Kuntz, Ludwig; Mennicken, Roman; Scholtes, Stefan

Working Paper

Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality

Ruhr Economic Papers, No. 277

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Kuntz, Ludwig; Mennicken, Roman; Scholtes, Stefan (2011) : Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality, Ruhr Economic Papers, No. 277, ISBN 978-3-86788-322-1, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI), Essen

This Version is available at:
https://hdl.handle.net/10419/61454

RUHR
ECONOMIC PAPERS

Ludwig Kuntz
Roman Mennicken
Stefan Scholtes

Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality

RWI #277

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #277

Responsible Editor: Christoph M. Schmidt

ISSN 1864-4872 (online) – ISBN 978-3-86788-322-1

Ruhr Economic Papers #277

Ludwig Kuntz, Roman Mennicken, and Stefan Scholtes

Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

ISSN 1864-4872 (online)
ISBN 978-3-86788-322-1

Ludwig Kuntz, Roman Mennicken, and Stefan Scholtes[1]

Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality

Abstract

We discuss the impact of organizational workload on professional service outcomes, such as survival rates in hospitals. The prevailing view in the literature is that service quality deteriorates when organizational workload increases. In contrast, we argue that the relationship between workload and service outcomes is nonlinear and that there is a quality-optimal workload level. Whilst outcomes deteriorate with increasing workload when workload levels are already high, they will improve if workload increases from a low level. We reach this hypothesis by combining three perspectives: (i) the queuing theory perspective, with its focus on congestion, (ii) a discretionary choice perspective, with a focus on decisions made by professionals in response to changes in workload, and (iii) an endocrinological perspective, with a focus on the subconscious effects of workload on worker performance through the cognitive impact of stress hormones. Using a patient census of 1.4 million patients in 624 departments across 101 hospitals, we provide empirical support for the nonlinearity hypothesis in the context of hospital survival rates. We further discuss the implications for hospital capacity planning and the wider implications for service operations management.

JEL Classification: I12, M11, M54

Keywords: Service quality; service outcomes; organizational workload; hospital capacity planning; behavioral operations; stress

August 2011

1 *Ludwig Kuntz, Faculty of Management, Economics and Social Sciences, University of Cologne; Roman Mennicken, RWI and Faculty of Management, Economics and Social Sciences, University of Cologne; Stefan Scholtes, Judge Business School, University of Cambridge. – All correspondence to Roman Mennicken, RWI, Hohenzollernstr. 1–3, 45128 Essen, Germany, E-Mail: roman.mennicken@rwi-essen.de..*

1. Introduction

Governments and health care professionals across the world are deliberating the rising cost of health care. Aging populations and unhealthy lifestyles are driving the relentless demand for ever more comprehensive health care services. Whilst cost containment has always been on the agenda, the recent economic recession has pushed it to the forefront. Health care providers are seeing their revenues fall in the wake of austerity measures implemented in a bid to reduce national deficits. As a result, hospitals and other health care organizations are having to make significant efficiency savings - and fast.

To cut staff is the hospital manager's knee-jerk reaction to mounting cost pressure. Over 50% of hospitals in a recent survey by the American Hospital Association had reduced staff to cope with the economic downturn (American Hospital Association 2011b). This is unsurprising, given that staffing represents the largest cost pool, with over two thirds of every hospital dollar spent on staff wages and benefits (American Hospital Association 2011a). Additionally, staff cuts can be quickly implemented across the organization through recruitment freezes and redundancies.

In contrast to many other industries, demand for health care services does not decline in economically challenging times. When hospital managers reduce staff numbers, workload will inevitably increase, leaving clinicians to wrestle with the corresponding impact on service quality. What is the nature of the relationship between organizational workload and the quality of hospital services? This is the contextual question we address in this paper.

Service quality is an abstract and multi-dimensional construct, in particular in the context of a hospital, with its complex range of services. To put our study in perspective, we identify three key dimensions of service quality, which are related but have different measurement foci. The first dimension, *congestion-related* service quality, is concerned with speed and throughput. Typical measures in the hospital

context are waiting times and length of stays. Congestion has been, and remains, a key concern in operations management (Hopp et al. 2007, Kc and Terwiesch 2009, Ramdas and Williams 2009, Kc and Terwiesch 2010). The second quality dimension, *perception-related* service quality, focuses on overall experience and outcome, as perceived by the service consumer. It can be gauged by direct measures, such as consumer feedback, and indirect measures, such as loyalty and advocacy.

This paper is concerned with a third quality dimension, *outcome-related* service quality (Ata and van Mieghem 2009, Wang et al. 2010, Kc and Terwiesch 2010, Anand et al. 2011). In contrast to perceived quality, service outcome is determined by objective third-party assessment - usually in the form of expert professional peers. This dimension is particularly relevant for complex professional services, such as health care, as consumers normally lack the requisite knowledge or experience to assess the quality of the service they receive. In our empirical study we are concerned with a particularly important outcome measure of hospital services: a patient's probability of surviving hospitalization.

The causal variable of interest here is organizational workload. Workload refers not just to work volume per se, but to work volume relative to a set of organizational resources that define its capacity. We use the term 'organizational workload' to denote the percentage utilization of an organization's service delivery capacity. Workload is not constant, but varies over time. The specific focus of this paper is on variation in workload between service episodes, with a typical patient stay as an exemplary reference period. The effects of variations in long-term average workloads between organizations have been discussed elsewhere, for example in relation to learning curve effects induced by high cumulative volume (Pisano et al. 2001, Halm et al. 2002) or chronic effects of stress and burn-out on productivity (Dahl 2011). In this empirical study we model differences in long-term average workload levels between organizations as organizational fixed effects over the observation period.

A common measure of organizational workload in the hospital context is bed occupancy. However, published numbers of certified beds are an unreliable indicator of hospital capacity. Certified beds can be unstaffed and effectively mothballed, whilst hospitals can also shift bed capacity between clinical departments. As we are concerned here with departmental workload, we will use an alternative and more general measure of capacity: the maximum number of patients treated in a department on any one day during the observation period. The department's daily workload is then the patient volume in the department on that day, expressed as a percentage of the capacity measure. The question we ask is how a patient's chances of in-hospital survival alter with the average daily workload that the department experiences during the patient's stay.

Several studies in the medical literature argue that clinical quality deteriorates as workload increases (Weissman et al. 2007, Schilling et al. 2010). However, a recent study of cardiothoracic patients in a US hospital failed to identify a significant effect of workload on in-hospital survival probability (Kc and Terwiesch 2009). We provide an explanation as to why a general quality deterioration hypothesis is difficult to maintain. We argue that the relationship between workload and service quality is best understood as a nonlinear phenomenon: increasing workload leads to improved quality when workload levels are low, whilst quality deteriorates when workload further increases from already high levels. As a consequence, it is possible to identify a quality-optimal workload level - a tipping point beyond which quality deteriorates, and often rapidly so.

2. Hypothesis development

We distinguish three partial effects by which workload variation can impact outcome-related service quality: (i) resource availability alters with workload, (ii) workers make conscious decisions in response to changing workload, and (iii) workload acts as a stressor and triggers a subconscious stress response in workers. Whilst

the first two perspectives have been widely studied in the operations literature, the stress response, which is particularly relevant to error propensity, has been neglected so far.

2.1. Congestion-induced effects

When workload increases, limited resources must be shared between a greater number of consumers, leading to increased congestion and waiting times. Longer idle waiting during a service episode does not only negatively affect the consumers' service perception, but can also have a detrimental effect on service outcomes if a consumer's condition deteriorates over time (Rosanio et al. 1999) or if she is exposed to environmental threats. For example, a disease may progress while a patient is waiting for treatment and the longer a patient remains in hospital, the greater the risk of contracting hospital acquired infections.

Waiting times increase nonlinearly with workload: a percentage point increase at low or medium workload levels will have a less pronounced effect on waiting times than a percentage point increase when workload is already high. This effect is captured in the waiting time formulas of queueing theory. These formulas can typically be decomposed as a product of a term that depends only on characteristics of the service process and is independent of the traffic rate, and a term of the form $\frac{\rho}{1-\rho}$, where ρ denotes capacity utilization. An example is the Pollaczek-Khinchin formula for the expected waiting time in an M/G/1 queue $W(\rho) = \alpha \frac{\rho}{1-\rho}$, where $\alpha = \frac{1+C^2}{2\mu}$ depends only on the service rate μ and the coefficient of service time variation C. As expected waiting times are proportional to $\frac{\rho}{1-\rho}$, they increase rapidly when utilization ρ approaches 100%.

