix and volume. Because Community Hospital has lower patient volume, hospitalists attend both medical and surgical patients; the latter require more interventions per day. In turn, the different mix of patients in Community Hospital means that average phase lengths are shorter, and hence a lower caseload allows those hospitalists to be more operationally effective (Hypotheses 2 and 3). In contrast, hospitalists at Academic Hospital are highly overloaded.

8.3.3. Optimal Staffing Level. The optimal staffing level depends on the magnitude of LOS reduction and its financial value to the hospital. We imputed similar nonhospitalist visit rates in the two hospitals. If hospitalists had similar caseloads in the two hospitals, then LOS reduction for a given patient would be similar regardless of hospital. However, neither the patients' characteristics nor the hospitalists' optimal caseload are similar in the two hospitals. In our case study, hospitalist care leads to larger LOS reductions at Community Hospital because of shorter average phase lengths and lower caseload. On the other hand, LOS reduction has more financial value at Academic Hospital, which is mostly capacity-constrained, compared with Community Hospital, which is mostly demand-constrained. Taking all these factors into account, we find that Community Hospital should provide hospitalist coverage for 79% of patients, and Academic Hospital should provide hospitalist coverage for 84% of its medical patients. We suggest that both hospitals are understaffed, but for different reasons. Community Hospital should increase their staffing so that they can increase hospitalist coverage to more patients. Academic Hospital should increase staffing to lower each hospitalist's caseload. Although this reduces the total hospitalist coverage, it leads to a lower hospital census by making hospitalists more responsive to the patient types that are more prone to delays.

9. Limitations and Future Research

Like all studies that rely on analytical models and secondary data sources, our study has some limitations. For analytical tractability, we assumed that the time allocated to each patient for visiting and coordinating care is independent of the hospitalist's caseload and the patient's conditions. The first part is supported by Tipping et al. (2010), who report the results of a time-motion study showing that caseload did not change the amount of time hospitalists spent with each patient. The second part implies that a patient's LOS reduction from being attended by a hospitalist is only affected by the number of patients in his or her hospitalist's caseload. Future studies can explore how the time spent per patient varies across patient types, and incorporate that into the caseload and case-mix design. We also assumed that hospitalists alternate between rounding and responding service modes, which is in line with our observations of and interviews with practicing hospitalists in the United States.

If hospitalists can respond in the middle of rounds, then hospitalist care will be more valuable than our study suggests. Future studies can investigate these nuances in hospitalist workflow and their impacts on the caseload and case-mix.

In this study, we only considered the impact of hospitalist caseload and case-mix on LOS, which is an outcome important to hospital management from an operations perspective. However, there are other aspects of a hospitalist's job such as noncare obligations (e.g., teaching) that may impact their interactions with patients, and the hospitalist model of care may affect outcomes other than LOS. For example, a hospitalist may enhance communications, and in turn improve nurse and patient satisfaction (Pressel et al. 2008, Fulton et al. 2011). This may have long-term financial implications for hospitals. If this is the case, then we may have underestimated the financial benefit of the hospitalist model in our case studies. Future studies can consider other aspects of hospitalists' work, and hospitalists' effect on nonoperational outcomes.

This study provides the first analytical results regarding the *ideal* caseload and case-mix for hospitalists. As an initial exploration of operations management in this domain, we scoped the work solely as a tactical study of caseload and case-mix. Another important area of research lies in the operational decisions of how to dynamically assign randomly arriving patients to hospitalists. Although our work provides guidance regarding what case-mix and caseload to target, it does not address the question of how to achieve those targets. Future studies should investigate implementable policies that drive toward the target caseload and case-mix while accounting for the stochasticity and complexities in daily hospital operations.

