# Separate & Concentrate: Accounting for Patient Complexity in General Hospitals

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

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

Sandra Sülz
Erasmus School of Health Policy and Management, Erasmus University Rotterdam, sulz@eshpm.eur.nl

Revised Version December 2017

Scholars have recently suggested the reorganization of general hospitals into organizationally separate divisions for routine and non-routine services to overcome operational misalignments between the two types of services. We provide empirical evidence for this proposal from a quality perspective, using over 250,000 patient records from 60 German hospitals across 39 disease segments, and focusing on in-hospital mortality as outcome. First, routine patients in the sample benefit from a high relative volume (focus) of their disease segment in their hospital, suggesting that routine services division should be organized as a set of disease-focused units (hospitals-within-hospitals). Second, after controlling for focus effects, mortality of routine patients is statistically unaffected by their hospitals volume in their disease segment, while mortality rates for complex patients are lower in hospitals that have a low volume in the patient's disease segment. This suggests that the reduced patient volume in the two separate divisions, relative to the whole hospital, will not impede quality for routine patients and may increase quality for complex patients. Finally, we provide evidence that non-routine service divisions can improve service quality for complex patients by adopting a disease-based rather than medical specialty-based departmental routing strategy for newly arriving patients. A counterfactual analysis, based on a simultaneous equations probit model that controls simultaneously for endogeneity of volume, focus, and routing suggests that the proposed reorganization could have reduced mortality in the sample by 13:43% (95% CI [6:87%; 18:95%]) for routine patients and by 11:66% (95% CI [6:13%; 16:86%]) for non-routine patients.

*Key words*: healthcare management, hospital, patient complexity, volume, focus, routing, solution shop, value-adding process, service quality, mortality

## 1. Introduction

In their seminall book *The Innovator's Prescription*, Christensen et al. (2009) call hospitals "*some of the most managerially intractable institutions in the annals of capitalism*". They identify the co-existence of two fundamentally mis-aligned operational models as a root-cause of the managerial complexity and suggest that general hospitals should be replaced by two types of organizations. One type, called "value-adding process clinic", delivers standardized, routine treatments for patients with well-diagnosed conditions at predictably high quality. The other type, called a "solution shop", organizes care for more complex patients who are non-standard because their conditions are ill-diagnosed on admission or because their morbidity patterns are too complex for the effective application of standard procedures. Solution shops concentrate on solution-finding, developing and testing hypotheses in search of the best treatment for the individual patient, while value-adding process clinics focus on solution-execution, leveraging standardization, scale, and focus.

While the theoretical arguments for an organizational separation of routine and complex services are compelling (Christensen et al. 2009, Edmondson 2012), there is scant direct empirical evidence to date in support of the proposed reorganization. Importantly, splitting a hospital into two organizationally separated divisions for routine and non-routine services will change the absolute and relative volumes of the disease groups in these two divisions, relative to the hospital as a whole. It is possible that the quality-effects associated with these volume changes counteract any operational benefits of the organizational separation. Using over 250,000 patient-level discharge records from 60 German hospitals across 39 disease groups with significant in-hospital mortallity risks, this paper provides evidence that this is not the case. In fact, we find that the effects of a hospital's absolute or relative volume in a disease group on its mortality rate in the disease group are different for routine and complex patients and that these differences support the separation model.

First, after controlling for hospital selection, we find no significant effect of a hospital's absolute volume in a disease group on mortality for routine patients with the disease, while mortality for complex patients in a disease group *increases* with a hospital's volume in that disease group, on average from 4.58% in low-volume hospitals to 6.08% in high-volume hospitals for a disease group. This suggests that splitting a hospital's volume in a disease segment across two divisions for routine and non-routine services, thereby lowering the volume of patients in the divisions relative to the hospital as a whole, will not harm routine services and may *improve* outcomes for complex patients.

