MANAGEMENT SCIENCE
Vol. 55, No. 9, September 2009, pp. 1486–1498
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5509 | 1486

DOI 10.1287/mnsc.1090.1037

Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations

Diwas S. Kc
Goizueta Business School, Emory University, Atlanta, Georgia 30322, diwas_kc@bus.emory.edu

Christian Terwiesch
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, terwiesch@wharton.upenn.edu

Much of prior work in the area of service operations management has assumed service rates to be exogenous to the level of load on the system. Using operational data from patient transport services and cardiothoracic surgery—two vastly different health-care delivery services—we show that the processing speed of service workers is influenced by the system load. We find that workers accelerate the service rate as load increases. In particular, a 10% increase in load reduces length of stay by two days for cardiothoracic surgery patients, whereas a 20% increase in the load for patient transporters reduces the transport time by 30 seconds. Moreover, we show that such acceleration may not be sustainable. Long periods of increased load (overwork) have the effect of decreasing the service rate. In cardiothoracic surgery, an increase in overwork by 1% increases length of stay by six hours. Consistent with prior studies in the medical literature, we also find that overwork is associated with a reduction in quality of care in cardiothoracic surgery—an increase in overwork by 10% is associated with an increase in likelihood of mortality by 2%. We also find that load is associated with an early discharge of patients, which is in turn correlated with a small increase in mortality rate.

Key words: health-care operations; productivity; service operations; optimal control of queues; quality
History: Received April 24, 2007; accepted April 3, 2009, by Linda V. Green, public sector applications. Published online in *Articles in Advance* July 6, 2009.

1. Introduction

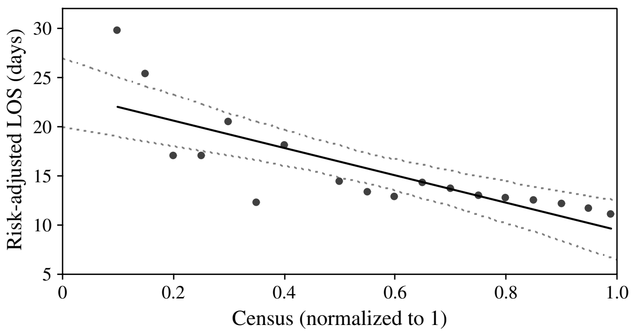
Over the last decade, most hospitals have witnessed a substantial increase in fixed costs, largely reflecting growing expenses for new technologies and liability insurance. Over the same period, hospitals also had to face a substantial decrease in per-case reimbursements, reflecting the transition from fee for service reimbursements to contractual reimbursements due to managed care. As a result of these two trends, hospitals have come under increasing pressure to operate at very high levels of utilization. From a macro perspective, high utilization is a desirable system property for a hospital and its employees, as it spreads the fixed cost over a larger volume of patients. However, recent research conducted with a more micro perspective (Green 2004) has demonstrated that operating at high levels of utilization has many operational implications, including long waiting times.

Most of these micro level models are based on queueing analysis (Green 2004, Smith-Daniels et al. 1988). Such models analyze patient flows and, in particular, patient waiting times based on information about the care capacity of the process, the variability of its service times, and the behavior of a stochastic demand for care. A high level of utilization (a high level of demand relative to the available capacity) leads to a dramatic increase in wait times and—if waiting is not feasible due to the emergency of the case or due to a limited amount of space—a reduction in patient flow (i.e., the number of patients cared for in a unit of time). Collectively, queueing analysis in health care has emerged as an active area of research with a clear potential for impacting health-care practice.

A central assumption in this existing literature is that the service time, i.e., the time it takes a resource to care for a patient, is independent of the state of the process including the current workload. In this paper we show this might not always be the case. Consider the data shown in Figure 1. As a motivating preview to one of our results, the figure shows the relationship between the risk-adjusted[1] length of stay of cardiothoracic surgery patients as a function of the

[1] In the medical literature, the risk-adjusted length of stay is computed by first determining how individual patient risk factors predict the length of stay, and then generating an expected length of stay for each patient based on their specific risk factors.

Figure 1 Length of Stay as a Function of Census



Notes. Census is defined as the number of patients in the cardiac unit at the time that a patient is admitted. Length of stay (LOS) is the total number of days a patient spends at the hospital. Dashed lines represent 95% confidence intervals.

workload in the cardiothoracic surgery unit[2] at the time of discharge. We observe a clear pattern indicating that the service time (duration the patient is in the unit) decreases with an increase in workload. The unit thus increases its throughput when it is busy. In other words, its level of care capacity seems to be adaptive to higher levels of workload. From an empirical perspective as well as from the perspective of hospital management, the data shown in Figure 1 raises a set of interesting research questions. (1) What drives this increase in processing speed? Is the hospital simply discharging patients prematurely, or is there evidence that the same work gets done faster? (2) Are there any implications for the quality of care provided? (3) Can the resources in the hospital sustain this increased service rate or does there exist an effect of overwork?

We address these three questions by conducting a detailed econometric analysis of two care processes in a major U.S. teaching hospital. In particular, we look at process and outcome data of some 3,000 cardiothoracic surgery patients. We measure the length of stay for each patient and relate it to a set of covariates, including current workload and the cumulative fatigue, or workload burden on service workers. We address the alternative explanation of Figure 1 that the hospital simply discharges patients prematurely in two ways. First, we look at risk-adjusted mortality data to investigate how workload and overwork lead to changes in mortality. Second, we also study another care process in the hospital that is not a medical process and does not provide the option of simply cutting the service time short at the potential cost of quality. In particular, we look at the service times of over 17,000 requests for patient transport and analyze how they change with workload and the subsequent overwork.

This research design allows us to make the following three contributions. First, we measure the performance of hospital employees and show that employees adjust their service rates with changing levels of load. This is, to the best of our knowledge, the first empirical test of the insights obtained from the optimal queueing control literature. In cardiothoracic surgery, we find that a 10% increase in load leads to a reduced length of stay (service time) of over two days (about 20%). Similarly, we find that patient transporters speed up their tasks by 30 seconds (about 2.4% of service time) if load increases by 20%. Second, our study investigates the impact of workload as well as overwork on the quality of care, a relationship that is potentially a matter of life or death in a hospital. *Overwork* is defined as the excess workload beyond an expected amount of workload over a given period of time. Specifically, we establish that patients admitted to an overworked unit are associated with an increased risk of mortality. On average, a 10% increase in overwork is associated with a 2% increase in risk of mortality. Third, we show that although hospital employees can respond to increased workload by increasing their productivity in the short run, such an acceleration in general is not sustainable. After a duration of exceptionally high workload, employees are subject to the aftereffects of overwork. This effect of overwork could outweigh the higher service rates discussed above. A sustained level of 1% above average load for a week in cardiothoracic surgery units leads to an average increase in length of stay of almost six hours (2%).