It is difficult to argue that congestion-related waiting, i.e., waiting that is unintended and not part of the service protocol, might systematically lead to *better* outcomes. If that were the case, such beneficial waiting times should be worked into

the service protocol. However, potential outcome deterioration as a consequence of congestion-related waiting is well documented. For stroke patients, for example, a long waiting time for treatment is associated with a considerably worsened prognosis (Hacke et al. 2004).

Whilst waiting can have a negative outcome effect, it is likely that this effect is negligible when waiting times are very short. This is an important assumption for our hypothesis development. Formally, if a function $f(W)$ describes outcome-related quality as a function of waiting time W, then we assume that $f'(W) \leq 0$ and, importantly, that $f'(0) = 0$. If waiting time increases with utilization, i.e., $W'(\rho) > 0$, then the partial effect of congestion on quality, $Q_C(\rho) = f(W(\rho))$, will satisfy the following conditions:

ASSUMPTION 1. *$Q'_C(0) = 0$ and $Q'_C(\rho) \leq 0$.*

If the marginal quality deterioration increases in absolute value with waiting time, i.e. if $f''(W) < 0$, and if waiting time increases in a convex manner with utilization ρ, i.e., if $W''(\rho) > 0$ as in the Pollaczek-Khinchin formula, then $Q''_C(\rho) < 0$, i.e., quality deteriorates more rapidly at higher workload levels. In the health care context such effects could occur, for example, when the marginal probability of contracting an infection increases with patient density (Archibald et al. 1997).

2.2. Discretion-induced effects

Classical queuing theory assumes that service provision is unaffected by workload, that variations in workload are buffered entirely by waiting times. This is unrealistic in the context of professional services, where workers have a degree of discretion over service provision. When professionals experience high workload they may decide to cut corners to reduce service times and improve throughput, accepting associated service quality compromises for individual consumers. In this context Hopp et al. (2007) refer to quality as an additional variability buffer.

Within the hospital context, Kc and Terwiesch (2009, 2010) demonstrate empirically that increased workload is associated with reduced service time, measured as a patient's length of stay in the hospital: when workload increases, patients are discharged sooner. Cutting corners in this way can clearly have detrimental effects on service outcomes. Hugonnet et al. (2007) study the relationship between workload and infection rates and conclude "*Low staffing level was followed only a few days later by the occurrence of infections. This suggests that under the pressure of increased workload, healthcare workers do not comply with infection control measures, such as hand hygiene, due to time constraints.*"

Hugonnet et al. (2007) also provide evidence that discretionary reduction in service provision and its associated negative effect on service outcomes will predominantly occur with high workloads. Since service length is not the focus of our paper, we only report briefly that this nonlinear effect of workload on service length could be confirmed with our sample of approximately 1.4 million patient episodes from 624 German hospital departments. Figures 1 and 2 summarize estimated logarithmic length of stay and 95% confidence intervals as a function of workload for all patients (full sample) and for a subsample of patients with high mortality risk. The details of the employed spline regression methodology, control variables, and the conditions included in the high-risk subsample are explained in Section 3. The estimated curves confirm the refined service length hypothesis: when workload is low, doctors do not appear to use their discretion over premature service completion. Increased capacity utilization leads to a moderate increase in length of stay, in line with the predictions of queueing theory. However, when workload is very high, the results confirm the findings of Kc and Terwiesch (2009, 2010): doctors increase patient throughput.

In some contexts, workers may also use their discretion when workload is low. Specifically, if organizations or individual workers benefit from recorded activity

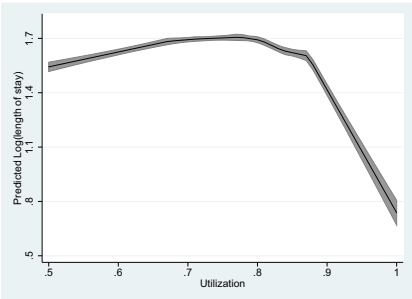

Figure 1 **Length of stay for full sample**

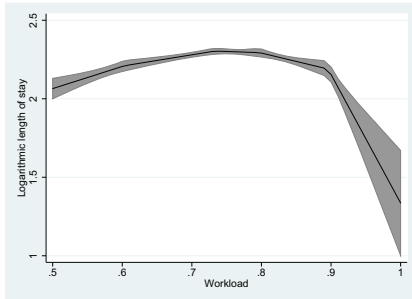

Figure 2 **Length of stay for high risk sample**

rather than outcomes and the need for specific activities cannot be directly assessed by the consumer or the payer, then workers may use their discretion in performing *more* activity than is strictly necessary. This effect is likely to occur when workload is low and much capacity is unused. In the health care context this phenomenon is known as over-treatment and is often attributed to perverse incentives created by payment contracts. The effect of over-treatment on clinical outcomes depends on the context: existing studies either fail to show a significant effect or confirm a negative effect on clinical outcomes (see e.g. Torres and Santiago (2004)).

In summary, the discretion perspective leads to the following general impact of workload on service quality: at low workload levels discretion is either not exercised, and subsequently service quality remains unaffected, or service quality increases with increasing workload as unnecessary activity is reduced. At high workload levels, however, increased workload leads to a deterioration in outcomes as professionals use their discretion and cut corners to improve throughput. This is summarized in the following assumption on the partial effect Q_D of discretion on outcome quality as a function of workload.

ASSUMPTION 2. $Q'_D(0) \geq 0$, $Q'_D(1) < 0$.

2.3. Stress-induced effects

Edmondson and Tucker (2001) distinguish between problems and errors in a service

process. They define a problem as a "*disruption of the worker's ability to execute a prescribed task*". In relation to workload, an inappropriate time allocation disrupts a worker's routine. In contrast, an error is defined as "*an execution of a task that is either unnecessary or incorrectly carried out and that could have been avoided with appropriate distribution of pre-existing information.*" Importantly, "*workers are well aware of the problems they encounter. In contrast, by definition, people are unaware of their own errors while making them.*" It is in the context of problem solving that workers use their discretion over service provision as discussed above. To develop a fuller picture of the effect workload has on outcome-related quality, we also need to address subconscious impulses that affect a worker's error making propensity. Drawing on the endocrinology literature, we argue that propensity for errors is *reduced*, and subsequently outcomes improve, when workload increases from a low level up to a certain threshold, beyond which error-making increases and service quality deteriorates.

There is ample evidence in the medical and psychological literature that workload acts as a stressor, i.e., increased workload leads to increased stress hormone levels. In a meta-analysis review of 208 laboratory studies, Dickerson and Kemeny (2004) found that performance tasks that contained uncontrollable elements and could be critiqued by others elicited a human cortisol response. Heightened workload increases the number of such tasks and decreases the level of control as time pressures mount. Sonnentag and Fritz (2006) review studies of the effect of day-to-day workload variation and conclude that cortisol secretion increases as short-term workload increases. This link is confirmed in a wide spectrum of professional service contexts, including air traffic controllers (Zeier et al. 1996), managers (Lundberg and Frankenhaeuser 1999), medical staff in neonatal and pediatric intensive care units (Fischer and et al. 2000) and ambulance service personnel (Backe et al. 2009).

The effect of stress hormone levels on performance has been a central topic in endocrinology since Selye's seminal proposal to study stress as an organism's generic, nonspecific response to different exogenous strains (Selye 1936, McEwen 2002). Of particular interest in this context is the relationship between stress hormone levels and cognitive performance. Following the discovery that some stress hormones, such as cortisol, can cross the blood-brain barrier and affect neurons directly via receptors (McEwen et al. 1968), researchers have made significant advances in explaining how stress hormones affect cognitive functions. Among other things, stress hormones have been shown to control the excitability of neurons in those regions of the brain that are central to memorizing and learning. In a recent review article Lupien et al. (2007) summarize the state of knowledge relevant to our study: "*We have shown here that the effects of stress hormones on human cognition are best understood in line with the inverted U-shape function between glucocorticoids and cognitive performance.*" The inverted U-shape is supported both by scientific theory, based on the interplay between two receptor types that differ significantly in their affinity for glucocorticoids (de Kloet et al. 1999) and by empirical evidence, based on randomized controlled trials (Lupien et al. 1999).