Second, we find that a hospital's degree of focus on a disease group, measured as the hospital's *relative volume* of the disease group, is associated with lower mortality in that disease group. However, this positive focus effect is confined only to routine patients in a disease group; risk-adjusted mortality drops on average from 2.09% in hospitals with a low degree of disease group

focus to 1.36% in hospitals with a high degree of focus. The data do not show a significant focus-mortality effect for complex patients in a disease group. These findings suggest that the routine services division in a hospital could be beneficially organized as *focused factories* (Skinner 1974), concentrating on a narrow range of related diseases, while the non-routine solution shop may maintain a broad spectrum of integrated services across all disease groups. Solution shops are therefore well aligned with the requirements of emergency medicine. Importantly, since routine patients benefit more from their hospital's focus on their disease group than from its absolute patient volume in the disease group, disease focused value-adding process clinics can be organized at hospital level, as hospitals-within-hospitals, and do not require a large-scale regional reorganization of an entire hospital system to achieve significant shifts in total patient volumes.

Third, concentrating on the organization of solution shops themselves, we study the process of routing patients into hospital departments. Depending on their departmental organization, hospitals may adopt two distinct strategies for routing newly arriving patients to a department. The first strategy is predicated on the ubiquitous specialty-based organization of hospitals into departments that offer specific treatment options, e.g. surgery in a general surgery department or a conservative treatment for the same condition in a medicine department. The routing process therefore has to incorporate an early judgment about the best treatment for the patient at the time of admission. This is a difficult decision to make, particularly for emergency patients or patients with multiple chronic illnesses. If these complex patients are routed to the wrong department, they either receive suboptimal treatment or have to be transferred to another department at a later stage, lengthening the search process for the best treatment. An alternative strategy routes all patients within a particular disease group into the same hospital department that is equipped to provide a full range of different treatments for the condition, across requisite medical specialties. This makes hospital departments demand-focused rather than supply-focused, and avoids the need to make potentially premature treatment decisions as part of the routing decision; the search process for the best treatment is deferred to the admitting department.

We refer to these two routing strategies as specialty-based and disease-based routing. In its pure form, disease-based routing leads to the routing of an entire disease group into the same hospital department. While pure disease-based routing is rare, we find that hospitals that concentrate a larger proportion of patients in a disease group in the same department, i.e. adopt a more disease-based routing strategy, have significantly lower mortality rates for complex patients in that disease group, with average mortality across disease groups falling from 6.49% in hospitals with a low degree of disease-based routing to 4.50% in hospitals with a high degree of disease-based routing. There is no evidence in the data that a greater degree of disease-based routing affects mortality for

routine patients, for whom specialty-based routing is less error-prone. This finding suggests that disease-based routing is preferable to specialty-based routing in solution shops.

To provide an illustration of the potential magnitude of the quality impact of the hospital reorganization studied in this paper, we perform a counterfactual analysis for the patient sample. Specifically, we assme that each hospital is split into two organizationally separate divisions for routine and non-routine patients and that routine services are delivered by value-adding process clinics, organized as disease-focused hospitals-within-hospitals, while the solution shops for the remaining non-routine patients adopt an improved disease-based departmental routing strategy. After controlling for endogenous patient selection for all three independent variables (absolute volume and relative volume in a disease segment, and the hospital's routing strategy), we estimate that this reorganization of the sample hospitals would have reduced mortality rates by 13.43% (95% CI [6.87%; 18.95%]) for routine patients and between 7.79% (95% CI [3.72%, 11.68%]) and 11.66% (95% CI [6.13%, 16.86%]) for non-routine patients, depending on the degree of disease-based routing adopted by the solution shop.

