Stress on the Ward: Evidence of Safety Tipping Points in Hospitals

Ludwig Kuntz
Faculty of Management, Economics and Social Sciences, University of Cologne kuntz@wiso.uni-koeln.de

Roman Mennicken
Rheinisch-Westfälisches Institut für Wirtschaftsforschung roman.mennicken@rwi-essen.de

Stefan Scholtes
Judge Business School, University of Cambridge s.scholtes@jbs.cam.ac.uk

Do hospitals experience safety tipping points as utilization increases and, if so, what are the implications for hospital operations management? We argue that safety tipping points occur when managerial escalation policies are exhausted and workload variability buffers are depleted. Front-line clinical staff is forced to ration resources and, at the same time, becomes more error-prone as a result of elevated stress hormone levels. We confirm the existence of safety tipping points for in-hospital mortality using discharge records of 82,280 patients across six high-mortality-risk conditions from 256 clinical departments of 83 German hospitals. Focusing on survival during the first seven days after admission, we estimate a mortality tipping point at an occupancy level of 92.5%. Among the 14.7% of patients in our sample who experienced occupancy above the tipping point during the first seven days of their hospital stay, high occupancy accounted for 17% of deaths. The existence of a safety tipping point has important implications for hospital management. First, flexible capacity is likely to be more cost-effective for safety improvement than rigid capacity as it will only be used when occupancy reaches the tipping point. Within our sample, flexible staffing saves 40% of the cost of a fully staffed capacity expansion, while achieving the same reduction in mortality. Second, reducing the variability of demand by pooling capacity in hospital clusters can greatly increase safety in a hospital system because it reduces the likelihood that a patient experiences occupancy levels beyond the tipping point. Pooling the capacity of nearby hospitals in our sample reduces the number of deaths due to high occupancy by 39.5%.

Key words: Hospital mortality; occupancy; utilization; variability buffers; stress; service quality

1. Introduction

Avoidable deaths occur in hospitals as the direct or indirect consequence of avoidable adverse events, such as medication errors, infections, delayed treatments, or technical complications during operations. The scale of the problem was prominently highlighted by the influential Harvard Medical Practice Study in the early 1990s (Brennan et al. 1991, Leape et al. 1991), which estimated that 6,895 of the 2,671,863 patients hospitalized in New York State in 1984 had died in hospital as a consequence of preventable adverse events – more than three times the New York State traffic death toll in the same year (US Dept. of Transportation 2012). This study ignited a major global effort to improve hospital safety, much of which was focused on the prevention of individual human errors through process redesign and the use of technology. For example, Bates et al. (1999) report an 81% reduction in medication errors through the introduction of computerized physician order entry systems. Leape et al. (1999) and Kucukarslan et al. (2003) report that the inclusion of pharmacists in ward rounds on ICUs and general medicine wards led to a 66% and 78% respective reduction in preventable adverse drug reactions. Comprehensive surgical safety checklists reduced the total number of surgical complications in a Dutch hospital from 27.3% to 16.7% (Eefje et al. 2010). Despite such impressive improvements, advances appear to be isolated and system-wide progress in reducing avoidable death rates remains slow (Leape and Berwick 2005, Landrigan et al. 2010).
The mortality effects of interventions at the level of a hospital or a hospital system are less well understood than interventions at the process level, despite the fact that the former are less dependent on the specific organizational context and therefore more easily scalable and more likely to help achieve the desired system-wide progress. An example in point is hospital capacity pooling, where nearby hospitals agree to short-term admission diversion or staff relocation protocols in response to local occupancy surges. While pooling does not affect average hospital utilization in the system, it reduces the variability in utilization levels of individual hospitals. Does capacity pooling reduce avoidable death rates in the system? The answer to this question depends on the nature of the relationship between occupancy levels and mortality. If this relationship is linear, then avoidable deaths in the system are not affected by pooling: Whatever is won by avoiding further increased utilization in a busy hospital is lost by increasing utilization and thereby mortality in the hospital that admits the diverted patients. However, if the relationship is nonlinear, capacity pooling can have a significant effect on avoidable deaths. We will argue in this paper that bed occupancy has a highly nonlinear effect on mortality: Mortality remains unaffected by occupancy up to a tipping point, beyond which it deteriorates rapidly with further increased occupancy levels. As a consequence, capacity pooling has the potential to reduce avoidable death rates across a health system as patient diversion reduces the propensity of a hospital to exceed the safety tipping point, while the mortality in a less busy diversion hospital is not affected as long as its occupancy level remains below the tipping point.

In order to provide evidence for a safety tipping point, we have to overcome three methodological challenges. First, avoidable in-hospital deaths are rare events. One has to combine data from multiple hospitals to assemble a sufficiently large and sufficiently homogeneous sample of patients to obtain the requisite statistical power. Second, the usually considered aggregate hospital occupancy levels are unlikely to be sufficiently closely linked with the operational decisions that affect individual patient care and which, as we will argue, are the reason for the existence of the tipping point. These decisions are taken at departmental level in a hospital. Hospital departments, as managerial units, are not standardized and not recorded as part of US or UK administrative patient records, which form the basis for most empirical studies of hospital operations. Departments are, however, recorded in a standardized way in German discharge records. In response to these two challenges, we have assembled a large multi-hospital data set from 83 acute hospitals in Germany.

The third methodological challenge is of a statistical nature: When occupancy is high, doctors may choose to discharge relatively healthy patients earlier than they would normally do to make space for newly arriving patients. They may well select patients for early discharge on health-related factors that are not recorded in the discharge record and thus not observable to the researcher. As a consequence, the patient pool in the hospital during periods of high occupancy is sicker in an unobserved way and these patients are therefore more likely to die - not as a consequence of high occupancy per se but as a consequence of early discharge decisions in response to high occupancy. We account for this endogeneity effect in our econometric models.

Using a sample of 82,280 patients with high mortality risk, we estimate a tipping point at an occupancy level of 92.5%; 14.7% of these patients experienced occupancy levels above the tipping point. Of the 4,247 deaths in our sample of 82,280 patients, this results in an estimated 78 deaths due to high occupancy, accounting for 1 in 55 deaths in the sample. However, 85.3% of patients in the sample never experienced occupancy above the tipping point. The effect for those patients who did is therefore significantly larger: occupancy accounts for 17% of deaths among these patients. This is clinically highly significant. We discuss potential interventions, specifically the effect of capacity increase, and the associated value of staffing flexibility, and of capacity pooling. Flexibly staffed capacity turns out to be 40% cheaper than rigid staffing for a commensurate mortality improvement. To estimate the effect of capacity pooling, we combine near-by hospitals in our sample to hospital clusters and estimate that 39.5% of the deaths due to occupancy in our sample could have been avoided if near-by hospitals had pooled their capacity.
The paper is organized as follows: After a review of the related literature we present the theoretical arguments for the existence of a safety tipping point. We then explain the context of our empirical study and the data, followed by a more detailed explanation of the econometric models with which we address the afore-mentioned methodological challenges. We then present the empirical results and discuss the implications of the safety tipping point for the use of flexible capacity and capacity pooling.

2. Related Literature

The nature of the association between system utilization and service quality in general, and hospital occupancy and mortality in particular is not yet well understood and the literature has hitherto produced inconsistent results. In an early paper, Oliva and Sterman (2001) built a dynamic model of a service organization to illustrate the complex interactions between service demand and managerial response and how these can lead to an erosion of service quality over time. More recently, a series of empirical studies have focused on operational efficiency and throughput in the context of hospitals. These studies acknowledge that a focus on throughput alone can have harmful consequences for clinical quality and therefore recommend to also investigate effects on clinical outcomes such as mortality rates (KC and Terwiesch 2009, 2012, Long and Mathews 2013, Berry Jaeker and Tucker 2012, Kim et al. 2013). KC and Terwiesch (2009), for example, complemented a throughput analysis of a sample of cardiothoracic patient records from a US hospital with a study of the effect of workload on mortality and found a significant effect of fatigue but could not identify a significant effect of bed occupancy on mortality. In contrast to these throughput studies, the present paper focuses fully on the relationship between occupancy and mortality and its managerial implications. Providing a separate, more detailed study on mortality effects is important as good clinical outcomes are the ultimate goal of hospitals and efficiency studies therefore need to take account of potential outcome effects in an appropriate way.

The focus of this paper is on system utilization as a cause of quality variation and complements recent studies that consider workload effects at the level of individual workers. Powell et al. (2012) show that doctors’ discharge coding behavior is affected by workload with detrimental effect for hospital reimbursement. Green et al. (2012) show that absenteeism rates are correlated with anticipated future nurse workload. Drawing on the theory of stress, Tan and Netessine (2012) show that workload has a curvilinear effect on waiters’ performance and illustrate that a reduction of staffing can in fact lead to an increase in revenues. In the same vein, Hopp et al. (2007) illustrate within a queuing model that when a server has discretion over service time in response to workload, increasing the number of servers may worsen congestion. We will incorporate these insights into the development of the tipping point hypothesis, which integrates the effects of excess capacity, managerial actions and individual worker responses to explain the organization-level effect of variation in system utilization on service quality.