In summary, there is evidence that hormone levels are a monotone function of workload and that cognitive performance has an inverse U-shaped relationship with hormone levels. As workload increases, stress hormone levels increase and it is therefore plausible to assume that a worker's propensity to make errors decreases with workloads at low workload levels and increases at high levels of workload.

It is not obvious how characteristics of worker error rates translate to quality effects at customer level, as workers use their discretion in deciding how much time they will spend with an individual customer. Whilst a worker's propensity to make errors may be increased at times of heavy workload, the same worker will have less time with each patient in which to make an error, as she shortens service times in

response to work pressure. To discuss this effect, we assume that worker i's errors occur as Poisson events at a rate $\lambda_i(x)$ at workload x. If $t_i(x)$ is the amount of time she spends with a specific customer during his service episode, then $t_i(x)\lambda_i(x)$ is the rate at which the customer experiences errors by worker i during his service episode. At low workload levels, the stress response leads to decreasing worker error rate $\lambda_i(x)$. If the worker's time with the customer does not increase, i.e., if $t_i'(x) \leq 0$, the beneficial effect translates to the customer's episode error rate, as the worker makes fewer errors across shorter time periods. At high workload levels stress leads to increased worker error rates, whilst at the same time workers may reduce their individual customer exposure times in response to increasing workloads. The episode error rate $t_i(x)\lambda_i(x)$ will only increase if the percentage rise in the worker's error rate exceeds the percentage reduction in her exposure time to the customer. These arguments extend to a customer's overall episode error rate across all workers involved in his service, as the overall episode error rate is the sum of the episode error rates $t_i(x)\lambda_i(x)$ over all workers i. The following proposition summarizes this conclusion. You will notice that the condition $\frac{\lambda_i'(x)}{\lambda_i(x)} \geq -\frac{t_i'(x)}{t_i(x)}$ is equivalent to $(t_i\lambda_i)'(x) \geq 0$ and captures the fact that the percentage reduction in exposure time to worker i is an insufficient counterbalance to the percentage deterioration of worker i's error rate.

PROPOSITION 1. *Suppose a customer is being served by m workers, individual worker errors follow m independent Poisson processes with differentiable error rates $\lambda_i(x)$ as a function of workload x, and that worker i spends time $t_i(x)$ with a customer during their service.*

1. *If $\lambda_i'(x) < 0$ and $t_i'(x) \leq 0$ for all i then the customer's episode error rate decreases with workload at level x.*

2. *If $\lambda_i'(x) > 0$ and $\frac{\lambda_i'(x)}{\lambda_i(x)} \geq -\frac{t_i'(x)}{t_i(x)}$ for all i, with strict inequality for at least one i, then the customer's episode error rate increases with workload at workload level x.*

For the development of our hypothesis we assume that the first part of the proposition applies to low workload levels and the second to high levels, leading to the following assumption on the partial effect of stress on outcome quality.

ASSUMPTION 3. *$Q'_E(0) > 0$ and $Q'_E(1) < 0$.*

2.4. Hypothesis

Having considered the effects of congestion, professional discretion and stress on outcome-related service quality and argued that at low workload levels waiting-related effects are insignificant, decision-related effects are either insignificant or positive, and stress-related effects are positive, we surmise that quality increases with workload at low workload levels. At high levels, however, all three effects point towards the negative and outcome-related quality deteriorates with increased workload. These effects are summarized in our central hypothesis.

HYPOTHESIS. *The workload of professional service organizations has a nonlinear effect on outcome-related service quality. At low workload levels an increase in workload results in a more positive outcome, whilst at high workload levels, an increase leads to a more negative outcome. There is an optimal workload level with regard to outcome-related service quality.*

We have made three important assumptions in the development of this hypothesis: (i) the quality deterioration effect of waiting is negligible when waiting times are very short, (ii) the time workers spend with customers decreases with workload and (iii) worker error rates deteriorate considerably with very high workloads. The first two assumptions relate to the hypothesized effect at low workloads: firstly, if quality deteriorates markedly with waiting, even when waiting times are very short, and an increase in workload leads to increased waiting even at very low workload levels,

then this negative effect could potentially dominate any beneficial stress effect at low workload levels. Secondly, if workers spend more time with customers when workload increases from low levels, specifically if the percentage increase in exposure time exceed the percentage improvement in error rates, then the stress-induced effect itself is negative from the outset. In both cases, workload would have a negative effect on outcomes even at low workload levels. The third assumption relates to the hypothesized effect at high workload levels: if error rates increase only marginally relative to reduced exposure times then the stress-induced effect remains positive at high workload levels and could, at least in principle, outweigh the negative effects of congestion and deliberate cutting of corners.

3. Empirical Study

3.1. Data

The data for this study consists of a patient census from 101 German hospitals. For 72 of these hospitals the database contains administrative hospital discharge records of all patients discharged over one year - either 2004 or 2005. For the remaining 29 hospitals all patients discharged during the two year period 2004-2005 are included. The database contains 1,415,754 cases across 624 hospital departments. The fact that the data constitute a complete census of the departments is significant as it allows us to calculate workload proxies at department level.

We use a patient's probability of in-hospital survival as a metric of outcome-related service quality (Gaynor et al. 2005, Huckman and Pisano 2006, Kc and Terwiesch 2009). The US Department of Health identifies six conditions "*for which mortality has been shown to vary substantially across institutions and for which evidence suggests that high mortality may be associated with deficiencies in the quality of care*". These are: acute myocardial infarction (AMI), congestive heart failure (CHF), gastrointestinal hemorrhage (GIH), hip replacement after fracture (HIP),

pneumonia (PNE) and stroke (STR) (Agency for Healthcare Research and Quality 2006). We decided to study the effect of workload on the survival of patients with these primary conditions, giving us a subsample consisting of 85,321 patient episodes across 393 departments in 93 hospitals.

Since we use discharge records, patients who are admitted during the study period but discharged outside of this window are not included in the data. Consequently, the end of the study period does not constitute a complete patient census and would lead to a bias in workload estimates. Similarly, if a patient was admitted before the study period, we cannot calculate the workload during their entire stay. To account for these censoring issues, we exclude patients who were admitted before the hospital's observation period or discharged during the final month of the observation period. This is prudent in light of an average length of stay of 11 days for the patients in our subsample. Additionally, we exclude departments where either no or all patients survived and departments with fewer than 20 patients over the course of the observation period. Following these exclusions the sample consists of 75,314 patient episodes across 243 departments in 87 hospitals.

3.2. Variables

Organizational workload during a patient episode. To compute workload as the percentage utilization of the department's capacity during a patient episode, we need to measure departmental capacity, i.e., the maximum number of patients that can be treated in a department on any one day. Whilst the natural measure is the number of staffed beds, this number is not publicly available. Public documents refer to the number of *certified* hospital beds. Interviews with hospital managers have revealed that this number can deviate significantly from the number of staffed beds and is not a reliable measure of operational capacity. In the absence of reliable

staffed bed numbers, we use the maximal number of inpatients in the department on any one day during the observation period as a measure of departmental capacity. We compute patient episode workload as the ratio of the average daily patient volume in the department during the episode to the department's capacity.

Patient risk factors. The discharge records contain several variables that allowed us to control for patient heterogeneity. Beside the primary medical condition and important individual risk factors (i.e., age, gender, emergency status), the presence of secondary diagnoses is a potential source of heterogeneity. To account for these comorbidities we adopted a standard approach, developed by Elixhauser et al. (1998), to the German ICD-10 system, following Quan et al. (2005). One of the original comorbidities in the Elixhauser model, HIV, had a very low incidence rate in our subsample and was therefore omitted and all patients with HIV comorbidity removed from the sample. A frequently applied alternative to the Elixhauser model is a comorbidity index developed by Charlson et al. (1987). Both models produced very similar estimates. We report only results using Elixhauser comorbidities.

Severity of patient-mix in the department. In addition to controls for the clinical conditions of patient i, we control for the severity of condition of the other patients in the department during patient i's stay. Following Weissman et al. (2007), we use the diagnosis related groups (DRG) in which patients are categorized for reimbursement purposes. Each DRG has an associated cost weight (CW) which reflects the treatment cost of a typical patient in this group in an average hospital. We calculated the average cost weight of all patients in patient i's department for each day of patient i's stay and averaged this number across patient i's length of stay. Following Weissman et al. (2007), we use this average cost weight as a control variable rather than as a second independent variable because it accounts not only for resources used in the department itself but includes costs incurred in the operating theater and other functional departments.