## 2. Literature review

### 2.1. Volume–quality effect

The relationship between total volume of activity and performance has been the subject of a long-standing scholarly debate related to economies of scale, learning, and experience. The positive association between cumulative experience and performance is referred to as an "empirical regularity" in Huckman and Zinner (2008) and has been replicated across different settings and across different levels of analysis, from the individual to the organization (Reagans et al. 2005), with respect to both productivity and quality performance (e.g. KC and Staats 2012, Thirumalai and Sinha 2011). The health economics and medical literature complement the management literature and provide evidence of a positive association between volume and clinical quality for a variety of clinical conditions and surgical procedures (e.g. Birkmeyer et al. 2002, Gaynor et al. 2005). However, while the sign of the association is generally reported as positive, some studies report a negative association (e.g Horwitz et al. 2015). In a systematic review of the literature Mesman et al. (2015, p. 1066) conclude that "after decades of research, only a few studies focus on the circumstances under which volume and outcome show a positive association as well as the underlying mechanisms." The mechanisms and organizational factors that affect the strength of the volume–quality relationship remain poorly understood.

The management literature on learning offers insights into potential contingency factors (Pisano et al. 2001, Theokary and Ren 2011), and management scholars have recently begun to unpick the notion of experience, distinguishing between focal, related, and unrelated experience (e.g. Boh

et al. 2007, KC and Staats 2012, Schilling et al. 2003), same-task and different-task experience (Staats and Gino 2012), experience from success or failure (KC et al. 2013), and firm-specific and non-firm-specific experience (Huckman and Pisano 2006). These studies suggest that the volume effect is contingent on volume characteristics. This paper contributes to this developing contingency theory of the volume–quality effect by showing that the strength and direction of the association between volume and outcome is significantly moderated by patient complexity. In a recent hospital level study, Theokary and Ren (2011) find that greater patient volume is associated with *decreased* process quality for hospitals with high teaching intensity. These large teaching hospitals are likely to treat more complex patients. We take this study further by analysing the moderating effect of complexity at the patient level, while controlling for teaching status through hospital fixed effects. This paper's emphasis on service quality and disease group volume, rather than total hospital volume, complements recent work by Clark (2012), who studies the effect of hospital volume and focus on costs and its interaction with patient comorbidity and provides evidence that hospital volume and hospital focus are associated with lower cost but that these cost benefits are diminishing with patient comorbidity.

### 2.2. Focus–quality effect

In a multi-service firm, increased volume in a customer segment leads, ceteris paribus, to the increased relative volume of that segment within the firm's service portfolio and, therefore, to an increase in the firm's degree of focus on that segment. The question therefore arises as to what extend an estimated volume effect is attributable to focus (relative volume) rather than scale (absolute volume). Following Skinner (1974)'s seminal paper, the performance effect of focus has received much attention in the operations literature and practice. Healthcare has been a particularly fruitful context for the study of focus effects on service quality (see e.g. Clark and Huckman 2012, KC and Terwiesch 2011). However, the causal effect of focus on service quality, and specifically on mortality in the hospital context, remains debatable. While there is evidence of a positive association between a hospital's focus on a disease group and service quality, several studies have highlighted the importance of testing and correcting for endogenous selection when the effect of hospital focus is evaluated as better outcomes may be attributable to the ability of these hospitals to attract a comparatively healthier patient pool ("cherry-picking") rather than their superior operational performance (Cram et al. 2005). In fact, using the context of cardiac services in California, KC and Terwiesch (2011) demonstrate that this selection effect can fully account for any observed superior performance of disease focus at the hospital level.

In response to inconclusive evidence about the causal effect of focus on service quality, scholars have begun to develop a contingency theory of focus in service industries, trying to understand

under which conditions focus has a beneficial effect on performance and when it may be less effective. Using the context of cardiovascular services, Clark and Huckman (2012) show that the quality effect of focus is amplified by the degree to which a hospital offers services in disciplines that are related to the focal service. This paper contributes to this emerging contingency theory of focus in the hospital context by providing empirical evidence that patient complexity is an important contingency factor for the focus–quality effect: Routine patients in a disease group are more likely to benefit from an increase in their hospital's focus on their disease group than complex patients.