Several recent studies in the medical literature have identified a link between hospital activity levels and mortality (see Kane et al. (2007) for a review). Schilling et al. (2010) explored the effects of hospital occupancy levels on admission, annual nurse staffing levels and seasonal factors on hospital mortality in a retrospective study of 166,920 emergency patients with high-risk conditions admitted to 39 Michigan hospitals between 2003 and 2006. The study found that admission on days when the hospital is in the top tertile of its occupancy range is associated with an elevated mortality risk. Needleman et al. (2011) studied the effect of nurse staffing levels below targets using 197,861 patient records from 43 clinical units of a US medical center and concluded that registered nurse staffing below target levels is associated with increased mortality. These studies model workload with a dichotomous ”high-low” variable, which does not rule out a linear relationship between occupancy levels and mortality. Our study goes beyond these papers in that we estimate a
continuous nonlinear occupancy model which supports the existence of a safety tipping point and discuss the managerial implications of the existence of a safety tipping point.

In summary, the main contribution of this paper is to point out that there are good operational reasons to expect the effect of occupancy on mortality to exhibit a threshold phenomenon: Occupancy has no discernable effect on mortality up to a tipping point, beyond which it affects mortality significantly. By not taking this phenomenon into account prior studies either could not detect an effect or overestimated effects at low occupancy levels and underestimated the severity of very high utilization. The tipping point effect has important implications for hospital management and planning. We discuss the cost-safety tradeoff of two managerial responses that can save lives: bed capacity increases at hospitals, both flexible and inflexible, and capacity pooling with neighboring hospitals.

3. The Tipping Point Hypothesis

In most countries hospital planning focuses on bed capacity as the primary metric for sizing hospital departments; requisite staffing and other resources are largely calculated on a per-bed basis, using ratios that depend on departmental characteristics (Rechel et al. 2010). Consequently, as the volume of patients in a department approaches full bed capacity, workload pressure builds up across the unit as all resources become stretched. Capacity utilization – measured as the percentage of beds occupied – is therefore a useful aggregate measure of workload pressure in hospitals.

Occupancy levels in acute hospitals show significant variation as demand for urgent care is unpredictable. Such variation is managed by drawing on variability buffers. The first and most obvious buffer is built-in excess capacity: Hospital plans are typically based on average bed occupancy levels in the order of 85%–90%, thus providing a capacity buffer for demand peaks (Green 2004). A second class of buffers relates to managerial actions when occupancy levels rise. Managers can ask staff to work overtime, deploy flexible staff from elsewhere in the hospital or hire temporary staff from nursing banks or medical locum agencies. In 2004–2005 UK hospitals spent 9.4% of their nursing budget on temporary nursing staff (Department of Health (UK) 2006). In addition, hospitals can manage demand by canceling scheduled elective cases at short notice in response to unexpected surges in emergency admissions. UK hospitals, for example, cancel between 0.7% and 1% of elective patient admissions at the last minute for reasons unrelated to the circumstances of the scheduled patient (Department of Health (UK) 2012). The third class of buffers relates to responses by front-line staff. First, doctors and nurses are willing to work harder and for longer in times of crises (Scott et al. 2006). In fact, many nurses choose their profession based on an intrinsic motivation to care for people in need. Doctors undergo a substantial socialization process during their long professional training (Laine and Davidoff 1996). The values and professional norms of healthcare workers instil a strong motivation and willingness to “go the extra mile”, which provides an important human variability buffer. Second, health care professionals are likely to ration access to care as system utilization increases and, in doing so, will give priority to sicker patients. KC and Terwiesch (2012) and Long and Mathews (2013) provide evidence of active rationing from busy intensive care units (see also Berk and Moinzadeh (1998) and Padma et al. (2004)). While rationing can have a negative effect on the less sick the ability to prioritize is an important variability buffer, helping to shelter the most critically ill. All these variability buffers are drawn on simultaneously as utilization levels increase and allow the organization to cope with a wide variation in occupancy levels while safeguarding the most critical aspect of clinical care: the avoidance of death.

However, as occupancy levels continue to rise the organization’s variability buffers become depleted; all beds are filled and additional patients need to “board” in other departments, no more elective patients can be canceled at short-notice and qualified agency staff are scarce or resources to hire them are limited. Yet demand for hospital care is at times unrelenting. Acute care hospitals cannot turn emergency patients away. When the variability buffers are depleted, resources need
to be rationed more aggressively; doctors and nurses begin to cut corners even for more seriously ill patients, using service quality as an implicit variability buffer (Oliva and Sterman 2001, Hopp et al. 2007).

In addition to cutting corners as a conscious response to excessive workload, doctors and nurses are exposed to workload-related stress, which causes their performance to deteriorate. Lazarus and Folkman (1984) point out that stress results from an “imbalance between demands and resources” and occurs when “pressure exceeds one’s perceived ability to cope”; this is precisely the case when workload becomes excessive and the ability to cope by exploiting buffers reaches its limits. This effect was pointed out by Piquette and Reeves (2009) who observed that, in the context of critical care, “individual distress occurred in unexpectedly high demands unmatched by appropriate resources.” At the biological level, workload stress leads to elevated stress hormone levels, specifically those of cortisol (Dickerson and Kemeny 2004, Sonnentag and Fritz 2006), which impairs the workers’ cognitive abilities, especially memory and attention, and the quality of their decision-making (Lupien et al. 2007). In addition, team work deteriorates: Piquette and Reeves (2009) observed that “emotional distress was strongly contagious to other team members, […] disruptive for team work and deleterious for individual and collective performance”. The consequent negative impact of stress on clinical outcomes is well documented in the medical literature. Dugan et al. (1996) show for example that nursing-related stress is strongly associated with the propensity of adverse incidents; Buckley et al. (1997) show that haste and stress were causative factors in 17% of 281 critical incidents.

In summary, as system utilization increases to moderately high levels, managers respond by exploiting resource buffers and well-motivated employees work harder. Quality of care can largely be maintained and safety is not negatively affected. However, at very high utilization levels, variability buffers are depleted and managerial response is inhibited. If utilization exceeds this critical tipping point, managers are unable to respond. The pressure is passed on to front-line staff, who are unable to escape it. They then respond in two ways: First, by consciously cutting corners, using quality as an implicit variability buffer; and second, by subconsciously committing more errors as a result of elevated stress hormone levels. As a consequence, quality of care and safety will deteriorate during periods of high utilization. The following empirical study will provide evidence for this tipping point phenomenon.

4. Empirical Study

4.1. The Sample Size Challenge

We conduct a patient-level empirical study of the association between occupancy levels at a hospital department during a patient’s stay and her probability of dying in hospital. Before we introduce the data and econometric model we will briefly discuss the magnitude of the statistical challenge, specifically the required sample size, as this informs our choice of data and models. The primary challenge is that any effect of high occupancy on mortality will be very small relative to biological and other clinical causes of death. Specifically, high occupancy will only have an effect on avoidable mortality, not on mortality per se. Estimates of avoidable deaths range widely in the medical literature, depending on the context; Gruen et al. (2006) studied a sample of trauma deaths and estimated that 2.64% were avoidable; Healey et al. (2002) investigated deaths in several surgical departments and found that between 19% and 44.1% were avoidable. Only a fraction of the observed deaths are avoidable and high occupancy will be a causal factor for only a fraction of these. Therefore, the event of interest – avoidable death due to high occupancy – is rare and statistical power is a concern.
A sample size estimation reveals the magnitude of the challenge. The simplest specification is a logistic regression model of the form \( \logit(Y_i) = \alpha + \beta X_i + \gamma Z_i \), where \( Y_i \) is a dichotomous variable indicating the death of patient \( i \), \( X_i \) is the occupancy covariate and \( Z_i \) is a vector of control covariates. For ease of interpretation, we assume that \( X_i \) is a dichotomous variable with value 1 if patient \( i \) experienced occupancy levels above a threshold beyond which we believe mortality will be affected. The logistic sample size formula of Hsieh et al. (1998) can then be used for a power analysis for a significant two-tailed Wald test for the null hypothesis \( H_0 : \beta = 0 \). Figure 1 shows the required sample sizes for 80% power at a 5% significance level as a function of the unknown effect size, i.e. the excess mortality of those patients who experience occupancy levels above the tipping point when the unconditional population mortality is 10%. If mortality for the subpopulation above the tipping point rises from 10% to 10.5% (.005 on the x-axis of Figure 1), i.e. there are 5 additional deaths per 1,000 admissions when occupancy levels are above the tipping point, and if 15% of patients experience such high occupancy levels, then the required sample size exceeds 150,000. The required sample size decreases if more patients are exposed to occupancy levels above the tipping point or if the effect size is larger; if 30% of the patient population experience occupancy above the tipping point and the mortality rate for these patients rises to 11% against the population average of 10% (.01 on the x-axis of Figure 1) then the required sample size is approximately 20,000. However, these sample size estimations are optimistic as they assume that the occupancy variable \( X_i \) and the controls \( Z_i \) are uncorrelated. The sample size will have to be even larger if this is not the case (Hsieh et al. 1998). Also, the required sample size increases if the patient population of interest has a lower mortality risk. In conclusion, large samples of high-risk patients are needed for mortality studies of the type we conduct in this paper. As only a relatively small proportion of any single hospital’s patients will have a sufficiently high mortality risk, such studies require data from multiple hospitals as in our sample.