To check for robustness, we have estimated models with alternative controls for patient severity across the department. The first is a staff cost weight (SCW), derived from the cost weight using labor cost rates published by the regulator for the German DRG system, InEK. The staff cost weight captures the labor intensity of a patient rather than the general resource intensity. As a second, more clinically focused measure of severity across the patient pool, we used the average Charlson comorbidity index of the other patients in the department (see Charlson et al. (1987)). The estimations of workload effects were robust within these three specifications. We only report results for staff cost weight as a departmental case complexity control.

Staffing and seasonal patterns. To control for varying staffing patterns, we follow Kc and Terwiesch (2009) and include a weekend and public holiday dummy variable. This controls for the so called 'weekend effect' (Bell and Redelmeier 2001). Since the average length of stay for patients in our sample is more than 10 days, patients who are admitted on weekends or public holidays are likely to experience more days with lower staffing levels than patients admitted during the week. We also included dummies for month-of-the-year to capture longer-term temporal factors, such as a demand spike during flu season. Finally, we include a year dummy for 2005 to control for systematic differences between the observation years.

Department fixed effects. There is an ongoing debate in the literature about the impact of hospital characteristics on quality of care (Shortell and Hughes 1988, McClellan and Noguchi 1998, Gaynor et al. 2005). To account for such effects in an aggregate manner, we include department-within-hospital fixed effects and cluster standard errors at hospital level.

Table 1 shows the summary statistics of survival, workload variables and patient risk factors.

Table 1 Descriptive Statistics for Main Variables

	Mean
Survival probability	0.906
Patient-level workload (fixed capacity during observation)	0.742
Patient-level workload (varying capacity during observation)	0.791
Fraction of patients admitted on weekends or public holidays	0.226
Fraction of male patients	0.497
Fraction of emergency admissions	0.536
Fractions of primary conditions	
Acute myocardial infarction (AMI)	0.191
Pneumonia (PNE)	0.204
Stroke (STR)	0.292
Congestive heart failure (CHF)	0.070
Hip replacement after fracture (HIP)	0.128
Gastrointestinal hemorrhage (GIH)	0.115
Age (in years)	68.32
	(19.67)
Patient episodes N	75,314

3.3. Model and estimation method

We estimate the survival probability P_{ijk} for patient i in department j of hospital k with a logit model

$$\text{logit}(P_{ijk}) = \beta_0 + \sum_{l=1}^{n} \beta_{1l} f_l(W_{ijk}) + \beta_2 R_{ijk} + \beta_3 S_{ijk} + \beta_4 D_{jk} + \epsilon_{ijk}. \tag{1}$$

W_{ijk} denotes the independent variable workload and the remaining variables are controls. The vector R_{ijk} contains risk control factors, specifically dummy variables for the main condition, with acute myocardial infarction (AMI) as the reference condition, and gender, age, age squared, emergency admission status and dummy variables for the Elixhauser comorbidities, as well as the staff case weight. The vector S_{ijk} contains seasonal dummy variables for admissions on weekends or public holidays, for month-of-the-year and for year of admission; D_{jk} is the vector of department dummy variables to capture department fixed effects.

To account for the hypothesized nonlinear relationship between workload and survival probability, we have estimated a linear spline regression model (Marsh and Comier 2001). This structure is captured in the term $\sum_{l=1}^{n} \beta_{1l} f_l(W_{ijk})$, where n

is the number of splines and the coefficients β_{1l} represent the slopes of the linear segments. The model assumes that workload has a piecewise linear effect on service quality. This guarantees that the effect changes continuously with varying workload values.

We apply spline models instead of the more ubiquitous polynomial models as piecewise linear functions are more sensitive to changing gradient signs over the range of the independent variable. For example, a square term in a polynomial can be significant because it contributes substantially to model fit over a local subregion. The effect of the square term, however, is maintained globally; it carries over to neighboring regions and introduces a spurious nonlinearity in regions where there is none. This is not the case with piecewise linear functions, where slopes can undergo discontinuous changes.

To simplify the interpretation of coefficients in spline regressions it is useful to specify a piecewise linear function as a linear combination of piecewise linear basis functions $f_i, i = 1, \ldots, n$, in the following way: let $0 < \mu_1 < \ldots < \mu_{n-1} < 1$ be a chosen set of nodes where the linear pieces are joined. The first function $f_1(x) = \min\{x, \mu_1\}$ equals x if $x \leq \mu_1$ and μ_1 if $x > \mu_1$; the last function $f_n(x) = \max\{x - \mu_{n-1}, 0\}$ equals $x - \mu_{n-1}$ if $x \geq \mu_{n-1}$ and zero if $x < \mu_{n-1}$. Between these two functions lie $(n-2)$ functions of the form

$$f_l(x) = \max\{\min\{x, \mu_l\} - \mu_{l-1}, 0\} = \begin{cases} 0 & \text{if } x \leq \mu_{l-1} \\ x - \mu_{l-1} & \text{if } \mu_{l-1} \leq x \leq \mu_l \\ \mu_l - \mu_{l-1} & \text{if } x \geq \mu_l. \end{cases}$$

Note that the function $\sum_{l=1}^{n} \beta_{1l} f_l(x)$ is linear if all coefficients β_{1l} coincide. The coefficient β_{1l} can be readily interpreted as the slope of the fitted function on the interval (μ_{l-1}, μ_l) with $\mu_0 = 0, \mu_n = 1$.

We estimate a logit model with dummy variables for hospital departments, as well as a conditional fixed-effects logit model. The conditional fixed-effects logit

model uses fewer degrees of freedom as it does not explicitly estimate parameters for the department dummy variables. This allows us to calculate a likelihood ratio (LR) chi-square test. The logit model with department dummy variables has the advantage that it allows us to calculate partial effects, while the conditional fixed-effect logit model would only allow the prediction of $P(1|\text{Fixed Effect}=0)$, which results in misleadingly low survival probabilities (Wooldrige 2002). The coefficient estimates obtained from these two models are very similar. As mentioned earlier, standard errors are clustered at hospital level to take account of the hierarchical nature of the model specification.

3.4. Model selection

To select an appropriate spline model, we follow Royston and Sauerbrei (2007), using the Stata command *uvrs*. We begin by choosing a maximum number n of linear spline pieces that we allow for the most complex spline model and fix the $(n-1)$-quantiles of the empirical workload distribution as the set of candidate nodes where spline pieces may be joined. We first estimate the model with n spline pieces. The model selection procedure then compares this benchmark model, in terms of fit, with estimations of simpler but increasingly more complex models with $k=0,1...,n-1$ spline pieces, where $k=0$ and $k=1$ correspond to a model without the workload variable and to a linear model, respectively. The procedure stops and selects the model with k spline pieces if the benchmark model with n pieces does not provide a significantly better fit, based on the chi-square statistic of log-likelihood differences. If the benchmark model fits significantly better, the procedure proceeds to splines with $k+1$ pieces. The benchmark model is selected if it fits significantly better than any of the other simpler models.

Since the candidate nodes are the $(n-1)$-quantiles, we have a choice between these nodes when specifying a spline with $k<n$ pieces. First, for a spline with $k=2$

pieces there are $n-1$ possible models, one for each of the candidate nodes. We estimate all $n-1$ models and select the model with the maximum likelihood function value. The node corresponding to the selected model is considered *identified* and will be kept as a node if we estimate splines with more pieces. If $k \geq 2$ and the procedure progresses from models with k pieces to models with $k+1$ pieces because the benchmark model still has a significantly better fit, then we have a list of $k-1$ identified nodes, which we include in all models with $k+1$ pieces. To specify the spline with $k+1$ pieces, we therefore only need to identify one new node from the remaining $n-k$ non-identified nodes. We estimate the respective $n-k$ models and again select the model with the maximum likelihood function value as the preferred model with $k+1$ pieces, to be compared with the benchmark model. We add the corresponding new node to the list of identified nodes.

The model selected by the above procedure depends on the maximal number n of spline pieces and the respective candidate nodes, which are set at the $(n-1)$-quantiles of the workload distribution. To find an overall best fitting model, we executed the procedure for $n = 1, \ldots, 20$ and selected the final model from the resulting 20 models on the basis of the Bayesian Information Criterion (BIC) as suggested by Long and Freese (2006).