### 2.3. Departmental routing and gatekeeping

Departmental routing relates naturally to the operations literature on gatekeeping. Gatekeepers, such as primary care physicians or emergency doctors in the healthcare context, are typically customers' first point of contact with the firm. Gatekeepers decide whether to serve customers themselves or refer them to other, more specialized service providers within or outside the organization (Shumsky and Pinker 2003). The operations literature on gatekeeping has been largely concerned so far with factors that affect referral rates, specifically, incentivization schemes (Lee et al. 2012, Shumsky and Pinker 2003) and workload (Freeman et al. 2016). In the context of a multi-service firm, such as a general hospital, the binary decision whether or not to refer a customer to a specialist (e.g. the emergency doctor's decision to admit the patient to the hospital) is only one aspect of the gatekeeping decision. Conditional on a positive referral decision, the gatekeeper has the choice between several possible referral departments. Alizamir et al. (2013) analyze a normative model to address this challenge in the context of congested systems. We complement this modeling study with empirical evidence that highlights the relevance of patient complexity for the choice of routing strategy in the context of hospital services. Specifically, we show that a hospital's strategy to route a larger proportion of a disease group into the same department benefits complex patients more than routine patients.

## 3. Hypothesis development

The paper's hypotheses relate to five concepts – patient segment, patient complexity, segment volume, segment focus, and segment concentration – which we will now explain before we develop the hypotheses. Multi-service firms, such as general hospitals, serve multiple customer segments. These are groups of customers who share a similar primary need, defined as the customer's reason for engaging with the organization. In the hospital context, the primary need is the reason why a patient is admitted to the hospital, such as pneumonia or a hip fracture. The World Health Organization's (WHO) International Classification of Diseases (ICD) provides a global standard that allows hospital patients to be grouped by the disease identified as the reason for hospital

admission. These groups of patients with the same primary need form natural *patient segments*. We will provide more detail on this segmentation in Section 4.2.

In addition, and orthogonal to the segment classification, customers within a segment can be classified by their degree of *complexity*. From an operations management perspective, the complexity of a customer's need can be expected to increase the uncertainty about the appropriate service process for the customer. We therefore use process uncertainty as the theoretical lens to develop the hypotheses. Process uncertainty refers to the degree of incompleteness of a firm's information at the beginning of the service episode about the "what, where, when, and how" of that customer's service process (Argote 1982, Larsson and Bowen 1989).

A customer's process uncertainty depends on (i) the total amount of information required before confident service decisions for that customer can be made and (ii) how much of that information is available at the start of the customer's service episode. The total amount of required information depends on the customer's needs and is likely to vary by customer segment, reflecting average information requirements for the customer's primary need, as well as within segments, reflecting customer idiosyncrasies. In the hospital context, variation among patients within a disease segment is often caused by comorbidities, typically chronic illnesses, which are ancillary needs that are not the cause of admission but may complicate the diagnosis and treatment of the primary need (Clark 2012, Gittell 2002). Comorbidities generally increase the total amount of information required before confident service decisions can be made.

The second determinant of process uncertainty – how much of the required information is available at the start of the service episode – is captured in the hospital context by the patient's admission status. Patients whose hospital admission was planned will typically have been seen by a hospital doctor as an outpatient before admission and will have a treatment plan in place when they are admitted. Conversely, emergency patients have no such plan and therefore, ceteris paribus, more information is missing at the time of admission. Patients with the highest degree of process uncertainty are emergency patients with multiple comorbidities, while patients with the lowest degree of process uncertainty have planned hospital admissions and no comorbidities. We define these as *complex* and *routine* patients, respectively. Other patients, including emergency patients with no or low comorbidity burden and planned patients with comorbidities, have a moderate degree of process uncertainty and serve as *benchmark* patients in the empirical study.