Figure 1  Required sample size for 5% significance and 80% power when population mortality is 10%

4.2. Data from German Hospital Departments
Earlier multi-hospital mortality studies have concentrated on the effect of aggregate hospital occupancy (e.g. Schilling et al. (2010)). In contrast, we measure daily occupancy at the level of departments within hospitals. This is important because managers and clinicians are most likely to respond to occupancy levels in their department. Therefore aggregate hospital occupancy measures the system load that is relevant for a particular patient with significant error, which leads to attenuation bias and underestimated effect sizes (Wooldridge 2002).

We use data from German hospitals, which are particularly suited for a multi-hospital department-level analysis. First, in contrast to US and UK discharge records, German records contain standardized department codes, including departmental referrals during a hospital episode.
Second, the organizational structure of German hospitals is firmly regulated, leading to a rigid and fairly homogeneous departmental organization across hospitals. The structural similarity begins at the top: Almost all German hospitals have the same top management team structure, consisting of a commercial director, a medical director and a nursing director, each with well-defined roles and obligations across the hospital. The departmental structure below the top team is also standardized, including general services, such as kitchen and laundry, large diagnostic divisions, such as radiology and pathology, and the clinical departments, including general surgery and general medicine, as well as specialist departments. These bed-bearing clinical departments are the focus of our study and system utilization is measured at the level of these organizational units. Importantly, every department has a clinical director – the Chefarzt – who, as lead physician, has the ultimate clinical responsibility for all patients and is the superior of all doctors in the department. The clinical director also has budgetary responsibility for her department. Although there is a cautious trend toward the use of interdisciplinary beds, the system remains rigid and the vast majority of patients and resources are managed at the level of these clinical departments. In particular, any responses to occupancy variations are most likely to be managed at departmental rather than hospital level.

Our initial database consists of standardized discharge records of 101 German hospitals, covering an observation period of one year – either 2004 or 2005 – for 72 hospitals and of the two years 2004-2005 for the remaining 29 hospitals. The database contains the records of all patients discharged during the observation period, totalling 1,415,754 patient records across 624 hospital departments. The fact that the entire patient population for each hospital department is included allows us to calculate daily midnight patient counts for each hospital department. However, because we use discharge records, patients who were admitted during the observation period but discharged after the end of this period are not included in our data. Therefore, the end of the observation period does not constitute a complete patient census and patient counts based on the data are too low. We therefore remove all patients who were discharged during the final month of the observation period of the hospital from further consideration. This is prudent in light of an average length of stay of 11 days in the sample. Similarly, if a patient was admitted before the observation period, we cannot calculate patient counts over their entire stay as we have no prior census information. We remove these patients as well.

As high occupancy is unlikely to impact the mortality risk of relatively healthy patients, we select a subsample of patients with high mortality risk. We focus on patients with six primary diagnoses identified by the US Department of Health as 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” (Agency for Healthcare Research and Quality 2006): Acute myocardial infarction (AMI), congestive heart failure (CHF), gastrointestinal hemorrhage (GIH), hip replacement after fracture (HIP), pneumonia (PNE) and stroke (STR). We discard the day of admission itself, and all patients who died or were discharged on the day of admission, because our focus is on occupancy effects in bed-bearing departments and we do not have workload metrics of emergency departments or information about arrival modes, which are likely to have a specific effect on survival on the day of admission. To further increase the homogeneity of the sample, we remove all hospitals that do not have emergency admissions, such as rehabilitation clinics, as such hospitals are unlikely to have critically ill patients. For departments with a small volume of our high-risk patients, mortality is rare and department fixed effects together with patient covariates can predict survival perfectly. This leads to numerical instability of the maximum likelihood optimization procedure, which we avoid by removing all patients who were in departments for which the department fixed effect leads to a perfect prediction. The remaining sample consists of 82,280 patients in 256 departments of 83 hospitals.
4.3. The Need for a Survival Analysis

We provide a first indication of the tipping point on the basis of raw data in Figure 2, which also illustrates the importance of controlling for time already spent in the hospital in the statistical analysis. The figure is based on a patient-day data set, where each observation corresponds to a record of one inpatient day for a particular patient, including an appropriate measure of occupancy experienced by the patient up to this observation day. Figure 2 is based on peak occupancy, defined as the maximum of all midnight occupancy levels experienced by the patient up to the observation day. We will discuss occupancy measures in more detail in Section 4.4. The left-hand graph of Figure 2 shows how mortality varies across the deciles of peak occupancy, initially decreasing but with a surprisingly marked increase at the 10th decile. The fact that mortality decreases with peak occupancy is to be expected: First, as the middle graph in Figure 2 shows, the mortality risk decreases with the length of stay as most deaths occur during the first few days of the stay, when the patient is most critically ill. Second, as illustrated by the right-hand graph, peak occupancy increases over time: the longer a patient stays in hospital the more likely it is that she will be exposed to high occupancy levels at some point during her stay. The combination of decreasing mortality and increasing peak occupancy with time results in a decrease of mortality with peak occupancy as indicated in the left-hand graph up to the 9th decile of peak occupancy. This expected trend, however, is starkly reversed at the 10th decile, in accordance with the tipping point hypothesis.

Figure 2 illustrates that the time already spent in hospital is an important control variable as it affects mortality as well as the likelihood of experiencing a period of high occupancy. This informs the choice of an appropriate model. The simplest mortality model would focus on patient episodes as the units of observation, survival as a dichotomous episode outcome, and an appropriate aggregate measure of occupancy during the patient episode as the independent variable of interest (e.g. KC and Terwiesch (2009)). Within this model, however, it is not possible to control for time spent in the hospital. Using length of hospital stay as a control variable is problematic because mortality curtails length of stay, which leads to reverse causality. For this reason, we have chosen a discrete-time survival model, based on patient-days as units of observation, which takes account of the effect of time spent in the hospital. We explain the model in more detail in Section 5.

Occupancy may affect patients differently at different stages of their hospital stay, leading to heterogeneity of effects. Specifically, patients are likely to be more critically ill during the early phase while they recover during the later phase of their stay. Convalescent patients are not as care-intensive, specifically with regard to monitoring requirements, and will be less vulnerable to deviations from optimal care. As shown in the death rate histogram of Figure 2, deaths are most likely during the first week of a patient’s stay, which is when occupancy induced adverse events are most likely to affect mortality and when these effects are most likely to be statistically detectable. In order to reduce sample heterogeneity, we therefore focus on the first week of stay as our observation period. All admitted patients are observed daily at midnight over the first seven days of their stay.
or up to their death or discharge if this occurs before the end of the seventh day in hospital. After the seventh day, patients are not further observed. Such prescribed follow-up periods are common in clinical and epidemiological studies and known as administrative or modified type I censoring (Klein and Moeschberger 1997). Table 1 contains summary statistics of the patient-day sample.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Patients</th>
<th>Percentage of full sample</th>
<th>Emergency admissions</th>
<th>Age (mean)</th>
<th>Mortality</th>
<th>Length of stay (days)</th>
<th>7-day mortality</th>
<th>7-day discharges</th>
<th>7-day patient-days</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>12,811</td>
<td>15.6%</td>
<td>53.1%</td>
<td>68.1</td>
<td>9.4%</td>
<td>10.0</td>
<td>5.9%</td>
<td>39.8%</td>
<td>69,985</td>
</tr>
<tr>
<td>CHF</td>
<td>17,852</td>
<td>21.7%</td>
<td>44.5%</td>
<td>74.9</td>
<td>9.0%</td>
<td>11.4</td>
<td>4.8%</td>
<td>29.7%</td>
<td>109,089</td>
</tr>
<tr>
<td>GHI</td>
<td>9,029</td>
<td>11.0%</td>
<td>48.5%</td>
<td>64.9</td>
<td>5.2%</td>
<td>9.2</td>
<td>2.8%</td>
<td>50.9%</td>
<td>48,128</td>
</tr>
<tr>
<td>HIP</td>
<td>7,974</td>
<td>9.7%</td>
<td>66.8%</td>
<td>73.5</td>
<td>5.3%</td>
<td>17.0</td>
<td>2.4%</td>
<td>10.9%</td>
<td>52,795</td>
</tr>
<tr>
<td>PNE</td>
<td>14,610</td>
<td>17.8%</td>
<td>50.5%</td>
<td>63.8</td>
<td>11.5%</td>
<td>11.7</td>
<td>6.2%</td>
<td>27.2%</td>
<td>91,179</td>
</tr>
<tr>
<td>STR</td>
<td>20,004</td>
<td>24.3%</td>
<td>61.9%</td>
<td>69.7</td>
<td>10.5%</td>
<td>13.5</td>
<td>6.5%</td>
<td>20.6%</td>
<td>126,404</td>
</tr>
</tbody>
</table>

Full sample 82,280 100.00% 53.7% 69.4 9.1% 12.1 5.2% 29.1% 497,580

4.4. Occupancy as a Time-varying Covariate
As occupancy refers to the percentage of used capacity, we need to first measure capacity for hospital departments. The natural measure is the number of beds in operation; however, this number is rarely available as public documents refer to the number of certified hospital beds. Interviews with hospital managers revealed that this number can deviate significantly from the number of beds in operation, which are fully resourced and readily available for patients, and is therefore not a reliable measure of operational capacity. In addition, certified bed numbers, while available in aggregate for hospitals, are not available at the department level, where we wish to measure occupancy. In the absence of reliable operational bed numbers, we therefore use the maximal daily midnight patient count in the department over the study period as a measure of the department’s capacity. We then calculate for every day of the patient’s stay the daily capacity utilization as the ratio of the midnight patient count at the beginning of the day and the department’s capacity.