4. Results

4.1. Results of Pooled Analysis

Table 2 contains a subset of the coefficient estimates obtained by the spline regression across the 6 conditions. The model selection procedure identifies two splines as the best model, i.e., more complex models did not improve the model fit significantly. The selected model identified the node at a workload of 87.1%.

The hypothesis is supported by both logit models: survival probability increases significantly with workload at low workload levels and decreases significantly at high levels. To illustrate the magnitude of this effect, we set all control variables to

Table 2 Selected coefficients of logit of survival odds

	Logit with dummies	Conditional Logit
$\min(\text{Workload}, 0.8712)$	1.130***	1.123***
	(0.190)	(0.189)
$\max(\text{Workload} - 0.8712, 0)$	-10.06***	-10.01***
	(1.365)	(1.359)
Severity of patient-mix	-0.0850	-0.0844
	(0.106)	(0.105)
Admitted on weekends or public holidays	-0.0599*	-0.0598*
	(0.0279)	(0.0278)
PNE	0.204**	0.203**
	(0.0740)	(0.0737)
STR	-0.0604	-0.0602
	(0.0912)	(0.0908)
CHF	0.764***	0.761***
	(0.134)	(0.133)
HIP	1.172***	1.163***
	(0.222)	(0.220)
GIH	1.046***	1.042***
	(0.0877)	(0.0872)
Observations	75,314	75,314
LR Chi-Square	.	18,392.1

Standard errors adjusted for clustering within hospitals

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

their mean values and varied workload. If workload is at a low level of 50%, average survival probability across the six conditions and the included hospital departments is estimated at 93.73% (95% CI: $93.18 - 94.28$). The probability increases to 95.78% (CI: $95.52 - 96.04$) at the estimated optimal workload level of 87%, beyond which estimated survival probabilities drop sharply to a 86.16% chance of survival at full capacity.

4.2. Results for Specific Conditions

In the pooled analysis we accounted for the six medical conditions through dummy variables and assumed a homogeneous nonlinear effect of workload. To obtain a more granular understanding of workload effects we examined the effect separately for each clinical condition. The model remains the same as in equation (1) with the

exception that R_{ijk} no longer includes dummy variables for the main conditions.

As we move from the pooled sample to individual conditions we lose sample size and therefore statistical power. Since we are attempting to explain changes in avoidable deaths through workload, rather than deaths per se, and avoidable mortality occurs much less frequently than all-cause mortality, a large sample size is required to detect a statistically significant signal (Peduzzi et al. 1996). We therefore focus in this section on the three conditions with the largest sample size, which also have the highest all-cause mortality rates - PNE, STR and AMI. The sample size and mortality rates for the remaining conditions CHF (N=4247, 7.7% mortality), GIH (N=7308, 5% mortality) and HIP (N=9243, 4.3% mortality) are too low to obtain robust significant results. The corresponding estimations identify only partial nonlinear effects for CHF (significant reduction in survival probability at high workload levels, but no significant effect at low levels) and no significant effect, linear or nonlinear, for GIH and HIP.

Table 3 Selected coefficients of logit of survival odds by condition

	AMI	PNE	STR
$\min(\text{Workload}, \text{Node})$		1.353**	1.730***
		(0.422)	(0.367)
$\max(\text{Workload} - \text{Node}, 0)$		-14.53***	-4.821***
		(3.113)	(1.202)
Workload	0.500		
	(0.454)		
Severity of patient-mix	-0.241^{+}	-0.220	0.0227
	(0.133)	(0.164)	(0.120)
Admitted on weekends or public holidays	-0.177*	-0.0529	0.0309
	(0.0873)	(0.0691)	(0.0560)
Observations	14,096	14,278	21,400
Node		0.8843	0.8253

Standard errors adjusted for clustering within hospitals

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3 contains the results of the logit model for the three included conditions. Figure 3 plots corresponding predicted survival probabilities with 95% confidence intervals, calculated by the delta method (Wooldrige 2002), with workloads ranging

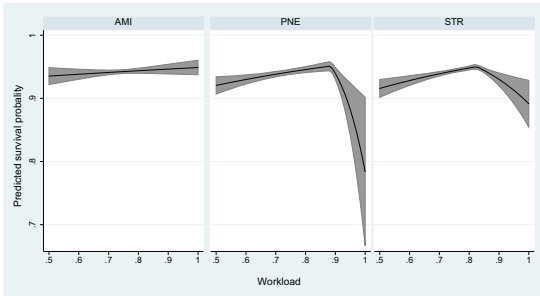

Figure 3 Predicted survival probabilities with 95% confidence intervals

from 50% to 100% and all other variables set to their means.

The results show a significant nonlinear effect, in line with our hypothesis, for pneumonia and stroke. It is somewhat surprising, at first glance, that we were not able to identify a significant workload effect, linear or nonlinear, for acute myocardial infarction (AMI), although the sample size is comparable to PNE and STR. To gain insight into the difference between these conditions, we interviewed doctors and nurses across hospitals and a chief medical officer of a UK strategic health authority. The interviewees were not surprised by the lack of a significant workload effect on AMI survival as, in their view, this was to be expected. Acute myocardial infarction (AMI) is a highly acute diagnosis and survival is mostly determined by correct clinical diagnosis and adequate, rapid treatment in the ambulance or emergency department, prior to ward admission (McNamara et al. 2006). Ward workload is therefore much less relevant than workload in the emergency department, which was not measured in our data. Additionally, our interviewees confirmed that ward treatment for AMI patients is fairly standardized and patients are monitored electronically so that a life-threatening deterioration in health quickly becomes obvious. In contrast, pneumonia and stroke patients require very intensive nursing care and deterioration in health may not be as obvious as in patients with AMI. In summary, ward workload can be expected to be a more relevant factor behind survival rates for pneumonia and stroke patients.

4.3. Effects of workload during phases of a patient episode

Differences in workload effects may occur across conditions, as illustrated above, but also across phases of a patient episode. Are effects more pronounced during certain phases of pneumonia or stroke episodes? Ideally, we would look to study daily workload effects in order to measure this; however, avoidable deaths will often be caused by an accumulation of suboptimal services on different days during the hospitalization and therefore significant individual day effects could only be identified with a substantially larger sample. Figure 4 illustrates the cumulative effect of workload over successive days. The estimates have been obtained by repeating the analysis in section 4.2 for workloads up to day d of a patient's stay. The curves correspond to workloads (i) on the day of admission, (ii) during the minimal period required to identify a linear spline with the model selection procedure outlined in section 3.4 (4 days for PNE and 6 days for STR), (iii) during the period up to the average length of stay for the condition in our data (11 days for PNE and 13 days for STR), and (iv) for the entire patient episode. You will notice that the survival optimal workload, once identified, is relatively stable as d changes.

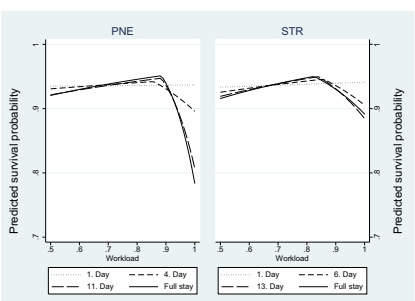

Figure 4 **Cumulative workload effect**

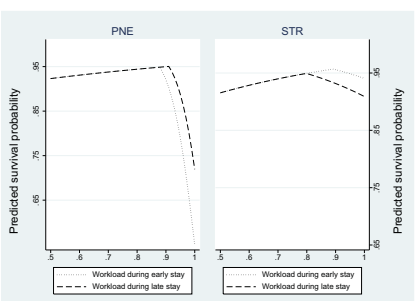

Figure 5 **Early versus late phase**

To verify whether workload during different phases of a patient episode has different effects on service outcomes, we split each patient episode into two phases of

equal duration and estimated the models with workload during the initial and final half of the patient stay. The results are shown in Figure 5. Interestingly, the gradients of the identified splines are similar for early phase and late phase workloads; however the tipping points after which quality deteriorates are different. In the case of pneumonia the first phase of a hospital episode appears to have a greater impact on survival probability, whilst the opposite is true for stroke patients.