Note that in the medical literature the term patient complexity has a narrower meaning. It refers to the number of different specialties that need to interact simultaneously during the care of a patient, e.g. a cancer patient with dementia requires an oncologist and a geriatrician. Clinical complexity is measured by the number of disease-relevant comorbidities of a patient, which is independent of the patient's admission status (emergency or elective). From an operations management

perspective, however, complexity relates to all the steps that the patient undergoes during a care process, not just the subset of simultaneously activated specialties. Take for example an otherwise healthy female patient who arrives in the emergency department with what appears to be abnormal vaginal bleeding. The emergency physician stabilizes her and refers her to the gynecology ward, where a gynecologist peforms a series of tests but cannot find a gynecological cause for the bleeding. She suspects a urinary problem and refers the patient to a urology ward, where she is treated. In this case the coordination between specialties was not required because of the patient's comorbidity pattern – there was none – but because of the uncertain cause of her symptoms. Two specialists had to coordinate in a sequential "hand-off" manner as fast as possible to resolve the uncertainty and identify approrpriate treatment. This dimension of complexity is not captured by the number of comorbidities but rather by the timely pattern of diagnostic steps. Such sequential search dynamics are characteristic of emergency services and can affect the operational complexity of patients as much as their comorbidity patterns. We therefore combine emergency status and comorbidity pattern in the definition of operational patient complexity. This approach is compatible with the differentiation between routine and complex operations in other knowledge-intensive industries (Edmondson 2012, p.32).

We study the service quality effects of three operational characteristics of a patient segment within a hospital and, specifically, how these effects vary between routine, benchmark and complex patients within the segment. The three variables of interest are: (i) the hospital's total *segment volume*, measured by the annual number of hospital admissions in the disease segment, (ii) the hospital's *segment focus*, measured by the relative annual segment volume in a disease segment as a percentage of the hospital's total annual patient volume across all segments, and (iii) the hospital's degree of *segment concentration*, measured as the percentage of patients in a disease segment routed to the hospital department that admits the majority of the hospital's patients in that segment.

### 3.1. Volume–quality effects

Most studies of volume–quality effects in the management, economics, and medical literature posit and confirm a positive association between volume and performance (e.g. Argote 1993, 2013, Gaynor et al. 2005, Pisano et al. 2001, Schilling et al. 2003). Two main mechanisms drive the volume–quality relationship. First, units that have a higher volume of activity benefit from scale economies, which reduces the average cost of outputs and, if the price is exogenous as in case of most hospital services, increases profitability and enables investment in quality that low-volume organizations will not be able to afford. Second, higher volume leads to more task repetition, which leads to more cumulative experience and thus, through individual and organizational learning, to better

performance (Huckman and Zinner 2008). Each member of the care team for a patient segment is likely to do a job better as volume increases, leading to higher quality of care in hospitals that serve a larger annual volume of patients in this segment.

However, increased volume may make coordination more difficult, which can diminish the benefits of scale (Clark 2012). We present three arguments why this coordination challenge is more pronounced for more complex patients.

First, research has shown that coordination mechanisms vary in their scalability. While impersonal (aka programmed) means of coordination, such as coordination through IT systems (e.g. electronic patient records, automated personnel scheduling) or scheduled meetings, scale up readily with increased volume of activity, this is not the case for personal and group-based (aka non-programmed) means. In a seminal study of employment security agencies Van de Ven et al. (1976, p. 331) observed that "*as unit size increases, the use of impersonal coordination increases significantly (...) while the use of horizontal channels and group meetings remains invariant with work unit size*".

The two types of coordination mechanisms – programmed and non-programmed – affect different types of patients differently. In a study of hospital emergency departments Argote (1982, p. 430) found that "*programmed means of coordination made a greater contribution to organizational effectiveness under conditions of low uncertainty than under conditions of high uncertainty. Conversely, nonprogrammed means of coordination made a greater contribution to organizational effectiveness when uncertainty was high.*" Since complex patients have a higher degree of process uncertainty than routine patients, this suggests that coordination challenges will reduce any beneficial effects of scale more markedly for more complex patients.