If hospital employees are indeed capable of adjusting their service rate as a function of the workload, this clearly has substantial implications for the management of care capacity. If service workers can adapt during periods of high workload by working faster, it may not be necessary to hire additional capacity during busy periods. Also, instead of relying on safety capacity to buffer against stochastic increases in demand, the hospital could rely on its staff's ability to temporarily accelerate their work. However, our empirical findings suggest additional managerial considerations that need to be made. Although such adaptive behavior from workers may appear desirable in the short run, one needs to also consider the quality and patient safety implications of such behavior. In addition, temporary worker speedup made come at the cost of future slowdown after the onset of fatigue. This could lead to a net total decline in performance. Decision makers should thus take into consideration the full set of possible implications of a temporary increase in service rates.

[2] The cardiothoracic surgery unit is the self-contained hospital unit that includes (i) admissions; (ii) diagnostic testing (catheterization lab, electrocardiogram, etc.); (iii) preoperative care, such as prepping the patient for surgery; (iv) surgery; (v) postoperative care (e.g., time in the intensive care unit); and (vi) discharge.

The remainder of this paper is organized as follows. In §§2 and 3, we present relevant literature and develop our hypotheses for a general model of service operations, respectively. We then operationalize our theory to the two hospital settings we study. Section 4 describes our research setting, the econometric model specification, and the results for our study of patient transporters. In §5, we report the same information for the cardiac surgery setting. We conclude with discussions and future avenues for research in §6.

2. Literature Review

The operations research literature has created a number of tools that directly or indirectly relate to the management of care capacity and its utilization (see Green 2004 for an overview). At the strategic level, decisions need to be made with respect to sizing the care capacity. This includes choosing occupancy rates (e.g., Smith-Daniels et al. 1988, Huang 1995, Green and Nguyen 2001) and making staffing decisions (e.g., Aiken et al. 2002, Kwak and Lee 1997, Green and Meissner 2002). At the tactical level, decisions need to be made with respect to scheduling and sequencing cases (e.g., Gerchak et al. 1996) as well as with respect to allocating capacity to various demand types (e.g., Green et al. 2006). Much of this prior body of literature, however, assumes that the service rate is exogenous to the level of capacity utilization. In this paper, we present and validate a framework of service operations where workers vary their service rates with the state of the system. There also exists a significant body of literature dealing with optimal payment systems for health services, as reported by Newhouse (1996). Many of these studies (e.g., Fuloria and Zenios 2001) explore the effect of various types of payment arrangements that incentivize health-care organizations into providing higher quality of services. Higher quality is often achieved only at a higher cost, of which workload and service rates are important contributors. This stream of literature seeks to examine how, in the presence of unobserved cost factors, appropriate incentives can still be provided to hospitals to induce higher quality. In addition to this general research on hospital operations, our analysis builds on two areas of prior research in operations management.

First we draw on the literature on the optimal control of queues. For example, Crabill (1972) and Bertsekas (2000) examine systems in which the service rate is adjusted dynamically as the queue length changes.[3] Some of these models study the dynamic control of a single-server queueing system that has Poisson arrivals and exponentially distributed service times. There are costs associated with an increase in the queue length and in an increase in the service rate. The objective is to choose the optimal service rate that minimizes the average sum of these two costs over a given planning horizon. In other words, a key objective of this body of literature is the development of service rate policies that effectively balance the costs of waiting with the costs of an accelerated service rate. Under relatively general assumptions, Stidham and Weber (1989) prove the existence of a stationary policy, i.e., one in which transition to a given state elicits the same service rate. Although closed form solutions for the optimal service rate as a function of queue length are not obtainable, George and Harrison (2001) develop a novel method for computing the optimal policy for the service time as a function of the queue length, subject to certain restrictions on the two cost functions. In such a setting, the optimal service rate is a nondecreasing function of the length of the queue. The intuition for the monotone policy is that working faster by a given unit rate has a bigger impact on total waiting cost when the queue is longer. In a similar vein, Berk and Moinzadeh (1998) also allow the service rate to vary, and normatively explore the impact of the option of a shorter service time on effective capacity. Even though the results in this body of literature are well established, there have been no empirical validations of this effect. We contribute to this line of research by providing explicit evidence of the adaptive behavior in two health-care services. For both services, although the underlying waiting costs and service rate costs are not estimated, we show that service rates increase when the load on the system increases.

Our work also extends prior studies of the impact of production system design on the productivity of employees. For example, using lab-based experiments, Schultz et al. (1998, 1999) consider serial production systems in which adjacent workers in a serial assembly line can observe each others' productivity, as measured by inventory levels between them. A key insight from this work is that workers tend to work faster or slower depending on the work in process inventory. Our objective in this paper is to demonstrate using actual operational data from a field based study at a hospital, that health-care delivery workers also demonostrate such adaptive behavior in response to the amount of workload. In addition, the previous studies have not considered the aspects of fatigue that accompany service rate acceleration, or the impact on the quality of service. Our study augments the existing body of work to include the dimensions of fatigue and quality. Powell and Schultz (2004) show that when assembly line workers adapt to variations

[3] Although previous work (e.g., Green 1984) models a queueing system that involves multiple servers, as far as we are aware, there are no established optimal policies on service rate when multiple servers are involved.

in load, they also improve the overall throughput of the system. One of the implications of our study is that the adaptive behavior of health-care providers increases the overall process flow of patients from the hospital.

3. Hypothesis Development

Our theoretical framework is based on the relationships between service times, workload, overwork, and service quality. All of these measures are defined for the discrete unit of work, denoted *i*. We define load ($LOAD_i = REQUESTS_i/RESOURCES_i$) as the total number of requests or jobs ($REQUESTS_i$) in the system divided by the total number of resources ($RESOURCES_i$) at that time that unit of work *i* is in the system. In other words, $LOAD_i$ provides a measure of the level of utilization of the system's resources that is connected with the unit of work *i*.

We define $SVCTIME_i$ to be the service time taken to process a request *i*. This definition of service time does not include any time spent waiting for the service to begin. Our hypothesis is that a higher workload leads to a reduction in service time; i.e.,

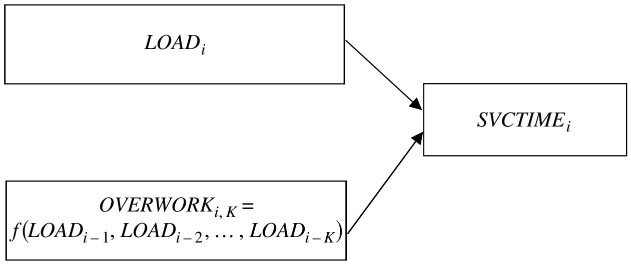
$$\frac{\partial SVCTIME}{\partial LOAD} < 0. \qquad (1)$$

Such a behavior can be rational from the worker's perspective if each service worker's utility is decreasing in the level of waiting time at a greater rate than the decrease in the utility associated with effort involved in obtaining a faster service rate, as theoretically established in the literature on the optimal control of queues.

Although productivity gains may be achieved in the short term as we hypothesize, high service rates may not be sustainable for longer periods of time. During periods of increased load, a worker may be motivated to work fast, but eventually fatigue effects may start to dominate, leading to increased service times. Early research in the field of ergonomics (Cakir et al. 1980) has shown that as fatigue rises, productivity falls. Tanabe and Nishihara (2004) use lab experiments to study changes in productivity and find that even though people are highly motivated in short term experiments, they become tired and performance deteriorates over a longer time frame as fatigue kicks in. Likewise, a key finding in the studies by Caldwell (2001) and Setyawati (1995) is that fatigued workers exhibit diminished productivity. Figure 2 summarizes these hypotheses.