5. Limitations of the empirical study

The first limitation of our study is the lack of direct information concerning the hospital departments' operational capacity. Our capacity metric is the maximum number of patients observed in the hospital department on any one day during the observation period. This measure is fixed for all patient episodes in the same department and relates to slow-changing capacity, such as the number of beds, number of doctors, or number of high-value medical devices. To check the robustness of our results we also computed episode workloads relative to an alternative, episode-specific measure of departmental capacity: the maximum number of patients in the department during the period from one month before the patient's admission to one month after discharge. This measure incorporates more flexible dimensions of capacity, such as access to nursing staff. The results for PNE and STR remain qualitatively unchanged and become quantitatively somewhat more pronounced. AMI now shows a nonlinear effect but the size is small, in keeping with our earlier argument that AMI survival is less affected by ward workload. In summary, whilst the quantitative effect of short-term capacity is different from that of long-term capacity, the qualitative results remain unchanged and also support the nonlinearity hypothesis.

A second limitation concerns potential demand endogeneity: workload is driven by demand, which may itself be affected by the quality of service outcomes. There are

several reasons why demand endogeneity is unlikely to be a significant confounding factor in our study. Firstly, mortality-related information was not easily available in Germany before 2005, the end of our observation period and therefore patients and referring doctors would not have had access to objective quality information to influence their hospital selection. Secondly, at least one of our conditions - stroke - is highly acute and hospital proximity is likely to be the prominent hospital selection criterion. Thirdly, there is some evidence in the literature that demand for hospital services is exogenous to survival probability, even for elective procedures. Using a 17-year panel, Gaynor et al. (2005) were unable to reject the hypothesis that volume is exogenous to survival in patients undergoing coronary artery bypass grafts. Finally, our data allows us to perform a simple exogeneity test: we calculated a standardized survival ratio for each department as the ratio of the observed number of survivals in the department to the predicted number of survivals of the department's patients across all departments in the sample, where the predicted number of survivals was calculated by estimating our basic model (1) without department fixed effects. The larger the department's ratio, the more favorable its survival rates compare to other hospital departments. We compared a department's standardized survival ratio with its demand growth over the observation period, using weekly admissions data. If demand were endogenous - that is, driven by survival rates - we would expect high standardized survival ratios to be associated with larger demand growth rates. However, the data for the 243 hospital departments in our study demonstrated no significant correlation between these two variables.

A third limitation of our study, in common with other studies of hospital outcomes, is the potential for omitted variable bias. For example, Schilling et al. (2010) refer in their study to staff skills, leadership and institutional aspects as possible omitted factors. Whilst we were unable to include actual departmental staffing levels over each patient's stay, the inclusion of department fixed effects should mitigate

this limitation somewhat in our study. Additionally, we do capture some variation in daily staffing levels with the weekend and public holiday dummy and seasonal dummy variables, as in Kc and Terwiesch (2009). Nevertheless, further research should address the effect of increased workload on actual staffing levels, and the resulting effect on service quality. This is particularly relevant for analyses at the level of a single hospital.

6. Hospital capacity planning

Our estimations led to a piecewise linear representation of service quality $Q(w)$ as a function of workload w during a service episode

$$Q(w) = \min\{\alpha_0 + \beta_0 w, \alpha_1 + \beta_1 w\}, \tag{2}$$

where $\beta_0 > 0$ and $\beta_1 < 0$. Optimal quality is delivered at workload level $w^* = \frac{\alpha_1 - \alpha_0}{\beta_0 - \beta_1}$ where the two linear pieces intersect. Table 4 shows relevant estimates of the econometric model (1) applied to stroke patients in two different hospitals in our data set - one hospital with 523 stroke patients in a dedicated stroke unit and a second hospital with 1,278 stroke patients across 5 departments. In both cases the observed

Table 4 Estimation results for stroke patients in two hospitals

	STR patients in stroke unit	STR patients across five departments
β_0	6.342*	7.134***
	(2.915)	(0.962)
β_1	-35.83**	-13.81*
	(13.00)	(6.172)
Quality-optimal workload w^*	0.843	0.822
Optimal survival probability $Q(w^*)$	0.981	0.967
	(0.006)	(0.005)
Average observed workload $\bar{w}$ across all patients	0.759	0.760
Observed survival rate across all patients	0.925	0.896
Observations	523	1278

Robust standard errors, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

average workload $\bar{w}$ is considerably lower than the quality-optimal workload w^*. It

is tempting to argue that the quality of service for stroke patients could be improve by cutting resources, as this would shift the average workload closer to the optimal level w^*.

In fact, practicing managers may argue that a change in resources to a fraction r of current levels would change workloads by the factor $\frac{1}{r}$ and therefore if $r = \frac{w^*}{\bar{w}}$ the average workload would shift to the desired level w^*. In the case of the two hospitals in Table 4 this would call for substantial cuts to 90% (stroke unit) and 92% (across departments) of current capacity. This argumentation, however, is seriously flawed on two counts: first, it ignores the effects of workload variability and second, it fails to take account of the effects of a change in resources on length of stay.

The effect of workload variability is conceptually similar to the classical newsvendor problem, but is complicated by the fact that acute hospitals cannot choose a fixed capacity constraint beyond which they will no longer admit patients. It would be near impossible to turn away patients in urgent need of care and as such, we have to accept the reality that the organization will serve demand beyond its planned capacity when the need arises. For this reason, it seems more appropriate to formulate the hospital capacity optimization problem in terms of choosing a resource vector R instead of choosing a fixed capacity limit beyond which demand is capped. The survival-optimization problem chooses resources R so as to maximize expected survival probability across patient episode workloads

$$\max_R \mathbf{E}_{F(R)}[Q(w)], \tag{3}$$

where $F(R)$ is the cumulative distribution function of episode workload when the organization has resources R. It is difficult to predict how workload distributions would alter with resource levels, even if demand remained unaffected. If we disregard the length of stay response to changing resources, the optimization problem (3) turns into a modified newsvendor problem (see appendix for details). The corresponding

survival-optimal resource levels for the two hospitals in Table 4, based on their empirical workload distributions, can be calculated as 98% (stroke unit) and 99% (across departments) of current levels, far from the earlier flawed advice to aim for 90% (stroke unit) and 92% (across departments) of current capacity. Therefore the effect of workload variation alone is sufficient to illustrate that average workload $\bar{w}$ may well be much lower than w^* at optimal resource levels.

The actual capacity optimization problem, however, is significantly more complex and needs to account for the additional length of stay response to resource changes. Specifically, congestion and professional discretion over service times exert counteracting effects. Figures 1 and 2 provide useful illustrations of this. When workload is low, only the congestion effect is active - clinicians are not yet under sufficient pressure to accelerate patient discharge. It is therefore plausible to assume that, as capacity is cut, low workloads increase more rapidly than high workloads because the counteracting effect of actively reduced service times is not yet being felt. At high workload levels, however, doctors may exercise their discretion and discharge patients earlier. Figures 1 and 2 provide some evidence that the combined effect of congestion and early discharge will still result in a reduction in length of stay at high workload levels. Therefore high workload levels can be expected to increase more gradually as capacity is cut. In summary, one can expect that capacity cuts lead to a shift in the workload distribution towards the right, as well as a compression of its shape. A comprehensive analysis of this effects is beyond the scope and data of the present study, however, and is left for future research.

7. Conclusions

The empirical results presented in this paper complement and refine recent research in the medical literature which argues that organizational workload can cause quality issues. Schilling et al. (2010), using the same medical conditions as in our

study, emphasize the effect of overcrowding in hospitals and emergency departments. Weissman et al. (2007) conclude that "*hospitals that operate at or near capacity [...] might consider re-engineering their structures of care to respond better during periods of high stress*" (p. 454). These studies are both based on aggregate samples across clinical conditions. In a more focused study of a cardiothoracic surgery unit, Kc and Terwiesch (2009) were unable to detect a significant impact of workload on in-hospital mortality. Our results, based on a sample across conditions and hospitals, suggest that the relationship between workload and in-hospital mortality is best understood as a nonlinear phenomenon. Linear models, as used in earlier studies, tend to underestimate the magnitude of outcome deterioration at very high workload levels. We also demonstrate that the impact of workload on mortality can be quite different for differing medical conditions.