In the survival analysis framework, where we observe patients daily, occupancy is a time-varying covariate. A critical question is how exposure to varying occupancy levels over time should be measured for an individual patient. The midnight patient count at the beginning of day \( t \) would seem the most natural candidate to affect mortality on day \( t \). However, with this occupancy metric we would only capture the immediate effect on the observation day; lagged effects, two or three days hence, would be discarded. In view of the rareness of avoidable deaths, as indicated in Section 4.1, it is unlikely that we will have enough power in our data to detect this effect. A second candidate would be the average occupancy experienced up to the beginning of day \( t \). However, the use of this metric is problematic within a tipping point model and is likely to lead to a spurious detection of a nonlinear effect: Time-averaging induces heteroscedasticity over time as the standard deviation of the average occupancy up to day \( t \) decreases with \( t \); very high or very low occupancy levels are therefore more likely for small \( t \). At the same time, mortality decreases with \( t \) as well, i.e., patients are more likely to die early during their stay, as indicated in the histogram in Figure 2. Therefore, mortality is naturally positively correlated with extremes of average occupancy, either very high or very low, as these extremes are more likely for small \( t \), when the variance of the time-averaged occupancy measure is larger. The time dummies in the survival model do not control for this effect because they only affects the intercept of the occupancy curve but not its shape. We would therefore estimate a nonlinear effect for time-averaged occupancy, even when occupancy has no effect.
We choose the *maximal* midnight occupancy level up to the beginning of day $t$ as the measure of the occupancy experienced to date by a patient in the hospital on day $t$. This *peak occupancy* metric has the advantage of being monotonically increasing over the stay in the hospital, which captures an important lag effect: Exposure to high occupancy on day $t$ can lead to death at a later day and cannot be "undone" by low occupancy after day $t$. This monotonicity property of the time-varying exposure also introduces a positive correlation with time, which is itself negatively correlated with mortality. Since the resulting correlation between peak occupancy and mortality is negative, the detection of a positive effect of peak occupancy on mortality beyond a tipping point will only become more difficult, which renders significant estimates conservative.

### 4.5. Control Variables

The need for risk-adjustment of patient-level data is comprehensively discussed in the literature (Iezzoni 2003). Our discharge records contain several variables that allow us to control for patient heterogeneity. Beside the primary medical condition and the individual risk factors age, gender and emergency admission, the presence of secondary diagnoses is an important source of heterogeneity. To account for these comorbidities we follow Needleman et al. (2011) and use indicator variables for a list of coexisting conditions (Elixhauser et al. 1998), adapted to the German context following Quan et al. (2005). In addition, we control for admission from another hospital with a dichotomous variable and for departmental transfers within the hospital during the stay of a patient with a time-varying exposure dummy which takes the value one on all days following the first transfer in the hospital.

We include day-of-stay dummy variables to model the baseline mortality hazard over the patient stay. Seasonal effects must also be controlled for as there might be times of the year when certain conditions occur more frequently or in a more severe way, e.g. through the winter months, and when occupancy in hospitals is also higher. Time-of-the-year can therefore confound results. To control for potential temporal correlations, we include dummy variables for admission month-of-the-year and for the 2005 observation year. To account for weekly patterns, such as differential staffing on weekends (KC and Terwiesch 2009), we include dummy variables for the weekday of the observation day. In addition, we control for the weekday of the admission to account for the so-called weekend-effect discussed in the medical literature (e.g. Bell and Redelmeier (2001)): patients who are admitted on weekends have a significantly higher mortality risk relative to weekday admissions, even after controlling for their individual risk factors.

Finally, we use department dummy variables to control for organizational heterogeneity in an aggregate way as departments will have differences in case-mix, size and staff endowment. We account for hospital effects by clustering error terms in the estimation at the hospital level.

### 5. Econometric Specification

We wish to estimate the association between the occupancy levels that patients experience during their hospital stay and the probability of in-hospital survival. As occupancy levels are most reliably calculated on the basis of midnight counts, a discrete-time survival analysis using patient-day observations is a natural modelling framework for this purpose. The population of interest in this study consists of patients who are admitted to hospital with one of the six high-risk conditions discussed in Section 4.2, and who survive until midnight of their day of admission to the hospital. By beginning our observation at midnight after admission we ensure that all observation periods have equal length. We follow patients up for seven days after admission and wish to estimate patient $i$'s discrete mortality hazard on day $t$ after admission

$$h_{it} = P[T_i = t \mid T_i > t-1, X_{it}], t = 1, \ldots, 7, \quad (1)$$
where $T_i$ denotes the day of death of patient $i$, counted from the day of admission, and $X_{it}$ is a covariate vector that is observable at the beginning of day $t$. The time-varying covariate vector $X_{it}$ includes dummy variables for each period $t$, which captures a baseline hazard model as a time dependent intercept, as well as a component $X_{jit}$ for peak occupancy experienced by patient $i$ up to the beginning of day $t$. The most common logit, probit and cloglog specifications for the discrete time hazards (Singer and Willett 2003) give very similar results. We report results for the probit specification

$$P[Y_{it} = 1 \mid X_{it}] = \Phi(X_{it}\beta), \tag{2}$$

where $Y_{it}$ is the observed dichotomous mortality variable, taking on the value 1 if patient $i$ dies on day $t$ and 0, otherwise, and $\Phi$ is the standard normal cumulative distribution function. This model admits a natural extension to test for the potentially confounding selection effect of early discharge as we will explain in Section 5.3.

5.1. The Tipping Point Model
We use a piecewise linear specification to estimate a potential tipping point with respect to occupancy. Specifically, assuming the peak occupancy that patient $i$ experienced during their stay up to day $t$ is stored in the $j$th component $X_{jit}$ of the covariate vector $X_{it}$, we use the parametric specification

$$\beta_{j1}X_{jit} + \beta_{j2}\max\{X_{jit} - \beta_{j3}, 0\} \tag{3}$$

to model the tipping point $\beta_{j3}$. Here, $\beta_{j1}$ is the slope of the line to the left of the tipping point and $\beta_{j2}$ captures the change in slope of the line as $X_{jit}$ exceeds the tipping point $\beta_{j3}$. We estimate all three parameters $\beta_{ji}, i = 1, 2, 3$.

The tipping point model is parsimonious with the minimum number of required parameters: One for the tipping point, and one for the behavior of the function on either side of the tipping point. It has several advantages over the more common polynomial specification of a nonlinear effect. First, it treats the tipping point explicitly as a parameter, which will allow us to estimate confidence intervals for the tipping point; second, its estimates have an immediate interpretation; third, the shape of the piecewise linear function can be asymmetric, with different slopes on either side of the tipping point. In contrast, polynomial models, with maxima and minima as candidates for tipping points, exhibit symmetric second order behavior and therefore symmetric shapes in the vicinity of these tipping points. The disadvantage of the piecewise linear model is that the term $\beta_{j2}\max\{X_{jit} - \beta_{j3}, 0\}$ in (3) renders the probit maximum likelihood problem non-concave. Fortunately, concavity is restored once the tipping point $\beta_{j3}$ has been fixed in (3). In order to optimize the likelihood function, we first estimated the remaining parameters repeatedly for a range of tipping points $\beta_{j3}$ and then used a procedure suggested by Muggeo (2003) to check optimality and estimate the standard error of the tipping point estimate.

5.2. Average Partial Effects
In view of the difficulty in interpreting coefficient estimates in generalized linear models, it has become customary to base statistical inference on average partial effects (APE) (Wooldridge 2002). Given probit estimates $\hat{\beta}$ based on (2), individual patient-level partial effect estimates $\nabla_X P[Y_{it} = 1 \mid X_{it}] = \phi(X_{it}'\hat{\beta})\hat{\beta}$ are aggregated to average partial effects

$$\text{APE}(\hat{\beta}) = \frac{1}{N} \sum_{i,t} \phi(X_{it}'\hat{\beta})\hat{\beta}, \tag{4}$$

where $\phi(z) = \Phi'(z)$ is the standard normal density and the sum is taken over all $N$ patient-days $(i, t)$ in the sample. The APE has a natural population-based interpretation as the proportional effect of a unit increase in a covariate across all patient-days (Wooldridge 2002). The asymptotic
variance-covariance matrix of the APE can be obtained via the delta method and is of the form $M \hat{V} M'$, where $\hat{V}$ is the variance-covariance matrix of the probit coefficient estimates,

$$M = \frac{1}{N} \sum_{i,t} \phi(X'_{it}\hat{\beta})(I - (X'_{it}\hat{\beta}\hat{\beta}'X_{it})),$$

and $I$ is the identity matrix (see Chapter 2.6.6. of Green and Hensher (2010)). In the context of the tipping point model, we are interested in the average partial effect of peak occupancy below and above the tipping point. In order to compute these average partial effects, we have to average over the appropriate subsample of patient-days with peak occupancy above and below the tipping point, respectively, rather than over the entire sample. Note, however, that model (3), does not provide a direct estimate of the slope above the tipping point $\beta_{j3}$; instead this slope is the sum of the two correlated estimates $\beta_{j1}$ and $\beta_{j2}$. Furthermore the slope $\beta_{j1}$ applies to occupancy below and above the tipping point. We therefore re-estimate the model for the already optimized tipping point $\beta_{j3}$, using the following reparametrization of (3):

$$\tilde{\beta}_{j1} \min\{x_{jxt}, \beta_{j3}\} + \tilde{\beta}_{j2} \max\{x_{jxt} - \beta_{j3}, 0\}. \tag{6}$$

In this model $\tilde{\beta}_{j1} = \beta_{j1}$ estimates the slope below the tipping point $\beta_{j3}$ while $\tilde{\beta}_{j2} = \beta_{j2} + \beta_{j1}$ estimates the slope above the tipping point. We cannot use this parametrization of the tipping point model for the estimation of the tipping point as Muggeo (2003)’s standard error estimation for the tipping point does not apply to this model. However, once the tipping point has been estimated on the basis of (3), we can fix it and re-estimate the remaining parameters using the new parametrization (6). We then calculate average partial effects associated with $\tilde{\beta}_{j1}$ and $\tilde{\beta}_{j2}$ via (4) and (5) by averaging over the relevant subsamples of patient-days with peak occupancy below or above the tipping point.