To study the phenomenon above, we construct the measure $OVERWORK_{i,K}$, which we define to be an increasing function in the difference between the observed $LOAD_i$ and the average over *K* units of time prior to the arrival of unit of work *i* in the system. In other words, when a unit of work *i* arrives at the system after a period of sustained levels of high *LOAD* for *K* units of time, our measure of $OVERWORK_{i,K}$ will be high. We argue that this holds for a broad set of values of *K* used to estimate *OVERWORK*. Based on the discussions above, we propose that the service time is increasing in the overwork; that is,

Figure 2 Effect of Load and Overwork on Service Time


$LOAD_i$
$SVCTIME_i$
$OVERWORK_{i,K} = f(LOAD_{i-1}, LOAD_{i-2}, \ldots, LOAD_{i-K})$

$$\frac{\partial SVCTIME}{\partial OVERWORK} > 0. \qquad (2)$$

We next consider the impact of the above effects of load, overwork, and service time on the quality of service (*QUALITY*), which is of paramount importance in health-care delivery. During periods of high *LOAD*, resources are more thinly spread out. We hypothesize that this decrease in the availability of resources can lead to a decline in quality, that is,

$$\frac{\partial QUALITY}{\partial LOAD} < 0. \qquad (3)$$

Similarly, we argue that a patient who is admitted to an overworked unit has a higher likelihood of encountering a quality lapse, as service workers who are more fatigued are more prone to making mistakes; that is,

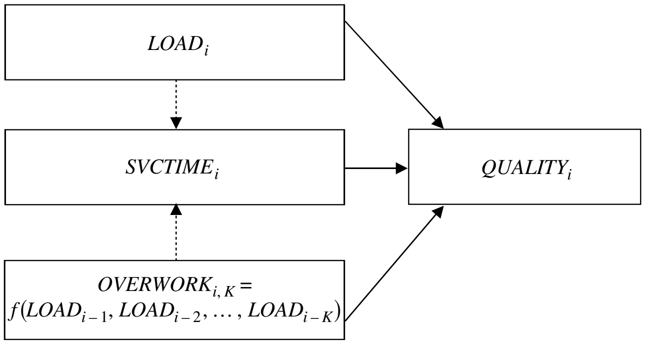
$$\frac{\partial QUALITY}{\partial OVERWORK} < 0. \qquad (4)$$

Finally, we hypothesize that when service times are decreased (after controlling for patient specific factors), and patients are discharged early, this could have an adverse impact on the quality of care; that is,

$$\frac{\partial QUALITY}{\partial SVCTIME} > 0. \qquad (5)$$

To test the hypotheses illustrated by Figure 3, we chose two vastly different kinds of services—patient transportation and cardiothoracic surgery—at a major U.S. teaching hospital. Patient transportation is, relative to other health-care tasks, simple, and the task of moving a patient from one part of the hospital to another is rather mechanical in nature. Typically each transport lasts less than half an hour. In sharp contrast, service workers in cardiothoracic surgery require advanced medical knowledge and

Figure 3 Effect of Load, Overwork, and Service Time on Quality



extensive training. The individual tasks in cardiothoracic surgery are more complicated, and the average patient length of stay is around two weeks.

A study looking at patient transport alone might be dismissed as not being applicable to more medical and diagnostic processes. A study looking at cardiac care alone might be dismissed with the claim that patients are simply discharged prematurely as opposed to receiving care at a faster service rate. Replicating our research design across these two different care processes hence increases the generalizability of our findings. We provide context-specific justifications for the hypotheses outlined, followed by our findings for the two studies.

4. Patient Transport Study

Patient transporters are hospital employees who perform the crucial role of taking a patient from one part of the hospital to another. The hospital that we study maintains a pool of between 2 and 26 transporters, depending on the time of day. When a patient is ready for transport, the nurse in charge of the hand-over submits an electronic request. The request then is placed in a queue to be processed by a dispatcher. When a transporter is available, the dispatcher assigns a transporter to a specific request. After a transporter arrives at the transport location, the transport process begins.

We operationalize the variables defined in the previous section as follows. $REQUESTS_i$ is the total number of transport requests, and $RESOURCES_i$ is the number of transporters working on the shift at the time that request i arrives. $LOAD_i$ is the fraction of transporters who were busy during the hour that transport i was started. So if 5 out of 10 transporters were occupied at the time that service i was rendered, $LOAD_i = 50\%$. Note that our definition of $LOAD_i$ corrects for anticipated increases in demand that were addressed by an increase in scheduled capacity. For example, the hours between 9 a.m. and 10 a.m. on a regular weekday, show three times more transport requests than there are between 9 p.m. and 10 p.m. However there are also two and one-half times more transporters staffed during the busier period. The $SVCTIME_i$ for each transport i is the time between the patient leaving the starting location and arriving at the final destination. This does not include any waiting time for the transporter to arrive.

Table 1 Transport Descriptive Statistics

Measure	Mean	Standard deviation	Median
SVCTIME (minutes)	12.6	7.75	10.35
LOAD	0.755	0.21	0.73
$OVERWORK_{K=4}$	0.001	0.21	0.02

We define $OVERWORK_{i,K}$ at the level of the transporter, and the measure for $OVERWORK_{i,K}$ is computed only if the transporter performing service i was on shift for each of the K periods prior to the start of service i. Let $t(i)$ be the time at which unit i arrives. To formalize the notion of overwork, we define $OVERWORK_{i,K}$, from time $t(i) - K$ up to time $t(i)$ as

$$OVERWORK_{i,K} = \frac{1}{N(K,i)} \sum_{j=i-N(K,i)}^{i-1} (LOAD_j - \overline{LOAD_{s(j)}}),$$

where $\overline{LOAD_{s(j)}}$ is the average load over the entire shift s, and $N(K, i)$ is the number of service requests during the last K periods up to $t(i)$. The K periods are measured in units of hours. For example, suppose that the expected load during a certain shift is four requests per worker every hour. However, suppose that for a particular hour proceeding request i ($K = 1$), the load has consistently remained at six requests per worker during which six requests happened to have been processed. The overwork, $OVERWORK_{i,1}$ associated with request i would then equal $(1/6)\sum_6(6-4) = 2$ requests per worker. In other words, the worker responsible for transporting request i has already experienced an additional load of two requests on average over this time period.

An average transport lasts 12.6 minutes, and the average load on transporters is 0.76. Table 1 provides descriptive statistics of the key variables of concern. To achieve parallelism with the cardiothoracic surgery study, we sought out possible measures of quality in patient transport. In speaking to the head of patient transport services, we found that one source of error involves the patient being transferred to the wrong location. The other potential error is a lapse in adherence to specific protocols (for example, with handling of equipment and supplies). However, these errors are not captured and collecting this data is not currently feasible. Thus, although desirable, the quality implications of speedup are not estimated.