We have highlighted the implications of these findings for quality-led hospital capacity planning. Since service quality is a nonlinear function of workload, quality cannot be optimized on the basis of average workload alone. As in the newsvendor problem, distributional characteristics do matter. Specifically, managers need to develop an understanding of how capacity changes affect workload distributions, and the crucial role of length of stay response, driven by the relationship between congestion effects and professional discretion over service completion.

The insights gained through this empirical study in a hospital context have implications for a wider spectrum of professional service organizations. We have argued conceptually and empirically that the effect of organizational workload on professional service outcomes is likely to be nonlinear: when organizational workload is already high, outcomes will deteriorate as workload is further increased. However, when workload is low or moderate outcomes do not deteriorate and may well improve as workload increases. The nonlinear phenomenon is conceptually explained by combining three perspectives, two of which - the effects of congestion and of

professional discretion over service provision - are well established in the literature. A new endocrinological perspective sheds light on the effect of workload on error rates, through the subconscious impact of stress hormones on a worker's cognitive performance. When workload is low, an increase in workload triggers a positive stress response, resulting in increased vigilance and improved individual and organizational performance. A worker's stress response at low to moderate workload levels acts as a variability buffer with respect to service outcome. However, the stress buffer clearly has its limits. When workload becomes too high and stress hormone levels exceed certain thresholds, a worker's cognitive performance begins to deteriorate and she becomes more error-prone. Furthermore, at high workload levels autonomous professionals will begin to take conscious decisions to improve throughput, using quality as a variability buffer, whilst congestion effects lead to significant waiting times with further detrimental impact on service quality. As a consequence of these mutually reinforcing effects, service quality can 'fall off a cliff' when workload exceeds a quality-tipping point.

This study has raised at least two potential questions for future research. First, how does the relationship between congestion and professional discretion over service completion affect workload distributions when resource levels are changed? And second, what factors affect the quality-tipping point and the extent of deterioration that follows when this point is passed? Finding answers to these questions will assist operations managers in hospitals, as well as other professional service organizations, in their attempt to drive efficiency whilst optimizing service quality.

References

Agency for Healthcare Research and Quality. 2006. Inpatient quality indicators. URL `http://www.qualityindicators.ahrq.gov/downloads/iqi/2006-Feb-InpatientQualityIndicators.pdf`.

American Hospital Association. 2011a. The cost of caring. URL `www.aha.org/aha/content/2008/pdf/08affordability-cost.pdf`.

American Hospital Association. 2011b. Hospitals continue to feel lingering effects of the economic recession. URL `http://www.aha.org/aha/content/2010/pdf/10june-econimpact.pdf`.

Anand, K.S., M.F. Pac, S. Veeraraghavan. 2011. Quality speed conundrum: trade-offs in customer-intensive services. *Management Science* **57**(1) 40–56.

Archibald et al. 1997. Patient density, nurse-to-patient ratio and nosocomial infection risk in a pediatric cardiac intensive care unit. *Pediatr Infect Dis J.* **16**(11) 1045–8.

Ata, B., J.A. van Mieghem. 2009. The value of partial resource pooling: should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

Backe et al. 2009. Assessment of salivary cortisol as stress marker in ambulance service personnel: comparison between shifts working on mobile intensive care unit and patient transport ambulance. *International Archives of Occupational and Environmental Health* **82** 1057–1064.

Bell, C.M., D.A. Redelmeier. 2001. Mortality among patients admitted to hospitals on weekends as compared with weekdays. *New England Journal of Medicine* **345**(9) 663–668.

Charlson, M.E., P. Pompei, K.L. Ales, C.R. MacKenzie. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**(5) 373–383.

Dahl, M.S. 2011. Organizational change and employee stress. *Management Science* **57**(2) 240–256.

de Kloet, E.R., M.S. Oitzl, M. Joels. 1999. Stress and cognition: are corticosteroids good or bad guys? *Trends in Neuroscience* **22**(10) 422–426.

Dickerson, S.S., M.E. Kemeny. 2004. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological Bulletin* **130**(3) 355–391.

Edmondson, A.C., A.L. Tucker. 2001. Why hospitals don't learn from failures: organizational and psychological dynamics that inhibit system change. *California Management Review* **45**(7) 55–72.

Elixhauser, A., C. Steiner, D.R. Harris, R.M. Coffey. 1998. Comorbidity measures for use with administrative data. *Medical Care* **36**(1) 8–27.

Fischer and et al. 2000. Experience and endocrine stress responses in neonatal and pediatric critical care nurses and physicians. *Critical Care Medicine* **28**(9) 3281–3288.

Gaynor, M., H. Seider, W.B. Vogt. 2005. The volume-outcome effect, scale economies, and learning-by-doing. *American Economic Review* **95**(2) 243–247.

Hacke et al. 2004. Association of outcome with early stroke treatment: pooled analysis of atlantis, ecass, and ninds rt-pa stroke trials. *Lancet* **363**(6) 768–774.

Halm, E.A., C.Lee, M.R. Chassin. 2002. Is volume related to outcome in health care? a systematic review and methodologic critique of the literature. *Annals of Internal Medicine* **137**(6) 511–521.

Hopp, W.J., S.M.R. Iravani, G.Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.

Huckman, R., G. Pisano. 2006. The firm specificity of individual performance: evidence from cardiac surgery. *Management Science* **52**(4) 473–488.

Hugonnet, S., J.C. Chevrolet, D. Pittet. 2007. The effect of workload on infection risk in critically ill patients. *Critical Care Medicine* **35**(1) 76–81.

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

Kc, D.S., C. Terwiesch. 2010. An econometric analysis of patient flows in the cardiac icu. *unpublished manuscript* .

Long, J., J. Freese. 2006. *Regression models for categorical dependent variables using STATA*. STATA Press.

Lundberg, U., M. Frankenhaeuser. 1999. Stress and workload of men and women in highranking positions. *Journal of Occupational Health Psychology* **4** 142–151.

Lupien, S.J., C.J. Gillin, R.L. Hauger. 1999. Working memory is more sensitive than declarative memory to the acute effects of corticosteroids: a dose-response study in humans. *Behav. Neurosc.* **113**(3) 420–430.

Lupien et al. 2007. The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition. *Brain and Cognition* **65** 209–237.

Marsh, L., D.R. Comier. 2001. *Spline regression models*. Quantitative applications in the social sciences, Sage Publications, Thousand Oaks.

McClellan, M., H. Noguchi. 1998. Technological change in heart disease treatment: does high tech mean low value? *American Economic Review* **88**(2) 90–96.

McEwen, B.S. 2002. *The end of stress as we know it*. National Academies Press.

McEwen, B.S., J.M. Weiss, L.S. Schwartz. 1968. Selective retention of corticosterone by limbic structures in rat brain. *Nature* **220** 911–912.

McNamara et al. 2006. Effect of door-to-balloon time on mortality in patients with st-segment elevation myocardial infarction. *Journal of the American College of Cardiology* **47**(11) 2180 – 2186.

Peduzzi et al. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* **49**(12) 1373–1379.

Pisano, P., R.M.J. Bohmer, A.C. Edmondson. 2001. Organizational differences in rates of learning: evidence from the adoption of minimally invasive cardiac surgery. *Management Science* **47**(6) 752–768.

Quan et al. 2005. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care* **43**(11) 1130–1139.

Ramdas, K., J. Williams. 2009. An empirical investigation into the tradeoffs that impact on-time performance in the airline industry. *Working Paper London Business School* .

Rosanio et al. 1999. Queuing for coronary angiography during severe supply-demand mismatch in a us public hospital: analysis of a waiting list registry. *Journal of the American Medical Association* **282**(2).

Royston, P., W. Sauerbrei. 2007. Multivariable modeling with cubic regression splines: a principled approach. *STATA Journal* **7**(1) 45–70.

Schilling, P.L., D.A. Campbell Jr., M.J. Englesbe, M. M. Davis. 2010. A comparison of in-hospital mortality risk conferred by high hospital occupancy, differences in nurse staffing levels, weekend admission, and seasonal influenza. *Medical Care* **48**(3) 224–232.

Selye, H. 1936. A syndrome produced by diverse nocuous agents. *Nature* **138** 32.

Shortell, S. M., E. F. Hughes. 1988. The effects of regulation, competition, and ownership on mortality rates among hospital inpatients. *New England Journal of Medicine* **318**(17) 1100–1107.