### 5.3. Early Discharge as a Competing Risk

If we find an association between occupancy levels and mortality, as hypothesized, then this association in itself is not yet evidence for a tipping point effect of occupancy on patient safety. In fact, there is an alternative explanation: The mortality hazard is estimated by considering the number of patients who die on day $t$ of their stay as a proportion of the total number of patients who are still in the hospital at the beginning of day $t$ and therefore at risk of dying during day $t$. We hypothesize that occupancy has an effect on the number of patients who die - but the observation could equally well be explained by an occupancy effect on the number of patients that are left at risk. Patients are discharged as their health status improves and this discharge trigger may well be affected by occupancy levels. When occupancy is high, doctors make space for new patients by discharging some patients earlier (KC and Terwiesch 2012). There is some evidence that doctors choose less ill patients for early discharge (Long and Mathews 2013). This is one of the reasons why we chose a seven day follow-up period for our analysis. Given an average length of stay of twelve days for the patients in our sample, we believe that relatively few of these patients will recover enough during the first seven days to be candidates for premature discharge in response to elevated occupancy levels.

We perform robustness checks to ascertain whether our results are affected by early discharge. Specifically, we estimate two models that allow for early discharge as a competing risk, to see whether this affects our results. The first model is based on the concept of a subdistribution hazard, which was introduced by Fine and Gray (1999) to deal with competing risks and has found wide-spread applications in epidemiology (Lau et al. 2009). When discharge is not considered as a competing risk, discharged patients are treated as censored on the day of discharge. This assumes that, conditional on covariates, discharged patients have the same mortality hazard as patients who
remain in the hospital. This is clearly a limiting assumption as discharged patients with the same covariates may well be healthier on factors observable to the discharging doctor but unobservable to the researcher. Rather than censoring discharged patients on the day of discharge, the Fine-Gray approach maintains the records of the discharged patients in the data beyond the time of discharge. This is operationalized by duplicating the patients’ records on the day of discharge and censoring the patient at the end of the seven day follow-up period. The approach therefore expands the risk set on day $t$ by all patients who were discharged prior to day $t$ and estimates the mortality hazard relative to this enlarged risk set. This estimates the subdistribution hazard, which is the probability that a patient dies on day $t$ given that she has survived up to day $t$ or has been discharged prior to day $t$ (Lau et al. 2009). By keeping discharged patients at risk, the approach removes the problem that an estimated change in the hazard may be due to the effect of occupancy on the risk set rather than on the number of patients who die. Estimation is achieved by applying model (2) to the expanded data set.

5.4. Early Discharge as a Selection Problem

In addition to the subdistribution approach, we extend the probit model (2) to account for discharge in a bivariate probit model with selection (Green 2003). As in the standard probit model, we assume that death occurs on day $t$ when a latent sickness status $Y^*_it$ becomes positive. Analogously, discharge on day $t$ is triggered by a second latent variable $S^*_it$ which is interpreted as the difference between the clinician’s utilities from keeping patient $i$ in hospital beyond day $t$ or discharging her on day $t$. If $S^*_it > 0$ then the patient is kept in hospital beyond day $t$, otherwise she is discharged on day $t$. Both latent variables are assumed to depend on individual covariates and random errors

\[
Y^*_it = X_it\beta + \epsilon_it
\]
\[
S^*_it = Z_it\gamma + \nu_it.
\]

As in the probit model (2), both error terms are assumed to follow standard normal distributions. However, some of the unobserved factors that explain the latent sickness status may also affect the discharge decision. We therefore allow for correlated errors and assume them to be sampled from a bivariate normal distribution with a correlation coefficient $\rho$ that needs to be estimated. By allowing for correlated errors in the simultaneous equations, we lift relevant unobserved information from the discharge equation to the mortality equation.

In a standard bivariate probit model, one observes each of the four combinations of the two events discharge and death. In our context, however, we do not observe the combined outcome death and discharge but only one of three event combinations

1. patient $i$ was discharged on day $t$
2. patient $i$ was not discharged on day $t$ and died on day $t$
3. patient $i$ was not discharged on day $t$ and survived day $t$.

This leads to a bivariate probit model with selection (see Chapter 21.6.4 of Green (2003)). This set-up allows us to test for confounding by early discharge in our sample by testing whether $\rho$ is significantly different from zero.

In order to estimate this model robustly, we need an instrumental variable for the discharge equation, i.e. a variable that is correlated with patient $i$’s probability of discharge but does not directly affect her probability of death on day $t$, conditional on the other covariates. We used the discharge rate amongst the other patients in the department on the day of interest as an instrument. Any systematic variation in discharge behavior should affect all patients, so there ought to be a correlation between the discharge rate of the other patient and patient $i$’s probability of discharge. However, it is unlikely that variation in the discharge rate of the other patients in the department is correlated with patient $i$’s probability of dying on day $t$. To make the instrument comparable between departments, we calculated the z-score of discharge numbers for each day in
each department by normalizing daily discharge numbers with respect to the average number of discharged patients and the standard deviation in the department during the observation period. Specifically, we calculated for patient $i$ in department $j$ on day $t$ of her stay the variable
\[ z_{ijt} = \frac{d_{ijt} - (\bar{d}_j - 1)}{\sigma_j}, \] (8)
where $d_{ijt}$ is the number of other patients (excluding patient $i$) discharged on day $t$ of patient $i$’s stay in department $j$, and $\bar{d}_j$ and $\sigma_j$ are the average and standard deviation of the daily discharge numbers in department $j$ during the observation period. We use $\bar{d}_j - 1$ instead of $\bar{d}_j$ in the numerator because the instrument relates to all patients except patient $i$.

6. Results
We estimated the discrete daily survival model (2) using the probit command and the bivariate probit model with selection (7) using the heckprob command in STATA, Version 12. Table 2 shows the estimation results for the variables of interest and a selection of control variables. The occupancy coefficients $\beta_{ijk}$ and $\tilde{\beta}_{ijk}$ refer to the tipping point parametrizations (3) and (6), respectively. The controls listed are the day of stay (Day 1 omitted), the primary condition (AMI omitted), the day of the week of admission (Sunday omitted), whether the patient has had an internal transfer within the hospital prior to the observation day (Int.Trans.), whether the patient had been referred from another hospital (Hosp.Ref.) and whether the patient was an emergency admission (Emerg.). The correlation coefficient $\rho$ refers to the estimated correlation between the error terms of the bivariate probit model with selection (7). The second column contains estimates of a null model with control variables only. As expected from Figure 2, the mortality risk reduces over time from Day 2 to Day 7. With the exception of stroke, all conditions have a significantly lower mortality risk than acute myocardial infarction (AMI). The estimations of the weekday of admissions coefficients confirm the so-called weekend effect discussed in the medical literature (e.g. Bell and Redelmeier (2001)): patients who are admitted on weekends have a significantly higher mortality risk relative to weekday admissions, even after controlling for the usual risk factors. This translates into significantly negative coefficients of Adm-Mon - Adm-Fri relative to the omitted dummy variable Adm-Sun for patients who are admitted on Sundays. Finally, internal transfers and emergency admissions are expected to be correlated with patient severity and are indeed significantly positively associated with mortality. The fact that a patient was referred from another hospital (Hosp.Ref. = 1) has no significant effect on mortality.

The third column of Table 2 includes peak occupancy up to the observation day as a linear variable. The estimated coefficient ($\text{beta}=0.101, p > 0.05$) is not statistically significant and affirms the difficulty of finding robust occupancy effects on mortality within a linear occupancy model, even in large samples (see also KC and Terwiesch (2009)).