4.1. Econometric Analysis

The variable *SVCTIME* does not take on negative values. Thus, we follow the commonly used approach of taking the natural logarithm of the variable to reduce the skewness in distribution. We specify our regression model as

$$\log(SVCTIME_i) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta}_1 + \beta_2 \log(LOAD_i) + \beta_3 OVERWORK_{i,K} + \varepsilon_i, \quad (6)$$

where ε_i is the mean zero error term. $\mathbf{X}_i$ consists of a set of variables that control for the underlying heterogeneity in patient characteristics or task characteristics, or both.[4] This includes indicators for the time of the day (*TIME*) and day of the week (*DAY*), which capture intertemporal differences in elevator availability and hallway traffic as well as specific information about the transport. Transporters may be required to use additional pieces of equipment (*EQUIP*) along the way, including intravenous medication, oxygen, and other supplies. Transports also vary in mode (*MODE*); some patients may require specialized telemetry beds, whereas others only need wheelchairs and transport beds. For example, a patient transport with a telemetry bed will take longer than with a wheel chair. For each transport *i*, we also correct for the person in charge of the transport (*NAME*), trip start (*START*), and end (*END*) locations, starting and ending locations for the transporter (*PATH*), and type of patient transported (*TRIP_TYPE*). Table A1 in the online appendix (provided in the e-companion)[5] provides a list of variables and controls ($\mathbf{X}_i$) for the econometric specification above.

As the load on the system increases in any given shift, the expected waiting times for transporters also tend to increase. Speeding up the transport time helps to somewhat mitigate the increase in waiting times. Thus, transporters (whose performance is constantly evaluated through a patient tracking system) have an incentive to speed up when the load on the system increases as outlined in (1). The coefficient of β_2 denotes the elasticity of service time with respect to load. A value of $\beta_2 < 0$ indicates that servers respond to high load by reducing the service time, providing support for (1).

To capture a potentially nonlinear relationship between *LOAD* and *SVCTIME*, we also created a categorical variable for *LOAD* for values in the ranges 0–0.3, 0.3–0.5, 0.5–0.65, 0.65–0.8, and 0.8–1 such that we had approximately similar numbers of observations within each range. We then estimated (6), replacing $\log(LOAD_i)$ with the categorical specification for $LOAD_i$.

Table 2 Effect of Load and Overwork on Transport Time

Coefficient	Model 1	Model 2
Intercept	2.73	2.13
	(0.74)***	(0.09)***
β_2	−0.17	−0.12
	(0.07)***	(0.07)*
β_3		0.09
		(0.05)**
R^2	0.62	0.62
F-statistic	4.5 ($p < 0.01$)	4.5 ($p < 0.01$)

Note. Dummy variables for X_i (provided in the online appendix) are not displayed.

***, **, and * denote statistical significance at the 1%, 5%, and 10% confidence levels, respectively. Standard errors are shown in parentheses.

Finally, as outlined in hypothesis (2), we expect *OVERWORK* to be negatively correlated with the transport time (*SVCTIME*). Patient transport is a physically demanding task, and after a few hours of transporting patients, transporters may exhibit symptoms of tiredness and fatigue. Thus, a positive value of β_3 suggests that overwork leads to a longer service time, providing support for the hypothesis outlined in (2).

4.2. Results

Table 2 summarizes the results of estimating the above regression model with service time as a dependent variable based on a sample of 17,000 patient transports. We find that the elasticity of load on service time is −0.12 (model 2). This amounts to approximately 2.4% faster service on average for a 20% increase in *LOAD*. This result provides support for our hypothesis that higher load leads to shorter service times.

Next, consider the effect of overwork. In estimating (6) above, we find that $K = 4$ yields the best model fit.[6] The regression results in Table 2 show that the coefficient for *OVERWORK* (β_3) has a value of 0.09 (p-value $= 0.05$). That, is a 0.1 unit increase in *OVERWORK* (or the equivalent of a sustained level of 0.1 additional load above the expected load for $K = 4$ hours) leads to an increase in service time by about 0.9%. This lends support to hypothesis (2) that overwork leads to an increase in the service time in patient transport. Our result is consistent with our interviews with patient transporters and their management who reported, based on their personal experience, that transporters visibly slow down at the end of busier shifts. At any given point in time, a worker is subject to the effects of both existing load, and fatigue effects arising from sustained load in the immediate past. We find that the correlation between

[4] The logarithmic transformation of *LOAD* captures the nonlinearity in the regression function and provides a better model fit, as demonstrated by the distribution of the residuals. *OVERWORK* by construction is not strictly positive.

[5] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

[6] Our estimations were performed with varying values for *K*. The final value of *K* that was chosen yields the best maximum likelihood value.

LOAD and *OVERWORK* is 0.295. Load and overwork have opposing effects, so at any given point in time, depending on the relative magnitudes of load and overwork, the net effect might be either a decrease or an increase in the service rate.

5. Cardiothoracic Surgery Study

Unlike patient transport, cardiothoracic surgery is a highly specialized service involving numerous care providers. In our analysis, we observe the lengths of stay and quality measure of patients who pass through a single cardiothoracic surgery unit. We observe the admission and discharge dates for each patient, which are used to compute the patient length of stay as well as the daily census. The index i identifies each unique patient admission. The associated service time $SVCTIME_i$ is the total length of stay for the patient from the date of admission to the discharge date. $REQUESTS_i$ measures the number of patients in the unit when patient i was admitted (the census) and $RESOURCES_i$ is the total bed capacity when patient i arrives at the cardiothoracic surgery unit. In our period of study, the total bed capacity remained unchanged. $LOAD_i$ is thus defined to be the census divided by the total bed capacity at the time that patient i is in the hospital. In our preliminary analysis (Figure 1), we looked at the effect of $LOAD_i$ at the time of admission. We also computed alternative measures of $LOAD_i$, including a measurement at the time of discharge, and at the midpoint of the patient's stay in the hospital. In addition, we also computed $LOAD_i$ over a nominal fixed length of stay for all patients.[7] We find that all four measures of $LOAD_i$ have a very similar effect on service time (Table A4 in the online appendix). For the remainder of this study, we compute $LOAD_i$ by using the daily average of load measured over the entire length of stay of patient i.

In contrast to our transport study, in the cardiac surgery study there exists no unique individual worker who performs all tasks related to a particular patient. Therefore, we estimate $OVERWORK_{i,K}$ at the level of the hospital unit using the daily load for K days prior to the admission day for patient i. Let $d(i)$ be the date on which patient i is admitted. We define $OVERWORK_{i,K}$, from time $d(i)-K$ up to time $d(i)$ as

$$OVERWORK_{i,K} = \frac{1}{N(K,i)} \sum_{j=i-N(K,i)}^{i-1} LOAD_j - \overline{DAILY\_LOAD_{d(j)}},$$

where $\overline{DAILY\_LOAD_{d(j)}}$ is the average load in the unit on the day of admission of patient j, and $N(K,i)$ is the number of patient arrivals during the last K periods up to $t(i)$. For example, a large positive value of $OVERWORK_{i,K}$ signifies that the unit has experienced high levels of load over the K days of observation prior to the admission of patient i.