10. Conclusion

Despite the growth of the hospitalist specialty, most academic studies of the hospitalist model have analyzed patient-oriented rather than system-oriented outcomes. To the best of our knowledge, we are the first to study hospitalist caseload and case-mix from an operational perspective. Our results partly comport with the generally held belief among clinicians that hospitalists are medically most effective when attending patients with complicated conditions. Such patients tend to have long service requirements. However, we show that for maximum operational effectiveness, hospitalists should also attend some patients that are operationally simple and have short service requirements to reduce discharge delays (a benefit not currently considered in the clinical literature). We find that hospitalist optimal caseload mostly depends on the frequency by which patients need interventions. This means that a hospitalist who cares for complex and long-stay patients may need to have a larger caseload than a hospitalist who cares for simple and short-stay patients. Combining these operational and clinical perspectives is key to maximizing the benefit of this growing medical specialty on hospital operations, while maintaining a high standard of care.

We conducted a case study of a small community hospital and a large academic hospital. The current practice used at each hospital is far from optimal. Community Hospital needs to staff more hospitalists to increase the proportion of patients covered. Small hospitals such as this one may find it difficult to justify having hospitalists for their patients, who are less sick than the typical patients seen at academic hospitals, for two reasons: they may underestimate the benefits of hospitalists in speeding up discharges, or they have not seen these benefits from their current hospitalists because of their suboptimal caseload and case-mix. On the other hand, Academic Hospital needs to staff more hospitalists but substantially lower its hospitalists' caseload, despite the fact that this reduces the proportion of patients covered. Academic medical centers with sicker patients such as this one may intuitively feel that all patients should be attended by a hospitalist. However, some patient types may not need many interventions and therefore could be attended by a nonhospitalist provider. Including these patients in hospitalists' caseloads limits their responsiveness. In our case study, we made another counterintuitive observation: the optimal hospitalist caseload at Community Hospital (where patients are less sick overall) is smaller than that of Academic Hospital. This is partly because of Community Hospital's patients' characteristics (shorter LOS and higher turnover), and partly because of its lower patient volume, which necessitates its hospitalists to have a mixed medical-surgical case-mix. Although these insights are from a case study of two specific hospitals, many hospitals in the United States have similar characteristics to one of these hospitals and hence can benefit from understanding the results.

Acknowledgments

The authors are grateful to the anonymous associate editor and reviewers for their helpful and constructive feedback, which greatly improved the paper.

Endnotes

[1] Our definition of *case-mix* should not be confused with *case-mix index*, which is a clinical term assigned to each diagnosis-related group to reflect the allocation of resources for treatment of patients in that group.

[2] In Appendix C.1 in the e-companion, we use a sample of detailed time-stamp data for hospitalist orders from a partner hospital and show that the time between consecutive hospitalist orders are exponentially distributed. This supports that our results in this section are not an artifact of adding a random noise to the procedure dates.

References

Bavafa H, Savin S, Terwiesch C (2013) Redesigning primary care delivery: Customized office revisit intervals and e-visits. Preprint, submitted December 6, http://dx.doi.org/10.2139/ssrn.2363685.

Campello F, Ingolfsson A, Shumsky RA (2016) Queueing models of case managers. *Management Sci.* 63(3):882–900.

Chan CW, Dong J, Green LV (2016) Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Oper. Res.* 65(2):469–495.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* 62(2): 462–482.

De Véricourt F, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

Dobson G, Tezcan T, Tilson V (2013) Optimal workflow decisions for investigators in systems with interruptions. *Management Sci.* 59(5):1125–1141.

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

Elliott DJ, Young RS, Brice J, Aguiar R, Kolm P (2014) Effect of hospitalist workload on the quality and efficiency of care. *JAMA Internal Medicine* 174(5):786–793.

Faddy M, Graves N, Pettitt A (2009) Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value Health* 12(2):309–314.

Fulton BR, Drevs KE, Ayala LJ, Malott DL Jr (2011) Patient satisfaction with hospitalists: Facility-level analyses. *Amer. J. Medical Quality* 26(2):95–102.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Glasheen JJ, Misky GJ, Reid MB, Harrison RA, Sharpe B, Auerbach A (2011) Career satisfaction and burnout in academic hospital medicine. *Arch. Internal Medicine* 171(8):782–790.

Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. *Oper. Res.* 56(6):1526–1538.