The results in the fourth column were obtained by estimating the tipping point model with the parametrization (3). This parametrization is particularly suited for the estimation of the standard error of the estimated tipping point. We first estimated the model for fixed tipping points with a 1% spacing at 85%, 86%, ..., 98%, resulting in a maximal likelihood at 92%, and then used the procedure suggested by Muggeo (2003) to further optimize locally in the vicinity of 92% and to calculate the standard error of the final estimate. The tipping point was estimated at 92.5%, with a 95% confidence interval of [89.0%, 96.0%]; 14.7% of the patients in the sample had experienced occupancy levels above the tipping point of 92.5% during their first seven days in hospital. While the linear effect of peak occupancy is insignificant ($\text{beta}=-0.00789, p > 0.05$), the slope changes significantly when peak occupancy exceeds the tipping point ($\text{beta}=2.083, p < 0.01$). The overall model fit is significantly improved relative to the linear model ($\text{deviance} = 17.0, p < 0.01$). These estimations support the tipping point hypothesis.
### Table 2  Seven day in-hospital survival estimates for selected covariates

<table>
<thead>
<tr>
<th></th>
<th>Null model</th>
<th>Linear model</th>
<th>Tipping point model I</th>
<th>Tipping point model II</th>
<th>Average partial effect</th>
<th>Selection model</th>
<th>Sub-distrib. hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{13}$</td>
<td>0.101</td>
<td>0.925</td>
<td>(0.0590)</td>
<td>(0.018)</td>
<td>-0.00789</td>
<td>-0.0533*</td>
<td>-0.00319</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>2.083**</td>
<td>-0.0017</td>
<td>(0.479)</td>
<td>(0.0258)</td>
<td>-0.00017</td>
<td>0.0469</td>
<td>0.0633</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>2.075**</td>
<td>0.046**</td>
<td>(0.450)</td>
<td>(0.0281)</td>
<td>2.028**</td>
<td>0.101</td>
<td>1.984**</td>
</tr>
<tr>
<td>Occ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.0272</td>
<td>(0.0493)</td>
</tr>
<tr>
<td>IV $z_{ij}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.473**</td>
<td>(0.0986)</td>
</tr>
</tbody>
</table>

#### Day 2
-0.105**  -0.108**  -0.107**  -0.107**  -0.108**  -0.214**  -0.0943**

#### Day 3
-0.170**  -0.175**  -0.174**  -0.174**  -0.174**  -0.349**  -0.159**

#### Day 4
-0.165**  -0.172**  -0.169**  -0.169**  -0.179**  -0.519**  -0.154**

#### Day 5
-0.191**  -0.199**  -0.197**  -0.197**  -0.198**  -0.692**  -0.179**

#### Day 6
-0.254**  -0.263**  -0.262**  -0.262**  -0.261**  -0.866**  -0.235**

#### Day 7
-0.274**  -0.283**  -0.283**  -0.283**  -0.198**  -3.299**  -0.237**

#### CHF
-0.173**  -0.173**  -0.173**  -0.173**  -0.169**  -0.331**  -0.152**

#### GIH
-0.384**  -0.384**  -0.384**  -0.384**  -0.381**  -0.619**  -0.393**

#### HIP
-0.611**  -0.611**  -0.611**  -0.611**  -0.610**  0.190*  -0.559**

#### PNE
-0.115**  -0.116**  -0.116**  -0.116**  -0.109**  -0.0784  -0.0802**

#### STR
0.0283  0.0283  0.0276  0.0276  0.0264  0.0930  0.0555

#### Adm-Mon
-0.0509*  -0.0534*  -0.0533*  -0.0533*  -0.0562**  0.114**  -0.0539*

#### Adm-Tue
-0.0781**  -0.0796**  -0.0781**  -0.0781**  -0.0819**  0.222**  -0.0722**

#### Adm-Wed
-0.0340  -0.0347  -0.0324  -0.0324  -0.0349  0.209**  -0.0259

#### Adm-Thu
-0.0585**  -0.0581**  -0.0561**  -0.0561  -0.0592**  0.221**  -0.0469

#### Adm-Fri
-0.0650**  -0.0623**  -0.0627**  -0.0627**  -0.0684**  0.328**  -0.0526**

#### Adm-Sat
-0.0287  -0.0259  -0.0272  -0.0272  -0.0292  0.119**  -0.0250

#### Int.Trans.
0.153**  0.152**  0.152**  0.152**  0.153**  0.220**  0.162**

#### Hosp.Ref.
-0.0185  -0.0184  -0.0188  -0.0188  -0.0246  0.225**  -0.0574

#### Emerg.
0.136**  0.136**  0.137**  0.137**  0.139**  0.0996**  0.137**

| $\rho$      | 0.47**      | (0.09)       |
| N           | 497,580     | 497,580      |
| log lik.    | -1891.9     | -21890.6     | -21882.1  | -21882.1  | -10252.13 | -22272.6 |

*p < 0.05, **p < 0.01, robust standard errors in parentheses, clustered by hospital.
6.1. Effect Size Estimations

In order to estimate average partial effects, we re-estimate the tipping point model with the alternative parametrization (6) as explained in Section 5.2. The results are contained in the fifth and sixth column of Table 2. The average partial effects allow us to estimate the size of the occupancy effect. Of the 497,580 patient-days in our sample, 51,318 (10.3%) were associated with historical peak occupancy above the estimated tipping point. The average historical peak occupancy on these days was 95.9% and the average daily mortality rate on these days was 0.00892, i.e., on average 892 of 100,000 patients who had experienced occupancy above the tipping point in the past died every day. The average partial effect estimate of peak occupancy above the tipping point is 0.046 with a 95% confidence interval of [0.024, 0.068]. This suggests that a reduction of peak occupancy above the tipping point by one percentage point, reducing average occupancy from 95.9% to 94.9%, would reduce the daily mortality rate on these days by 0.00046 (95% CI=[0.00024, 0.00068]). In aggregate, the daily loss of 892 patients per 100,000 patients would be reduced by 46 patients (95% CI=[24, 68]), a reduction of the death rate by 5.2% (95% CI=[2.7%, 7.6%]). This is a clinically significant effect associated with a modest one percentage point reduction in peak occupancy above the tipping point.

The tipping point model suggests an alternative estimation of the total number of avoidable deaths in the sample: Our model allows us to predict the mortality hazard on each observed patient-day for varying levels of peak occupancy. For each patient-day with a peak occupancy value above the estimated tipping point level of 92.5% we can therefore compare the predicted mortality hazard \( h_i \) in (1) with the predicted hazard \( h'_i \) if peak occupancy is reduced to the tipping point level. Summing up the differences between these two hazards over all patient-days in the sample provides an estimate of lives saved if no patient had experienced peak occupancy above the tipping point. For our sample this results in 78 saved lives of 4247 deaths, an overall reduction in mortality by 1.8%. In other words, 1 in 55 deaths in the sample is accounted for by occupancy above the tipping point. Since peak occupancy can only have an effect on avoidable deaths, it is instructive to relate this estimate to avoidable death estimates in the medical literature. Such estimates vary significantly, depending on the context (see Lessing et al. (2010) for a review). Combining an extreme estimate of a 44.1% avoidable death rate (Healey et al. 2002) with the estimated 1 in 55 deaths due to high occupancy results in a combined estimate of 4.2% avoidable deaths due to high occupancy. Combining our estimate with a more realistic assumption that 1 in 10 deaths are avoidable in our high-risk sample would suggest that 18% of these avoidable deaths are associated with peak occupancy above the safety tipping point.

So far, we have only discussed the size of the effect of high occupancy on all patients. However, 85.3% of patients in our sample did not experience occupancy above the tipping point. If we relate the estimated 78 saved lives to the 458 death among the 12,130 patients who experienced occupancy levels above the tipping point, the size of the effect becomes substantially larger: 17% of deaths of these patients could have been avoided if no patient had been exposed to occupancy above the tipping point.

7. Robustness Checks

7.1. Proportional Hazards

The survival model assumes that the effect of occupancy is the same during all days of a patient’s stay in the hospital. To test this proportionality assumption, we divide the seven day observation window into an early phase from the first to the third day of stay and a late phase from the fourth to the seventh day and code phase-dependent slopes on either side of the estimated tipping point, using a phase dummy variable (Singer and Willett 2003). We use the parametrization (6), i.e., the comparator model is the model in the fifth column of Table 2. The estimated coefficients for peak occupancy below the tipping point are 0.035 (sd=0.072, \( p > 0.1 \)) for the first phase and -0.085
(sd=0.087, \( p > 0.1 \)) for the late phase, while the coefficient estimates for peak occupancy above the tipping point are 2.103 (sd=0.809, \( p < 0.01 \)) for the early phase and 2.171 (sd=0.664, \( p < 0.01 \)) for the late phase. A Wald test does not reject the equal coefficient hypothesis at the 10% significance level. We are therefore satisfied that the proportionality assumption is tenable in our case.

7.2. Expanded Observation Period

One reason why we chose the short seven day follow-up period is that we expect the non-proportionality assumption to be violated for longer observation periods because patients are likely to be more critically ill during the early phase of their stay and are often convalescent and therefore not as vulnerable and not as care-intensive during the later phase. To validate this, we estimated a model over a 14-day observation window. We estimated a significant tipping point at 90.8% peak occupancy. However, allowing for different slopes in the first and second week of the patient’s stay, as in Section 7.1, significance is only maintained for the first week while slope estimates for the second week become insignificant. A Wald test rejects equality of the coefficients and therefore proportionality of the effect across two weeks. However, we cannot conclude that there is no occupancy effect on mortality during the second week but only that it is likely to be different. The fact that mortality is considerably lower during the second week (see Figure 2) makes the detection of a tipping point effect statistically more difficult. In light of Section 4.1 our study is likely to be underpowered for an analysis of the effect beyond the first week of stay. This justifies our focus on the first week of a patient’s stay.