Table 3 Cardiothoracic Surgery Descriptive Statistics

Measure	Mean	Standard deviation	Median
SVCTIME (days)	12.98	10.69	7
LOAD	0.78	0.086	0.79
$OVERWORK_{K=7\text{ days}}$	0.005	0.07	0.01
MORTALITY	0.068	0.255	0

As indicated in the descriptive statistics (Table 3), we see that the average length of stay for a patient undergoing cardiothoracic surgery is 12.98 days. The standard deviation of 10.7 days indicates significant variability in length of stay, which is partly due to the heterogeneity amongst patients. The average load of 0.78 is comparable to the average load seen by transporters.

5.1. Econometric Analysis

We test hypotheses (1) and (2) using the econometric specification:

$$\begin{aligned}\log(SVCTIME_i) = {} & \gamma_0 + \mathbf{Y}_i\boldsymbol{\gamma}_1 + \gamma_2 \log(LOAD_i) \\ & + \gamma_3 OVERWORK_{i,K} \\ & + \gamma_4 MON\_WED + \varepsilon_i. \qquad (7)\end{aligned}$$

$\mathbf{Y}_i$ includes a set of variables that control for the underlying heterogeneity in patient characteristics, as well as temporal factors such as month of admission. The patient population includes cardiac patients that vary widely in length of stay and risk levels. To account for cardiothoracic surgery specific factors that influence the $SVCTIME_i$ and outcome, as measured by the occurrence of postsurgery mortality ($MORTALITY_i$), we include several clinical preoperative risk factors including age (AGE_i), sex (SEX_i), race ($RACE_i$), emergency status ($EMER_i$), and various specific medical comorbidities and complicating factors to correct for patient level heterogeneity. We use two commonly used medical estimates of patient-specific risk. The measure $EUROSCORE_i$ is estimated on a zero to one scale and captures the preoperative level of patient risk based on a number of individual patient risk factors. A similar risk score, developed by the New York Heart Association ($CLASS\_NYHA_i$), was also available for each individual patient. We also observe the type of procedure ($PROCEDURE_i$)[8] performed, as this has a significant bearing on the length

[7] We thank the review team for suggesting the various measures of *LOAD*. The online appendix includes our results for impact of the various measures of *LOAD* on the service times.

[8] We did not observe individual surgeons involved in the procedures. However, each cardiothoracic procedure is highly specialized and is performed by either one or two surgeons. For instance, mitral valve procedures are operated by only one surgeon. Thus, *PROCEDURE* also serves as a proxy for the surgeon.

of stay. For example, a patient will need a longer recovery time following a combined valve and bypass surgery compared to a single bypass surgery. We correct for temporal factors that could affect the length of stay (through staffing shortages, during holiday season, and on weekends, for example) by using indicator variables denoting month ($MONTH_i$) and day of week (MON_WED_i) of admissions. Finally, we also observe incidences of a patient having to be reintubated ($RE\_INTUBATED_i$). Reintubation[9] occurs if a patient is put on ventilator support for a second time. Tables A2 and A3 in the online appendix provide detailed definitions of all operational and medical variables.

Because hospitals often cite bed capacity as the primary reason for the inability to admit new patients, we believe that bed capacity utilization is a significant driver of admission and discharge decisions, and ultimately determines a patient length of stay. In other words, when the system is busy, beds are in greater demand. Consequently, there is pressure to discharge patients faster to increase bed capacity. The effect of an increase in load on reducing the length of stay thus makes hypothesis (1) appear tenable in the context of a cardiothoracic surgery unit. The coefficient of γ_2 denotes the elasticity of service time with respect to load. A value of $\gamma_2 < 0$ indicates that the unit responds to high load by reducing patient length of stay, providing support for the hypothesis outlined in Equation (1).

Prior research investigating the performance of health workers has investigated the effect of worker fatigue on clinical decision making and outcome. In particular, fatigued and overworked medical residents and nurses have been observed to create more medical errors in diagnosis and treatment (e.g., Scott et al. 2006). For example, Gaba and Howard (2002) point out that most studies on fatigue show impairment of clinically relevant tasks. We argue that fatigue could impact the length of stay in two ways—either because the decision maker would like to take more time to make the discharge decision,[10] or because fatigued workers are more prone to making medical errors. We hypothesize that such errors lead to complications that call for additional rework, which would further lengthen a patient's stay. Hypothesis (2) is supported if fatigue leads to an increase in the patient's length of stay. Recall that K is the duration of units of time over which high load brings about a noticeable amount of fatigue. The value of K that yields the best model fit for specification (7) is chosen as the period of time over which $OVERWORK_K$ is estimated. The coefficient γ_3 captures overwork effects. A positive value of γ_3 suggests that a sustained period of high load leads to a longer service time, providing support for the hypothesis outlined in Equation (2).

The prior medical literature relies on self-reported measures of fatigue. In this paper, we show that our objective, census-based measure of overwork, also increases the length of stay. This suggests that overwork could be used as a proxy for a measure of the level of fatigue, where self-reported values are unavailable or biased.

We also examine whether staffing levels could affect the length of stay of patients. Although we do not directly observe the daily staffing levels in our data set, we note that medical care providers, including nurses, anesthesiologists, and residents typically work regular weekly schedules. Consequently, any variations in the level of medial staff are "seasonal" on a weekly basis. That is, the staffing level changes can be controlled for by simply accounting for the day of the week. In our preliminary analysis, we find that the number of staff does not vary greatly during weekdays. However, staffing levels are slightly lower during weekends. We also find that the average length of stay is slightly less than two weeks. This means that a patient admitted on a weekend would have stayed, on average, two weekends in the hospital, whereas a patient admitted early in the week would most likely have spent only one weekend. Given that the weekday staffing levels are higher than weekend staffing levels, the patient who ends up spending two weekends experiences more days with fewer support staff. Thus, by explicitly controlling for the day of week of admission, we account for the weekly schedule-related changes in the level of staffing that could drive the observed length of stay effects. In the econometric specification above, $MON\_WED = 1$ if a patient was admitted on either a Monday, Tuesday, or Wednesday, and $MON\_WED = 0$ otherwise. γ_4 estimates the effect of a weekend or near-weekend admission on increasing the length of stay.

We next consider the effect of load and overwork on the quality of service. In health-care operations, medical outcome is commonly used as a measure of quality of service. Compared to patient transport, outcomes are much more important and also more accurately quantifiable in the case of cardiothoracic surgery. Our focus, with respect to quality, is to investigate if and to what extent process variables such as workload and

[9] Intubation is the placement of a flexible plastic tube into the trachea to protect the patient's airway and provide a means of mechanical ventilation. If a patient is intubated again (or reintubated), it is an indicator of increasing patient severity, and possibly longer length of stay.

[10] In discussions with medical staff, we noted that doctors are more likely to prescribe medical tests when discharge and diagnosis decisions become difficult.

overwork are significant covariates when predicting mortality.