Second, high quality care for complex patients is likely to rely on effective informal dialogic coordination practices, where "*action, communication and cognition are essentially relational*" (Faraj and Xiao 2006). This is challenging in hospitals, where teams are often fluid and coordinated in ad-hoc groupings or "scaffolds", i.e. meso-level structures that define bounded roles and lead to shared responsibility (Valentine and Edmondson 2015). As volume increases, staff numbers increase, which leads to more fluctuation in fluid teams. As a consequence, staff need to form and maintain more relationships to achieve effective dialogic coordination. However, research in social anthropology, initiated by Dunbar (1992), suggests that humans can only maintain a limited number of stable social relationships, with estimates in the range of 100–200 relationships (see e.g. Gonçalves et al. (2011) for recent evidence based on Twitter activity). This imposes a limitation on the scalability of processes that rely on effective dialogic coordination, such as care processes for complex patients.

Our final argument is of a structural nature. Increased volume may lead to increased structural differentiation as more granular subdivisions of work tasks into different jobs and groups are put in place (e.g. Beyer and Trice 1979, Tolbert and Hall 2009). This increases the range of specific tasks and requires additional task interfaces, leading to an increased need for coordination. Drawing on Grant (1996)'s study of knowledge integration, Clark (2012, p. 88) argues that "*hospital volume might diminish the efficiency of 'wide-ranging' coordination by reducing the level of common knowledge (through finer division of labor) and by encouraging an increasingly siloed structure (through greater structural differentiation).*" Since complex patients are more likely to require access to more tasks and "wide-ranging coordination" than routine patients, coordination issues caused by structural effects are more likely to impede services for these patients.

In summary, while we follow the literature in positing a beneficial quality effect of segment volume for routine patients in the segment, we hypothesise that these benefits are less pronounced for complex patients than for routine patients.

HYPOTHESIS 1. *A hospital's total volume in a disease segment has a positive effect on service quality for routine patients in the segment. The quality benefits of volume are weaker for complex patients than for routine patients in the segment.*

### 3.2. Focus–quality effects

Huckman and Zinner (2008, p.179) point out that "*the benefits of focus (. . . ) are attributable not simply to repeating routines, but to limiting the number of different routines pursued within a site or organization*". While in many industries the narrowing of tasks, and the associated reduced organizational complexity, is an important mechanism by which focus improves performance, this conceptualization of focus does not translate easily to the context of general hospitals. As McDermott and Stock (2011) point out, the public mandate of general hospitals to provide comprehensive hospital services for their catchment population limits their ability to narrow their service spectrum. McDermott and Stock (2011) suggest that focus in hospitals is better understood as emphasis: A hospital's focus on a disease segment is the consequence of a strategic prioritization of this segment. The hospital seeks to excel in this segment by expanding resources to serve these patients. Indeed, this conceptualization of focus as emphasis is consistent with the prevalent measurement of focus in the hospital context as *relative* patient volume (see e.g. Clark and Huckman 2012, KC and Terwiesch 2011).

If we follow the literature and its conceptualization of focus as relative volume, then focus and volume are closely related. An increase in the absolute volume of a patient segment in a hospital causes, ceteris paribus, an increase in the relative volume of the segment in the hospital. Conversely, when we hold the volume of a focal segment constant, an increase in the hospital's focus on this

segment can only occur through a reduction in volume of the non-focal segments. Therefore pure focus effects are caused by spillovers from changes in patient volumes outside the focal segment. When the patient volume in other segments is reduced, there is less opportunity to get distracted by performing tasks that are unrelated to the focal segment, which is known to improve performance (Schilling et al. 2003). However, as Clark and Huckman (2012) point out, the limited volume of other patients comes at a cost as it limits the availability of complementary services that, while not needed for the more routine patients in the focus segment, may be required to serve the ancillary needs of more complex patients. In addition, as hospitals become more focused on a specific disease segment they are more likely to draw organizational boundaries around this segment, which can impede access to knowledge outside these boundaries that may be required for complex patients. We therefore expect focus to be beneficial for routine patients but that the beneficial effect is reduced for complex patients in a disease segment.