7.3. Results of the Selection Model

Columns 7 and 8 of Table 2 relate to the selection model explained in Section 5.4. In contrast to mortality, occupancy is less likely to have a lagged effect on discharge. Instead of historical peak occupancy we therefore use occupancy on the day prior to the observation day (called “Occ” in Table 2) in the discharge equation. In addition, we use the instrument \( z_{ijt} \) defined in (8). The results provide some evidence that early discharge in response to elevated occupancy levels occurs as the estimated correlation coefficient between the error terms is positive and significant (\( \rho = 0.47, p < 0.01 \)). This is to be expected as unobserved factors that increase a patient’s probability of being selected into the sample on the next day, i.e. not discharged home, are likely to be related to poor health and therefore increase the patient’s mortality risk. Including the selection equation, however, does not substantially change the peak occupancy slope coefficients \( \beta_{11} \) and \( \beta_{12} \) to the left and right of the tipping point, respectively, when compared to the single-equation model estimates in column 5 of Table 2. We therefore conclude that early discharge bias is small and that we can safely work with the simpler single-equation probit model, which has the advantage of being easier to communicate.

7.4. Subdistribution Hazards

The final column of Table 2 provides the results of the subdistribution hazard estimation explained in Section 5.3. This model accounts for discharge in a different way, namely as a competing risk to mortality. The fact that patients are discharged based on their health status leads to a systematic thinning process of the sample that is correlated with the mortality process: The mortality risk of those patients who are discharged cannot be assumed to be the same as the mortality risk of patients who remain in the hospital. Therefore, they should not be treated as censored, which is what the probit model in the fifth column of Table 2 assumes. The subdistribution approach takes the competing risk of discharge into account by estimating the subdistribution hazard, i.e. the probability that patient \( i \) dies on day \( t \), conditional on being in the hospital at the start of day \( t \) or having been discharged before day \( t \). This is operationalized by keeping patients who are discharged before the end of the observation period of seven days in the data by duplicating the
observation on the discharge day sufficiently often (Lau et al. 2009). This duplication of discharged patient records explains the increased sample size to \( N = 557,828 \) patient-days. The slope of peak occupancy below the tipping point is estimated to be \(-0.000319 (sd = 0.0633, p > 0.1)\) while the slope above the tipping point is \(1.984 (sd = 0.446, p < 0.01)\). These estimates are very similar to the estimates in the comparator model in the fifth column of Table 2, which did not take the competing risk of discharge into account. This provides additional evidence that any bias due to discharge is small during the first week of hospital stay for our high-risk sample and that the simpler model that does not account for competing risks provides adequate estimates of the effect of peak occupancy on mortality hazards.

7.5. Multiple Tipping Points and Smooth Splines
To test whether models with multiple tipping points or smooth curves might fit the data better, we used a model selection procedure suggested by Royston and Sauerbrei (2007), implemented in the Statan command `uvrs`, which chooses amongst alternative spline models with multiple breakpoints. We allowed for a maximum of 10 spline pieces with 9 breakpoints located at the 10th, 20th, ..., 90th percentiles of peak occupancy. The algorithm chooses the best fitting model by comparing successively increasingly complex spline models, i.e. models with an increasing number of breakpoints across the possible locations, with the most complex model, i.e. the model with 9 breakpoints. The algorithm stops when this most complex model does not provide a significantly better fit at the 5% significance level, based on the chi-square statistic of log-likelihood differences. If all goodness-of-fit tests are significant, the most complex model is chosen. We first estimated linear spline models, i.e. continuous models with linear pieces between the breakpoints. This method identified the piecewise linear model with a single tipping point at the 90th percentile of peak occupancy as the best fit. Repeating this procedure with cubic instead of linear splines, allows for nonlinearity between breakpoints but forces smoothness at breakpoints. The resulting best-fitting cubic spline had also a single tipping point at the 90th percentile of peak occupancy but had lower likelihood than the piecewise linear model. The advantage of nonlinearity between breakpoints was insufficient to compensate for the forced smoothness at the breakpoints. We are therefore satisfied that the piecewise linear model with a single tipping point is appropriate.

8. Managerial Implications of the Tipping Point
Our empirical study provides evidence that occupancy levels above a tipping point are associated with a substantial increase in in-hospital mortality. If the tipping point is reached frequently, the hospital will experience a sustained quality problem, which may even lead to its closure (Ruef and Scott 1998). Two natural managerial levers are capacity increases and capacity pooling with near-by hospitals. The first will reduce occupancy levels across the board, the second will reduce the variability of occupancy levels. Both actions imply that fewer patients exceed the occupancy tipping point. We estimate the effects of these interventions on the basis of our data and discuss the value of flexibility in the context of capacity increase.

8.1. Rigid Versus Flexible Capacity
In this section we analyze the effect of a 1% increase of hospital capacity in the system on mortality and the associated cost. We consider two options: A rigid capacity expansion with fully staffed beds, and a semi-flexible capacity expansion, where beds are resourced fully with the exception of staffing, which is flexibly deployed in response to occupancy surges. An increase of capacity by 1% reduces peak occupancy for all patient-days by a factor 1/1.01. Our model allows us to predict the corresponding changes in daily mortality hazards which we sum up across the sample to obtain the number of saved lives in our sample. An increase of capacity across the sample by 1% reduces the number patients who were exposed to occupancy above the tipping point from 12,130 to 10,118, a
reduction by 16.6%. The model predicts that 18 deaths could have been saved with a 1% increase in capacity, amounting to 3.9% of the 458 patients who died after having experienced occupancy above the tipping point. Note that these 18 saved patients account for 23.7% of the 78 patients that could have been saved if no patient had been exposed to occupancy levels above the tipping point (see Section 6.1). For the cost-benefit analysis, we annualize the number of saved deaths per hospital as observation periods differ by hospital in the sample, resulting in 18.75 lives saved per annum in the hospitals in our sample.

We estimate the annual costs of a 1% increase in capacity using national average costs of the (German Bureau of Statistics 2013) and department-specific staffing information from published hospital reports. We differentiate between clinical staff costs, related to doctors and nurses, and other infrastructure and overhead costs of capacity, such as beds, space or support services. We consider two options: fixed staffing and flexible staffing. The fixed staffing option assumes that both clinical staff and other infrastructure costs in the department are increased by 1%, while for the flexible staffing option only infrastructure costs are increased by 1% and clinical staff costs are only increased by 1% on days when occupancy is above the tipping point.

For the fixed capacity option, we first increase medical staffing in all departments by 1%. For the 14 departments where we did not have staffing information, we used the mean of the staffing per capacity of the other departments of the same type. In aggregate, a 1% increase in clinical staffing in all departments in our sample requires 52 doctors and 162 nurses. Based on national average costs this results in total cost of 13.7M Euro. Taking account of departmental capacities and department-type specific national cost averages, we calculated costs for support services (radiology, pathology, anaesthesia) of 4.5M Euro, administrative overheads and logistics (e.g. kitchen services, energy, building maintenance) of 9.9M Euro, and cost of capital of 2.4M Euro. Capital costs are based on investment costs of 200,000 Euros per bed and a depreciation period of 25 years (Bavarian Ministry of Finance 2013). All other on-costs were calculated on the basis of national average costs of the three most frequent conditions (diagnosis related groups) in the department (InEK GmbH 2013). In summary, the estimated total annual cost of a 1% capacity expansion in the sample departments amounts to 30.5M Euros, of which 13.7M Euros are costs of medical staff. In relation to the 18.75 saved lives associated with a 1% capacity increase, this amounts to a cost of 1.63M Euros per live saved. This is a very conservative estimate of the benefits of a 1% capacity increase as it is likely that for each avoidable death there are many more adverse events that do not result in death but result in harm and associated additional medical, legal and reputational costs.

The tipping point phenomenon suggests to install semi-flexible capacity and employ this capacity only when occupancy reaches the safety tipping point. We can estimate the associated costs by assuming that the required infrastructure is installed but its medical staffing remains flexible. The total costs of capacity without medical staffing is 16.8M Euros. Since only 2.5% of departmental days in the sample had occupancy above the tipping point, this reduces staffing costs to 0.5M Euros, giving a total cost of 17.3M Euros, or 0.92M Euros per saved life; semi-flexible capacity is more than 40% cheaper than fixed capacity and achieves, due to the tipping point characteristic, the same mortality reduction.

### 8.2. Capacity Pooling

As capacity increase is associated with high costs, we study the effect of pooling as a potentially less costly alternative to managing the effect of the safety tipping point. Pooling reduces the variability of demand and therefore of occupancy levels, which in turn reduces the propensity of a patient experiencing occupancy levels above the tipping point. Pooling is often implemented at the hospital level through cooperation agreements. Such agreements can include transfers of patients before or following admission, as well as transfers of staff to cover shortages at a partnering hospital. Hitherto, the main focus of such measures was cost-reduction. However, such cooperations can also
positively affect service quality by reducing the proportion of days with occupancy levels above the tipping point. We can estimate this effect within our sample. In order to achieve synergies from pooling, especially for emergency patients, pooled hospitals should be in close proximity so that ambulance diversions in response to occupancy levels do not cause inappropriate delays. We use German zip-codes to estimate the distances between the hospitals in our sample. This allows us to group the 83 hospitals into clusters. We do this step-wise, starting from single hospital clusters, by merging two clusters if they contain two hospitals that are less than 30 km apart. This leads to 43 hospital clusters for our sample, of which the largest consists of 18 hospitals. The maximal distance between any two hospitals within any of the clusters is 53 km. We then recalculate daily occupancy levels for each department type across the hospitals in the clusters by pooling departmental capacities: For each day of the year we added midnight patient counts across departments of the same type in the cluster and divided these by the sum of the capacities of these departments to obtain cluster occupancy levels. Pooling reduced the 51,318 patient-days with peak occupancy above the estimated tipping point to 32,317 patient-days, a reduction by 37%. At the patient level, 12,130 patients were exposed to occupancy above the tipping point on some day of their first seven days in the hospital. After pooling, this number reduced to 7,716 patients. We can calculate the number of saved lives by calculating for each patient-day the difference between the model predicted mortality hazard for the realized peak occupancy level and for the pooled peak occupancy and summing up the differentials over all patient-days. This resulted in an estimated 30 lives saved by pooling. As we had estimated earlier that 78 patients could have been saved if no patients had been exposed to occupancy above the tipping point, this occupancy effect can be reduced by 39.5% by pooling alone.