In this setting, a large body of medical literature has statistically analyzed variables that influence the risk-adjusted mortality score (Nashef et al. 2002, Kurki 2002, EuroSCORE 2007). Following a long line of medical research in cardiac surgery, the EuroSCORE model is one such statistical model that attributes a mortality score to a set of patient level risk factors. Specifically, the EuroSCORE model takes a number of medical covariates, such as gender, age, medical conditions, and procedure specific attributes, such as the nature of the procedure, and links them to the binary outcome of mortality using a logit regression. That is, the EuroSCORE model is essentially a logistic regression model with the dependent binary variable as quality of care and the independent variables as the preoperative and procedure-specific risk factors. The online appendix lists the set of independent variables used by the EuroSCORE model. In our analysis we augment the EuroSCORE model to examine the effect of additional covariates such as load and overwork on the mortality rate.

We study two mechanisms in which process variables might affect mortality. First, workload and overwork might impact the risk of mortality during the hospitalization of the patient. For example, Needleman et al. (2002, 2006) found that a higher number of hours of care by registered nurses per patient is associated with better care. Aiken et al. (2002) report that higher patient to nurse ratios are linked with higher patient mortality and failure to rescue among surgical patients. Following this prior work, we argue that for intensive care patients such as those in a cardiothoracic unit, a decrease in the time that doctors and nurses have available on a per-patient basis leads to an increase in risk-adjusted mortality during the hospitalization. Define the binary variable $MORTALITY\_IH_i$ such that $MORTALITY\_IH_i = 1$ if the ith patient died during the hospitalization and $MORTALITY\_IH_i = 0$ otherwise.

To test hypotheses (3) and (4) using in-hospital mortality as a measure of quality, we augment the EuroSCORE model by including the variables *LOAD* and *OVERWORK* as additional covariates. We consider the effect of *LOAD* and *OVERWORK* on all postoperative, in-hospital mortalities. This leads to the following logistic regression model:

$$\begin{aligned}&\operatorname{logit}[\Pr(MORTALITY\_IH_i)]\\ &\quad = \mu_0 + \mathbf{Z}_i \boldsymbol{\mu}_1 + \mu_2 LOAD_i + \mu_3 OVERWORK_{i,K}, \quad (8)\end{aligned}$$

where μ_0 is the base-line rate of in-hospital mortalities. $\mathbf{Z}_i$ includes the 19 medical covariates that are used in the EuroSCORE model to predict patient mortality.[11] A positive value of μ_2 in (8) would provide support for the hypothesis outlined in (3), indicating that patients entering cardiac surgery at a time when the unit is highly utilized face a higher mortality risk. Likewise, a positive value of μ_3 would provide support for the hypothesis outlined in (4), indicating that patients entering cardiac surgery at a time when the resources have been exposed to an extended period of high workload (i.e., are overworked) face a higher mortality risk.

Second, process variables might also impact mortality after the hospitalization of the patient, i.e., the mortality of patients who have already been discharged. We use the postdischarge mortality as an additional measure of quality. Define the binary variable $MORTALITY\_PD_i$ with $MORTALITY\_PD_i = 1$ if the ith patient died within 30 days postdischarge and $MORTALITY\_PD_i = 0$ otherwise. Just as we hypothesized for the in-house mortalities, we aim to analyze if an increase in load or the cumulative effect of overwork leads to an increase in probability of postdischarge mortality. Unexpected complications might be overlooked by a busy or overworked workforce.

In addition to validating (3) and (4), there exists another effect of process variables on mortality that is unique to the postdischarge mortality. A high workload might induce the hospital to discharge patients early; this in turn might increase the odds of mortality. However, to examine the effect of early discharge on mortality rate, it is not enough to simply observe the relationship between mortality and length of stay. This is because a longer hospital stay could be associated with increased case severity and a higher likelihood of mortality. On the other hand, a shorter length of stay due to an earlier discharge could lead to a lower quality of care, resulting in an increased likelihood of mortality. Our objective is to identify this second effect. To do so, we need to separate the confounding effect of severity of illness on the length of stay.

We do this by first computing the predicted length of stay for case i, $\widehat{SVCTIME}_i$. Among cardiothoracic surgery patients, medical risk factors such as patient age, sex, various comorbidities, and procedure type are considered to be significant predictors of length of stay. We estimate this risk-based expected length of stay ($\widehat{SVCTIME}_i$) using such medical risk factors. We then compute the variable $EARLYDIS_i$ as the difference between the actual length of stay ($SVCTIME_i$) and the predicted length of stay ($\widehat{SVCTIME}_i$):

$$EARLYDIS_i = \widehat{SVCTIME}_i - SVCTIME_i.$$

[11] Service time is not included in this empirical specification because for in-hospital mortalities, the patient discharge decisions and hence length of stay are not explicit decision variables.

The variable $EARLYDIS_i$ then captures changes in the length of stay caused by nonmedical risk factors. In particular, we hypothesize that an increase in load leads to an early discharge. To establish an increase in load leads to an early discharge, we use the following econometric specification:

$$EARLYDIS_i = \kappa_0 + \mathbf{Y}_i\boldsymbol{\kappa}_1 + \kappa_2 \log(LOAD_i) + \tau_i, \qquad (9)$$

where τ_i is the random error term. A positive value of κ_2 suggests that an increase in load leads to an early discharge. This in turn could impact mortality. By definition, early discharges only influence postdischarge mortality.

Next, to demonstrate that an increase in mortality occurs due to an early discharge, we use $EARLYDIS_i$ in a new logistic regression:

$$\begin{aligned}\text{logit}\,[\Pr(MORTALITY\_PD_i)] &= \eta_0 + \mathbf{Z}_i\boldsymbol{\eta}_1 + \eta_2 LOAD_i + \eta_3 OVERWORK_{i,K}\\ &\quad + \eta_4 EARLYDIS_i. \qquad (10)\end{aligned}$$

Positive values for η_2 and η_3 suggest that load and overwork directly contribute to an increase in mortality. Positive values for κ_2 and η_4 suggest that load indirectly contributes to mortality by inducing early discharges.

5.2. Results

We estimate our models based on a sample of 2,740 patients corresponding to all admissions in our study period from 2003 through 2006. Table 4 summarizes the regression results with length of stay as a dependent variable. We find that the length of stay decreases when the load on the system increases. For example, as indicated by the estimation using models (1) and (2), a 10% increase in load on average, leads to a shorter length of stay by 20%. Given that the average length of stay is around two weeks, this amounts to a significant reduction in length of stay of almost 2.5 days on average. However, the variation in *LOAD* in cardiothoracic surgery is relatively low compared to transport service, as indicated by the standard deviations in the descriptive statistics. Thus, only a relatively small fraction of the sample experiences load related changes of more than one day.

Table 4 **Effect of Load, Overwork, and Early Week Admission on Patient Length of Stay**

Coefficient	Model 1	Model 2	Model 3
Intercept	2.21 (0.09)***	2.13 (0.09)***	2.6 (0.07)***
γ_2	−2.07 (0.26)***	−2.08 (0.26)***	−0.58 (0.11)***
γ_3	2.27 (0.36)***	2.28 (0.36)***	
γ_4	−0.09 (0.03)***		−0.09 (0.03)***
R^2	0.24	0.24	0.22
F-statistic	38.3 ($p < 0.01$)	39.89 ($p < 0.01$)	37.6 ($p < 0.01$)

Note. Dummy variables for Y_i (provided in the online appendix) are not displayed.