HYPOTHESIS 2. *A hospital's focus on a disease segment has a positive effect on service quality for routine patients in the segment. The quality benefits of focus are weaker for complex patients than for routine patients in the segment.*

Note that Hypothesis 2 is related to but different from Hypothesis 3 in Clark and Huckman (2012). Both are moderating hypotheses on the quality benefits of focus on a patient segment. However, while Clark and Huckman (2012) posit that a hospital's increased focus on *another* segment increases the quality benefits of focus for the average patient in the focal segment, provided the other segment is related to the focal segment, we ask which patients *within* the focal segment accrue more benefits from focus, independently of any other segment's change in focus.

### 3.3. Routing–quality effects

The process of routing customers through service systems to a server with the appropriate degree of specialization is often referred to as gatekeeping (Shumsky and Pinker 2003). In multi-service firms, such as hospitals, gatekeepers perform three dependent tasks. First, they diagnose the customer's primary need, thereby allocating them to a customer segment. Second, they decide whether to serve the customer themselves or refer her to a more specialized department within the organization. Third, if a referral is necessary, they decide which department to refer the customer to. This third aspect of the routing process – the choice of the most appropriate department – is our focus in this study.

When a gatekeeper makes this departmental choice, she will have assigned the customer to a customer segment and will know from experience which department tends to receive most of that segment's customers. Without additional customer-specific information, we can expect the gatekeeper to believe that this "default department" is the best choice for the customer, and this

belief will be stronger the greater the segment's concentration in that default department. The proportion of the firm's segment customers routed to the segment's default department – which is our segment concentration measure – is the segment's "baserate" and measures the strength of the gatekeeper's prior belief that the segment's default department is the appropriate department for a segment customer. We will argue that a higher baserate can be expected to lead to fewer routing errors and, in consequence, to better service quality, and that this beneficial effect of a high baserate is more pronounced for complex patients.

We explain our main argument in the context of a stylized but realistic gatekeeping process. Suppose the gatekeeper has identified a customer's segment and has to decide which department should serve the customer. The gatekeepers "first impression" is that the customer should be sent to the segment's default department because that serves most of these customers. However, following Hopp and Lovejoy (2013, p. 533 ff), we assume that the gatekeeper will seek confirmation of this first impression by gathering additional relevant customer-specific information through a test. If, on the one hand, the test is positive, indicating that the default department is indeed the best choice for the customer, the gatekeeper's first impression is confirmed and we can expect her to assign the customer to the default department. In this case, the probability of an assignment error (false positive) is smaller the higher the segment's routing baserate (Hopp and Lovejoy 2013, p. 531). If, on the other hand, the test is negative, indicating that the default department is inappropriate for the patient, the gatekeeper has two conflicting pieces of evidence: her first impression, indicating that the customer should be referred to the default department, and a negative test result, indicating that he shouldn't. The higher the baserate, the more surprising the negative test result will be for the gatekeeper, and the more surprising the new evidence is, the more likely it is that the gatekeeper will conduct additional tests before deviating from the first impression (Hopp and Lovejoy (2013, p. 533), Rabin and Schrag (1999)). This behavior is to be expected as a consequence of confirmation bias, the human tendency to "*accept confirming evidence at face value while scrutinizing dis-confirming evidence hypercritically*" (Lord et al. 1979, p. 2099). Therefore, the higher a segment's routing baserate the larger the likelihood of additional testing after a negative first test and therefore the smaller the likelihood of a departmental allocation error following a negative test (false negative). In summary, both false positive and false negative errors decrease with increasing baserate.