Sonnentag, S., C. Fritz. 2006. Endocrinological processes associated with job stress: catecholamine and cortisol responses to acute and chronic stressors. P Perrewé, D Ganster, eds., *Employee Health, Coping and Methodologies*. Emerald, 1–59.

Torres, A., E. Santiago. 2004. Diagnosing ventilator-associated pneumonia. *New England Journal of Medicine* **350**(5) 433–435.

Wang, X., L.G. Debo, A. Scheller-Wolf, S.F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Science* **56**(11) 1873–1890.

Weissman et al. 2007. Hospital workload and adverse events. *Medical Care* **45**(5) 448–455.

Wooldrige, J.M. 2002. *Econmetric analysis of cross section and panel data*. MIT Press, Cambridge, Massachusetts.

Zeier, H., P. Brauchli, H.I. Joller-Jemelka. 1996. Effects of work demands on immunoglobulin a and cortisol in air traffic controllers. *Biological Psychology* **42** 413–423.

Appendix. Modified newsvendor solution

We assume a hospital department is currently run with a planned maximal daily capacity of C patients. Workload is measured relative to planned capacity: an average daily patient volume n in the department during a patient's stay leads to workload $w = \frac{n}{C}$. Had the planned capacity been rC, the workload would have been $\frac{w}{r}$. Here we assume, crucially, that length of stay is unaffected by changes in planned capacity. With this assumption, the optimization problem (3) is therefore simplified to

$$\max_{r>0} \mathbf{E}[Q(\frac{w}{r})], \tag{4}$$

where the expectation is taken over the distribution F of workloads w at current capacity C.

PROPOSITION 2. *Suppose (2) describes the relationship between workload and service quality in a department. Let F be the cumulative distribution function of patient episode workload W at current capacity and suppose a change in capacity by a factor r changes workload to $\frac{W}{r}$. Then the optimal solution of (4) is $r^* = \frac{s^*}{w^*}$, where $s^* \in (0,1)$ is a root of the monotone function $\beta_1 \mathbf{E}[W] + (\beta_0 - \beta_1)\mathbf{E}[I_{[0,s]}W]$. The root is unique if F is strictly monotone at s^*.*

Proof. We may assume w.l.o.g. that F has the support $[0,1]$. Since $s = w^* r$ the maximum of $Q(\frac{w}{r})$ is achieved at $w = s$ and we obtain

$$\begin{aligned}
\mathbf{E}[Q(\frac{w}{r})] &= \int_0^1 Q(\frac{w^* w}{s})dF(w) = \int_0^s (\alpha_0 + \beta_0 \frac{w^* w}{s})dF(w) + \int_s^1 (\alpha_1 + \beta_1 \frac{w^* w}{s})dF(w) \\
&= \int_0^s (\alpha_0 + \beta_0 w^*(1 + \frac{w-s}{s}))dF(w) + \int_s^1 (\alpha_1 + \beta_1 w^*(1 + \frac{w-s}{s}))dF(w) \\
&= \int_0^s (\alpha_0 + \beta_0 w^*)dF(w) + \frac{\beta_0 w^*}{s}\int_0^s (w-s)dF(w) + \int_s^1 (\alpha_1 + \beta_1 w^*)dF(w) + \frac{\beta_1 w^*}{s}\int_s^1 (w-s)dF(w) \\
&= Q(w^*) + \frac{w^*}{s} g(s).
\end{aligned}$$

Here we have used the fact that $\alpha_0 + \beta_0 w^* = \alpha_1 + \beta_1 w^* = Q(w^*)$, due to the definition of w^*. The function $g(s)$ is of the form

$$\begin{aligned}
g(s) &= \beta_0 \int_0^s (w-s)dF(w) + \beta_1 \int_s^1 (w-s)dF(w) \\
&= \beta_1 \int_0^1 (w-s)dF(w) + (\beta_0 - \beta_1)\int_0^s (w-s)dF(w) \\
&= \beta_1(\mathbf{E}[W] - s) + (\beta_0 - \beta_1)(\mathbf{E}[I_{[0,s]}W] - sF(s)).
\end{aligned}$$

Solving (4) is equivalent to maximizing $\frac{g(s)}{s}$. Integration by parts implies $\int_0^s w dF(w) - F(s)s = -\int_0^s F(w)dw$. Therefore $g(s) = \beta_1(\mathbf{E}[W] - s) - (\beta_0 - \beta_1)\int_0^s F(w)dw$ and $g'(s) = -\beta_1 - (\beta_0 - \beta_1)F(s)$. The function $\frac{g(s)}{s}$ has

a maximizer in the interval $(0,1)$ because its derivative has the same sign as $g'(s)s - g(s) = -\beta_1 \mathbf{E}[W] - (\beta_0 - \beta_1)\mathbf{E}[I_{[0,s]}W]$ and the latter function is monotonically decreasing in s with values $-\beta_1 \mathbf{E}[W] > 0$ at $s = 0$ and $-\beta_0 \mathbf{E}[W] < 0$ at $s = 1$. □

Additional statistics of the data

Table EC.1 Descriptive statistics for both models of risk-adjustment

	Mean	Standard Deviation
Elixhauser comorbidities		
Congestive heart failure	0.217	0.412
Cardiac arrhythmias	0.239	0.426
Valvular disease	0.072	0.259
Pulmonary circulation disorders	0.020	0.141
Peripheral vascular disorders	0.062	0.242
Hypertension, uncomplicated	0.426	0.495
Hypertension, complicated	0.074	0.263
Paralysis	0.176	0.380
Other neurological disorders	0.130	0.336
Chronic pulmonary disease	0.100	0.301
Diabetes, uncomplicated	0.152	0.359
Diabetes, complicated	0.097	0.295
Hypothyroidism	0.030	0.171
Renal failure	0.114	0.318
Liver disease	0.032	0.175
Peptic ulcer disease excluding bleeding	0.006	0.080
AIDS/HIV	0.001	0.029
Lymphoma	0.006	0.075
Metastatic cancer	0.013	0.114
Solid tumor without metastasis	0.026	0.160
Rheumatoid arthritis/collagen, vascular diseases	0.013	0.112
Coagulopathy	0.031	0.173
Obesity	0.088	0.284
Weight loss	0.024	0.153
Fluid and electrolyte disorders	0.156	0.363
Blood loss anemia	0.012	0.108
Deficiency anemia	0.017	0.129
Alcohol abuse	0.039	0.194
Drug abuse	0.005	0.069
Psychoses	0.006	0.076
Depression	0.037	0.189
Charlson-Index dummies		
Charlson-Index = 0	0.274	0.446
Charlson-Index = 1 od. 2	0.412	0.492
Charlson-Index = 3 od. 4	0.224	0.417
Charlson-Index $\geq$ 5	0.089	0.285
N	75314	

Table EC.2 All measures of workload and case complexity

	Mean	Standard Deviation
Workload measures		
Workload	0.742	0.117
Workload at admission	0.746	0.132
Time variant workload	0.791	0.102
Time variant workload at admission	0.796	0.122
Case complexity measures		
Staff case weight	1.001	0.948
Total case weight	1.888	1.693
Average Charlson Points	1.738	0.631
N	75314	

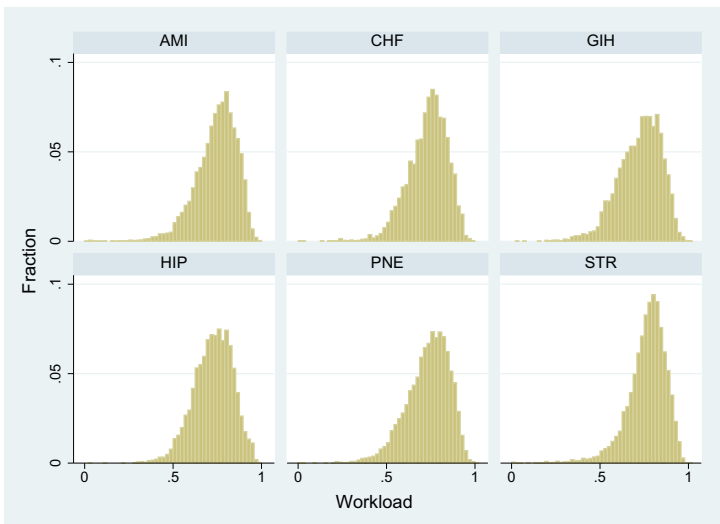

Figure EC.1 Histogram of occupancy during patient's stay by main condition