***, **, and * denote statistical significance at the 1%, 5%, and 10% confidence levels, respectively. Standard errors are shown in parentheses.

We also observe the effect of overwork in cardiothoracic surgery. In estimating (7), we find that $K = 7$ yields the best model fit.[12] As Table 4 illustrates, a 0.01 unit increase in $OVERWORK_K$ leads to a 2% (six hours) increase in the length of stay. Overall, we find that overwork has an important bearing on the performance of the cardiac unit and that high service rates cannot be sustained for longer periods of time, as postulated by the hypothesis in (2). In addition, we find that a weekday admission ($\gamma_4 = 0.09$) is associated with a shorter length of stay. Specifically, a patient who is admitted close to a weekend has a longer length of stay by about 9%. One explanation for this is that staff levels are lower during the weekends. As a result, many services such as imaging, diagnostic testing, and surgical services are curtailed. This means that a patient who is admitted close to a weekend is more likely to wait until the next weekday before full services can be rendered. In particular, nonscheduled patients admitted through the emergency department also have to wait before they can be added to the surgical schedule. This has the effect of increasing overall length of stay for patients who are admitted closer to a weekend.[13]

Now, we turn to the impact of the process variables on mortality and first examine the in-hospital mortalities (Table 5). We do find that overwork has a statistically significant effect ($\mu_3 = 3.53$, p-value $= 0.01$), supporting the hypothesis outlined in Equation (4) that patients admitted to an overworked unit are at increased risk of mortality. In particular, a 10% increase in *OVERWORK* is associated with a 2.2% increase in mortality rate.[14] This result is consistent with findings in the medical literature linking fatigue to a decrease in quality of care. However, our measure of fatigue (or overwork) is obtained from observed workload, whereas the previous studies relied on self-reported measures from service workers. The effect of load is not statistically significant at the 10% level.

[12] Our estimations were performed with varying values for K. The final value of K that was chosen yields the best maximum likelihood value.

[13] We thank the department editor for pointing out how staffing difference between weekdays and weekends could impact patient length of stay.

[14] The increase in probability was estimated using $\partial p/\partial\, OVERWORK = \mu_3 p(1-p)$ with the average mortality rate of $p = 0.068$.

Table 5 In-Hospital Mortality Results

Coefficient	Estimate	Odds ratio	95% wald confidence limit
Intercept	-1.3 (0.83)*		
μ_2	-4.09 (3.52)	0.017	0.002, 0.127
μ_3	3.53 (1.4)***	34.37	1.89, 622
Likelihood ratio (χ^2)	234.6 ($p < 0.0001$)		

Note. Dummy variables for Z_i (provided in the online appendix) are not displayed.

***, **, and * denote statistical significance at the 1%, 5%, and 10% confidence levels, respectively. Standard errors are shown in parentheses.

Next, we look at the results for the postdischarge mortality (Table 6). The coefficients η_2 and η_3 are not statistically significant at the 10% level, suggesting that overwork and load do not directly impact the postdischarge patient mortality. Thus, there is no support for the hypotheses indicated by (3) and (4) when tested on the postdischarge patient mortality. However, the coefficient for *EARLYDIS* (κ_2, Table 7) is estimated to be 7.1 (p-value $= 0.01$), providing strong evidence that an increase in load leads to an early discharge. In particular, a 10% increase in load leads to an early discharge by 0.7 days on average. When we examine the effect of early discharges on the postdischarge mortality rate, we find that the coefficient η_4 has an odds ratio that is close to 1 (coefficient $= 0.13$, odds ratio $= 1.14$), suggesting a small increase in odds of mortality associated with an early discharge. However, the probability of a 30-day postdischarge mortality of any randomly selected patient is less than 1%. Consequently, the corresponding increase in odds by a factor of 1.14 due to an early discharge by a day, is small (less than 1 in 1,000 cases is associated with an early discharge induced mortality). Furthermore, early discharges by more than a day would require the load to increase by more than 13%. Such variations of the load above the mean of 0.78 are infrequent, as indicated by the low standard deviation. Thus, this effect is statistically significant, as hypothesized, but small in absolute magnitude.

Table 6 Postdischarge Mortality Results

Coefficient	Estimate	Odds ratio	95% wald confidence limit
Intercept	-3.31 (2.62)		
η_2	-4.02 (3.38)	0.018	(0.001, 13.606)
η_3	9.2 (5.44)	$>$999	(0.111, 9.19)
η_4	0.131 (0.04)***	1.14	(1.043, 1.246)
Likelihood ratio (χ^2)	48.1 ($p < 0.01$)		

Note. Dummy variables for Z_i (provided in the online appendix) are not displayed.

***, **, and * denote statistical significance at the 1%, 5%, and 10% confidence levels, respectively. Standard errors are shown in parentheses.

Table 7 Early Discharges Resulting from Increased Load

Coefficient	Estimate
Intercept	1.97 (0.67)
κ_2	7.1 (1.4)***
R^2	0.06
F-statistic	5.57 ($p < 0.01$)

***, **, and * denote statistical significance at the 1%, 5%, and 10% confidence levels, respectively. Standard errors are shown in parentheses.

In summary, we find two effects related to quality. First, overwork leads to an increase in the in-hospital mortality rate. Second, increased levels of load lead to early discharges, which in turn is associated with a small increase in the postdischarge mortality rate.

6. Discussions and Future Research

Prior research has assumed that the service rate in a service operations facility is independent of the level of load on the system. We present a model of service worker productivity that includes the effect of load and (over time) the subsequent overwork on service rates. We also consider the quality implications of variable service rates. For the two vastly different services in our study, we find that resources in hospitals are sensitive to their levels of load and that service workers can adapt to system needs by expending more effort to increase the service rate as required.

Various researchers (e.g., Dranove 2002) have reported that hospitals, like most financially oriented entities, have an incentive to increase profits when possible. For instance, Friedman and Pauly (1983) and Anderson and Steinberg (1984) have shown that hospitals exhibit a profit-maximizing response to changes in reimbursement terms. In the United States, the diagnosis related group (DRG) for the diagnosis of the patient at the time of discharge determines the amount that the hospital is paid (Federal Trade Commission and Department of Justice Report 2004).[15] Hospitals receive this payment regardless of the realized cost of care; thus, each additional increase in length of stay beyond the standard expected stay generates zero or minimum marginal revenues.[16] Under DRG-based payment, hospitals have an incentive to increase admissions (Friedman and Pauly 1983).

[15] Each DRG has a payment weight assigned to it, which reflects the average cost of treating patients in that DRG.

[16] Certain hospitals receive an adjusted payment in excess of the standard DRG amount. Actual outlier adjustments are specific to a DRG and are typically made to teaching hospitals and hospitals that treat a disproportionate number of low income patients.