Green LV, Savin S, Savva N (2013) "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Sci.* 59(10):2237–2256.

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: Incorporating workload heterogeneity into the newsvendor problem. *Manufacturing Service Oper. Management* 14(1):99–114.

Howrey BT, Kuo YF, Goodwin JS (2011) Association of care by hospitalists on discharge destination and 30-day outcomes after acute ischemic stroke. *Medical Care* 49(8):701–707.

Kaiser Family Foundation (2016) Hospital adjusted expenses per inpatient day. Accessed February 14, 2019, http://kff.org/other/state-indicator/expenses-per-inpatient-day/.

Kara A, Flanagan ME, Gruber R, Lane KA, Bo N, Kroenke K, Weiner M (2019) A time motion study evaluating the impact of geographic cohorting of hospitalists. *J. Hospital Medicine* 14:E1–E7.

Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW (2007) Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care. *J. Amer. Medical Assoc.* 297(8):831–841.

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

Kuo YF, Goodwin JS (2010) Effect of hospitalists on length of stay in the medicare population: Variation according to hospital and patient characteristics. *J. Amer. Geriatrics Soc.* 58(9):1649–1657.

Lindenauer PK, Chehabeddine R, Pekow P, Fitzgerald J, Benjamin EM (2002) Quality of care for patients hospitalized with heart failure: Assessing the impact of hospitalists. *Arch. Internal Medicine* 162(11):1251–1256.

Lindenauer PK, Rothberg MB, Pekow PS, Kenwood C, Benjamin EM, Auerbach AD (2007) Outcomes of care by hospitalists, general internists, and family physicians. *New England J. Medicine* 357(25):2589–2600.

Liu N, Ziya S (2014) Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production Oper. Management* 23(12):2209–2223.

Luo J, Zhang J (2013) Staffing and control of instant messaging contact centers. *Oper. Res.* 61(2):328–343.

Lurie JD, Wachter RM (1999) Hospitalist staffing requirements. *Effective Clinical Practice* 2(3):126–130.

Marshall AH, McClean SI (2004) Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Sci.* 7(4):285–289.

Meltzer D (2001) Hospitalists and the doctor-patient relationship. *J. Legal Stud.* 30(S2):589–606.

Michtalik HJ, Yeh HC, Pronovost PJ, Brotman DJ (2013a) Impact of attending physician workload on patient care: A survey of hospitalists. *JAMA Internal Medicine* 173(5):375–377.

Michtalik HJ, Pronovost PJ, Marsteller JA, Spetz J, Brotman DJ (2013b) Developing a model for attending physician workload and outcomes. *JAMA Internal Medicine* 173(11):1026–1028.

Molinari C, Short R (2001) Effects of an HMO hospitalist program on inpatient utilization. *Amer. J. Management Care* 7(11): 1051–1060.

Ozen A, Balasubramanian H (2013) The impact of case mix on timely access to appointments in a primary care group practice. *Health Care Management Sci.* 16(2):101–118.

Palmer HC, Armistead NS, Elnicki DM, Halperin AK, Ogershok PR, Manivannan S, Hobbs GR, Evans K (2001) The effect of a hospitalist service with nurse discharge planner on patient care in an academic teaching hospital. *Amer. J. Medicine* 111(8):627–632.

Payne K, Marshall AH, Cairns KJ (2011) Investigating the efficiency of fitting Coxian phase-type distributions to healthcare data. *IMA J. Management Math.* 23(2):133–145.

Phy MP, Vanness DJ, Melton LJ, Long KH, Schleck CD, Larson DR, Huddleston PM, Huddleston JM (2005) Effects of a hospitalist model on elderly patients with hip fracture. *Arch. Internal Medicine* 165(7):796–801.

Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G (2012) The relationship between inpatient discharge timing and emergency department boarding. *J. Emerging Medicine* 42(2):186–196.

Pressel DM, Rapport DI, Watson N (2008) Nurses' assessment of pediatric physicians: Are hospitalists different? *J. Healthcare Management* 53(1):14–24.