8.3. Staff-to-bed Ratios and the Position of the Tipping Point

We end this section with a brief discussion of the effect of staff-to-bed ratios on the safety tipping point. The staff-to-bed ratio is an important management variable in a hospital and the subject of much debate in the medical literature. A recent large-scale study by Needleman et al. [20] provides evidence that shift staffing under target levels is associated with increased mortality. The results in a review of Lang et al. (2004) suggest that higher nurse staffing levels are associated with lower failure-to-rescue rates, lower inpatient mortality, and shorter hospital stays. A review article by Pronovost et al. (2002) concentrates on physician staffing levels and suggests that high-intensity staffing is associated with reduced hospital and intensive care unit mortality, and hospital and intensive care unit length of stay. The staff-to-bed ratio affects a hospital department’s ability to service expected demand and provides a buffer against variability in demand and in staffing itself (Bassamboo et al. 2010). Staffing levels per bed differ by department type as they admit patients with different acuity levels and care requirements. However, they also differ between departments of the same type. This source of variation allows us to investigate whether more highly-staffed hospital departments of the same type are less prone to the tipping point effect. Specifically, we would expect these departments to experience tipping points at higher occupancy levels than their low-staffed counterparts. To investigate this hypothesis, we use departmental staffing from standardized hospital quality reports published in 2004, which are available for 79 of the 83 hospitals in our sample. This reduces the data set slightly to 484,415 patient-days of 79,855 patients in 244 departments. We use the reported numbers to categorize departments as high-staffed or low-staffed, where a department was categorized as high-staffed if its staff-to-bed ratio was above the median ratio for departments of the same type in our sample. As before, we use the maximal midnight count of patients during the observation period as a proxy for the number of beds in a department; staff refers to both doctors and registered nurses.

In a standard linear model, differential effects of peak occupancy for low and high-staffed departments would be estimated by interacting the peak occupancy \( X_{jit} \) experienced by patient \( i \) up to
day $t$ of her stay with a staffing dummy variable $S_{it}$ that takes the value 1 if the patient’s department on day $t$ of her hospital stay has one of the departments with high staff-to-bed ratio, and 0 otherwise. Including direct effects of $X_{jit}$ and $S_{it}$, this leads to the linear term

$$\beta_0 S_{it} + \beta_{j1} X_{jit} + \beta_{j2} X_{jit} S_{it} = \beta_0 S_{it} + \beta_{j1} (X_{jit} S_{it} + X_{jit} (1 - S_{it})) + \beta_{j2} X_{jit} S_{it},$$

where $\hat{\beta}_{j2} = \hat{\beta}_{j1} + \hat{\beta}_{j2}$. Using the latter parametrization, tipping point model (6) can be used for both variables $X_{jit} (1 - S_{it})$ and $X_{jit} S_{it}$ to estimate nonlinear occupancy effects in the low- and high-staffed departments, respectively, leading to the model

$$\beta_0 S_{it} + \gamma_{j1} \min \{X_{jit} (1 - S_{it}), \gamma_{j2}\} + \gamma_{j2} \max \{X_{jit} (1 - S_{it}) - \gamma_{j3}, 0\} + \delta_{j1} \min \{X_{jit} S_{it}, \delta_{j3}\} + \delta_{j2} \max \{X_{jit} S_{it} - \delta_{j3}, 0\},$$

where $\gamma_{j3}$ and $\delta_{j3}$ are the tipping points for peak occupancy in the low-staffed and high-staffed departments, respectively.

In order to optimize the likelihood for the two tipping points we use a two-dimensional grid search with a spacing of 0.001 over the range of the peak occupancy variable $X_{jit}$, resulting in tipping point estimates of 86.5% for the low-staffed and 93.9% for the high-staffed departments. The model with a single tipping point for this subsample (equivalent to tipping point model I in Table 2) led to a tipping point estimate at 92.6%. Since the model with a single tipping point is nested in the model with two tipping points, with the constraint $\gamma_{j3} = \delta_{j3}$, a likelihood ratio test can be used to test whether the estimated tipping points are indeed different. The likelihood ratio test was not significant ($p=0.18$). This is not surprising, given the relatively large standard error estimated for the tipping point $\beta_{j3}$ in Table 2. Nevertheless, although our sample size does not give us enough power to distinguish the tipping points statistically, the estimated direction of the difference is as expected and the estimates therefore provide some initial evidence that bed staffing levels are important management lever to move the safety tipping point. Further research with a larger sample or more granular staffing data is necessary to quantify this effect more precisely.

9. Conclusion

Hospitals cannot turn away patients with acute conditions and therefore have to deal with surges in demand, leading to spikes in occupancy levels. When occupancy is very high, the managerial ability to respond by exploiting variability buffers becomes constrained as these buffers become depleted. The strain is passed on to employees, who are forced to ration limited resources to cope with excessive demand, while stress impairs their cognitive abilities. In combination, these effects lead to safety tipping points in hospitals. Neither the organization nor its clinical staff are able to absorb a further increase in occupancy beyond the safety tipping point without significant deterioration of the quality of care. Our empirical analysis demonstrates that such tipping points exist. In our sample, patients’ mortality risk begins to increase significantly with occupancy when occupancy levels exceed a tipping point of 92.5%. Our results provide ammunition for operations managers when their finance colleagues argue that capacity can be reduced while activity levels are maintained. When this is done, more patients will experience an unsafe day in the hospital, i.e. a day when occupancy levels exceed the safety tipping point. In our sample 14.7% of patients experienced days with occupancy above the estimated tipping point and 17 of 100 deaths among these patients in our sample are accounted for by high occupancy in their department.

The existence of safety tipping points is important. Earlier studies had neglected this phenomenon and did either not find a relationship between occupancy and mortality or exaggerated the effect at low occupancy levels and underestimated the effect at high levels. This is particularly
relevant in the debate about capacity pooling as the associated reduction in occupancy variability reduces the propensity of a patient to experience occupancy above the tipping point. In a simulation of capacity pooling, we have estimated that in our sample 39.5% of the deaths that are accounted for by occupancy could have been avoided if capacity had been pooled. This significant safety effect of capacity pooling is not apparent in a linear occupancy model, where a gain from avoiding high occupancy in one hospital is offset by the loss in increasing occupancy in another. This adds an important safety dimension to the cost-reduction benefits of capacity pooling of healthcare services.

A further important implication of the safety tipping point is that it has a marked effect on the value of semi-flexible capacity, specifically capacity with flexible medical staffing. Within the context of our data, we have argued that an increase of capacity with flexible staffing, triggered by occupancy levels, can be 40% cheaper than rigid capacity and achieve the same safety improvement in terms of mortality reduction.

We have pointed out the statistical challenges in estimating occupancy tipping points with respect to mortality, specifically the fact that avoidable mortality is a rare event that requires large samples. We had to assemble a multi-hospital department-based dataset for our analysis. There are, however, other less severe but more frequent indicators of quality deterioration, such as readmission to hospital, to operating theater or to ICU, patient falls, medication errors, or patient complaints that are routinely recorded by hospitals (KC and Terwiesch 2012, Kim et al. 2013, Long and Mathews 2013). While they are rarely associated with a patient’s death these events are indicative of poor clinical quality and are likely to be affected by occupancy levels. These events are sufficiently frequent to offer promise for department-level tipping point analyses, similar to the analysis conducted in this paper. The strength of a department-level analysis would be further enhanced by including data on day-by-day variation of staffing levels, which is currently not available in a sufficiently standardized form for multi-hospital studies. Results of such analyses could be very powerful in providing sorely needed evidence to guide the design of departmental escalation policies and process re-engineering efforts, as summarized by feedback we received from the clinical director of a large medical department: “I wonder what the position of the tipping point is related to. The tipping point is the point at which further reductions in staff are associated with worse outcomes. If we could identify what factors altered the tipping point, we might be some way to understanding how to improve outcomes with less staff – to increase efficiency. What are those factors - cultural, technological, skill mix, experience?” More research is required to answer this question.

Acknowledgments
We received valuable feedback on earlier versions of the manuscript from Carri Chan, Michael Freeman, Paul Kattuman, Stelios Kavadias, Christoph Loch, Vincent Mak, Nektarios Oraiopoulos, Nicos Savva and Christian Terwiesch, as well as participants of the 2011 Behavioral Operations Conference at Insead. We would like to thank Christian Rossbach of Activa GmbH for his support in assembling the staffing data.

References


Kane et al. 2007. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. Medical Care 45 1195–1204.