At high levels of load, a hospital's ability to admit new revenue generating patients is reduced. In this paper, we do not empirically examine the underlying incentive schemes. However, it is conceivable that a hospital facing a high load may have a financial incentive to reduce the duration of stay for patients who can be safely discharged earlier, to make room for new admissions. As we demonstrate from our analysis, this could have negative consequences for the patient's quality of care. Any potentially conflicting economic and service quality incentives need to be further examined empirically. With the advent of new policy changes in reimbursements to hospitals such as pay for performance (where a hospital's reimbursement is tied directly to its outcomes), there is greater need to empirically examine the role of hospital operations on its financial health.

We also show that increases in productivity cannot be sustained over a long period of time. Traditional wisdom has been that services should operate at close to full utilization to take advantage of capacity costs. However, sustained levels of high utilization results in overwork and the resultant decrease in productivity may offset any cost savings from operating at high utilization. In many service operations, the impact of high system load on the quality of service is a significant consideration for service managers. In our analysis of cardiothoracic surgery, we find that overwork increases the likelihood of mortality—a finding that is consistent with prior medical literature. We also identify a small decline in service quality which is correlated with an accelerated service rate (or early discharge).

We found the area of hospital operations to be a fruitful area to create a framework of service worker productivity. Future research needs to investigate if and how our findings apply to other services. For example, the impact of load (and queue length) on quality of inspections is of paramount importance in areas such as airport baggage screening (Jacobson et al. 2003) and in port security (Bakshi et al. 2008). One could also expect the effect of load and overwork to impact quality of service in a variety of applications including call centers and financial services (e.g., loan underwriting). Based on our interactions with the medical and business professionals at our research site, we also encountered a great interest to explore questions beyond the research presented in this paper. Future research could also look at accounting for other factors, beyond those used in the study that could affect the case severity. An extreme application in which the interaction between workload, fatigue, and early discharge is especially of interest to the medical community is the intensive care unit, and we hope that future research can extend our analysis to this important area of health-care operations.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The authors are grateful to the cardiac anesthesiologists and executives at the teaching hospital where this study was conducted. The authors thank the *Management Science* review team as well as Chris Lee, Marcelo Olivares, and Stefanos Zenios for their insightful and constructive comments on an earlier version of this paper.

References

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *J. Amer. Medical Assoc.* **288** 1987–1993.

Anderson, G., E. Steinberg. 1984. Hospital readmissions in the medicare population. *New England J. Medicine* **311** 1349–1353.

Bakshi, N., N. Gans, S. E. Flynn. 2008. Measuring the operational impact of container security initiative. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Berk, E., K. Moinzadeh. 1998. The impact of discharge decisions on health care quality. *Management Sci.* **44**(3) 400–415.

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA.

Cakir, A., D. J. Hart, T. F. Stewart. 1980. *Visual Display Terminals: A Manual Covering Ergonomics, Workplace Design, Health and Safety, Task Organization*. John Wiley & Sons, New York.

Caldwell, J. A. 2001. The impact of fatigue in air medical and other types of operations: A review of fatigue facts and potential countermeasures. *Air Medical J.* **20**(1) 25–32.

Crabill, T. B. 1972. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Sci.* **18**(9) 560–566.

Department of Justice and Federal Trade Commission. 2004. Improving health care: A dose of competition. A Report by the Federal Trade Commission and the Department of Justice, July 2004.

Dranove, D. 2002. *The Economic Evolution of American Health Care: From Marcus Welby to Managed Care*. Princeton University Press, Princeton, NJ.

EuroSCORE model. 2007. Retrieved April 23, http://www.euroscore.org.

Friedman, B., M. V. Pauly. 1983. A new approach to hospital cost functions and some issues in revenue regulation. *Health Care Financing Rev.* **4**(3) 105–114.

Fuloria, P. C., S. A. Zenios. 2001. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Sci.* **47**(6) 735–751.

Gaba, D. M., S. K. Howard. 2002. Fatigue among clinicians and the safety of patients. *New England J. Medicine* **347**(16) 1249–1255.

George, J. M., J. M. Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Oper. Res.* **49**(5) 720–731.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3) 321–334.

Green, L. 1984. A multiple dispatch queueing model of police patrol operations. *Management Sci.* **30**(6) 653–664.

Green, L. V. 2004. Capacity planning and management in hospitals. M. L. Brandeau, F. Sainfort, W. P. Pierskalla, eds. *Operations Research and Health Care: A Handbook of Methods and Applications*. Kluwer Academic Publishers, Norwell, MA, 15–42.

Green, L. V., J. Meissner. 2002. Developing insights for nurse staffing. Working paper, Columbia Business School, New York.

Green, L. V., V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *Health Services Res.* **36**(2) 421–442.

Green, L. V., S. Savin, B. Wang. 2006. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1) 11–25.

Huang, X. A. 1995. A planning model for requirement of emergency beds. *J. Math. Appl. Medicine Biol.* **12** 345–353.

Jacobson, S. H., J. Virta, J. M. Bowman, J. E. Kobza, J. J. Nestor. 2003. Modeling aviation baggage screening security systems: A case study. *IIE Trans.* **35**(3) 259–269.

Kurki, T. S. 2002. Prediction of outcome in cardiac surgery. *Mount Sainai J. Medicine* **69**(1–2) 68–72.

Kwak, N., C. Lee. 1997. A linear programming model for human resource allocation in a health-care organization. *J. Medical Systems* **21** 129–140.

Nashef, S. A., F. Roques, B. G. Hammill, E. D. Peterson, P. Michel, F. L. Grover, R. K. Wyse, T. B. Ferguson. 2002. Validation of European system for cardiac operative risk evaluation (EuroSCORE) in North American cardiac surgery. *Eur. J. Cardiothoracic Surgery* **22** 101–105.

Needleman, J., P. Buerhaus, S. Mattke, M. Stewart, K. Zelevinksy. 2002. Nurse-staffing levels and the quality of care in hospitals. *New England J. Medicine* **346**(22) 1715–1722.

Needleman, J., P. Buerhaus, M. Stewart, K. Zelevinksy, S. Mattke. 2006. Nurse staffing in hospitals: Is there a business case for quality? *Health Affairs* **25**(1) 204–211.

Newhouse, J. P. 1996. Reimbursing health plans and health providers: Efficiency production versus selection. *J. Econom. Literature* **34** 1286–1263.

Powell, S. G., K. L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Sci.* **50**(8) 1095–1105.

Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The effects of low inventory on the development of productivity norms. *Management Sci.* **45**(12) 1664–1678.

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Sci.* **44**(12) 1595–1607.

Scott, L. D., A. E. Orgers, W. Hwang, Y. Zhang. 2006. Effects of critical care nurses' work hours on vigilance and patient's safety. *Amer. J. Critical Care* **15**(1) 30–37.

Setyawati, L. 1995. Relation between feelings of fatigue, reaction time and work productivity. *J. Hum. Ergol.* (*Tokyo*) **24**(1) 129–35.

Smith-Daniels, V., S. B. Schweikhart, D. E. Smith-Daniels. 1988. Capacity management in health care services: Review and future research directions. *Decision Sci.* **19** 889–919.

Stidham, S., Jr., R. R. Weber. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Oper. Res.* **37**(4) 611–625.

Tanabe, S., S. Nishihara. 2004. Productivity and fatigue. *Indoor Air* **14**(Suppl. 7) 126–133.