Rachoin JS, Skaf J, Cerceo E, Fitzpatrick E, Milcarek B, Kupersmith E, Scheurer DB (2012) The impact of hospitalists on length of stay and costs: systematic review and meta-analysis. *Amer. J. Management Care* 18(1):e23–e30.

Roy A, Heckman MG, Roy V (2006) Associations between the hospitalist model of care and quality-of-care-related outcomes in patients undergoing hip fracture surgery. *Mayo Clinic Proc.* 81(1):28–31.

Roytman MM, Thomas SM, Jiang CS (2008) Comparison of practice patterns of hospitalists and community physicians in the care of patients with congestive heart failure. *J. Hospital Medicine* 3(1):35–41.

Shi P, Chou MC, Dai J, Ding D, Sim J (2015) Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Sci.* 62(1):1–28.

Southern WN, Berger MA, Bellin EY, Hailpern SM, Arnsten JH (2007) Hospitalist care and length of stay in patients requiring complex discharge planning and close clinical monitoring. *Arch. Internal Medicine* 167(17):1869–1874.

Sox HC (1999) The hospitalist model: Perspectives of the patient, the internist, and internal medicine. *Ann. Internal Medicine* 130(4, part 2):368–372.

Sprivulis PC, Da Silva JA, Jacobs IG, Jelinek GA, Frazer AR (2006) The association between hospital overcrowding and mortality among patients admitted via western Australian emergency departments. *Medical J. Australia* 184(5):208–212.

State Inpatient Database (2015) Healthcare Cost and Utilization Project, 2005–2009. Agency for Healthcare Research and Quality, Rockville, MD, www.hcup-us.ahrq.gov/sidoverview.jsp.

Stevens JP, Nyweide DJ, Maresh S, Hatfield LA, Howell MD, Landon BE (2017) Comparison of hospital resource use and outcomes among hospitalists, primary care physicians, and other generalists. *JAMA Internal Medicine* 177(12):1781–1787.

Taheri PA, Butz DA, Greenfield LJ (2000) Length of stay has minimal impact on the cost of hospital admission. *J. Amer. College Surgeons* 191(2):123–130.

Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* 60(6):1574–1593.

Tipping MD, Forth VE, O’leary KJ, Malkenson DM, Magill DB, Englert K, Williams MV (2010) Where did the day go? A time-motion study of hospitalists. *J. Hospital Medicine* 5(6):323–328.

Today’s Hospitalist (2018) Hospitalist career and compensation survey. Accessed February 7, 2019, https://www.todayshospitalist.com/salary-survey-results.

Trzeciak S, Rivers EP (2003) Emergency department overcrowding in the United States: An emerging threat to patient safety and public health. *Emerging Medicine J.* 20(5):402–405.

Wachter RM (2014) Hospitalist workload: The search for the magic number. *JAMA Internal Medicine* 174(5):794–795.

Wachter RM, Goldman L (1996) The emerging role of “hospitalists” in the American healthcare system. *New England J. Medicine* 335:514–517.

Wachter RM, Goldman L (2016) Zero to 50,000—The 20th anniversary of the hospitalist. *New England J. Medicine* 375(11): 1009–1011.

Wang WY, Gupta D (2014) Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing Service Oper. Management* 16(3):439–454.

White HL, Glazier RH (2011) Do hospitalist physicians improve the quality of inpatient care delivery? A systematic review of process, efficiency and outcome measures. *BMC Medicine* 9:58.

Wright PD, Bretthauer KM, Côté MJ (2006) Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages. *Decision Sci.* 37(1):39–70.

Yankovic N, Green LV (2011) Identifying good nursing levels: A queuing approach. *Oper. Res.* 59(4):942–955.

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Oper. Management* 16(2):283–299.

Yousefi V, Chong CA (2013) Does implementation of a hospitalist program in a Canadian community hospital improve measures of quality of care and utilization? An observational comparative analysis of hospitalists vs. traditional care providers. *BMC Health Services Res.* 13(1):204.

Zacharias C, Armony M (2017) Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.* 63(11):3978–3997.