Alizamir et al. (2013) confirm that this positive effect of high baserates on the accuracy of gatekeeping decisions is maintained in the context of a congested system, where diagnostic accuracy must be traded off against increased congestion as more testing leads to longer service times. They endogenize the testing decision in a model and show that the gatekeeper's optimal policy is to continue testing so long as the current belief falls into a specific interval, which depends on the

state of the system at the time of the decision. For a fixed system state, this interval widens with an increase in the a priori probability of a default customer type, which equates in our context to the segment's baserate. Therefore, when the system is in a specific state and the available information is inconclusive, an optimizing gatekeeper will conduct more testing the higher the baserate. The underlying intuition is that a gatekeeper faced with inconclusive evidence is able to spend more time with the customer in hand when the next customer has a higher a priori probability of a clear-cut decision and is therefore less likely to require a long diagnostic process. In the hospital context, we expect this beneficial effect to be more pronounced for complex patients, who require more time for an accurate departmental allocation in the first place. In the interest of brevity, we refrain from formulating and testing a hypothesis related to transfer errors in this paper and refer the interested reader to the supplementary material (Anonymized 2016, Chapter 7.3, Table 14), where we present empirical support that (i) complex patients whose hospitals have a higher degree of departmental concentration for their segment experience fewer in-hospital departmental transfers within the first seven days of their hospital stay and (ii) that this beneficial effect of departmental concentration deteriorates with the complexity of the patient.

Fewer departmental routing errors translate directly into quality benefits. Indeed, if a customer is not allocated to the best department for her needs, then she will either remain in that department and receive suboptimal, lower-quality service than in the correct department or the error will be corrected by transferring her to the correct department, leading to delayed treatment. Emergency patients are particularly vulnerable to delays of treatment, while this is less likely to be quality-critical for the non-urgent patients. Transfers themselves may also result in poor service quality as they require customer-specific information to be passed from the referring to the admitting department – a process that must be coordinated and is prone to ambiguity and misinterpretation. Indeed, problem-solving processes depend on an adequate information flow, which has been shown to be disrupted by handoffs, leading to poorer performance (Rathnam et al. 1995). In the service setting, the consequences of communication breakdowns and information loss have received considerable attention, especially in the healthcare domain, where "*handoffs during (...) transition of care is a point of vulnerability*" (Ong and Coiera 2011, p. 283). The service deterioration effect of handoffs becomes more pronounced the more information needs to be passed between departments. We therefore expect information loss to be more likely for complex patients than for routine patients since the patients' ancillary needs require more information to be passed on. In addition, our complex patients are emergency admissions and will often require a fast response, making quality effect of service delays following a departmental transfer even more detrimental. In summary, as increased segment concentration reduces departmental allocation errors, service quality should increase when segment concentration is higher, and this beneficial effect should be more pronounced for complex patients.

Hypothesis 3. *A hospital's degree of disease segment concentration has a positive effect for complex patients in the segment. The quality benfits of disease segment concentration are weaker for routine patients than for complex patients in the segment.*

This hypothesis is further supported by organizational arguments. A referral to the segment's default department is essentially a delegation of decisions affecting service delivery to the departmental level – the gatekeeper no longer needs to determine the type of service the customer will receive, as would be the case if she deviates from the default decision. In taking the default decision, the gatekeeper therefore exploits the "*well-known rule of thumb that a decision should be delegated to the lowest level that has the information necessary to make the decision*" (Mihm et al. 2010, p. 843). When a larger proportion of segment patients are allocated to the same default department, it is more likely that the relevant problem-specific knowledge and requisite range of competencies required for a disease segment are available in that department. Therefore, the more concentrated the segment, the more likely it is that the default department will assume "*natural lead function(s), which then (possibly unintentionally) contribute to faster search and solution quality*" (Mihm et al. 2010, p. 842).

Higher segment concentration also means that once decisions concerning the required service tasks are made, the interdependencies among service tasks are more likely to reside within the default department, reducing reliance on cross-departmental collaboration and increasing solution quality further. This departmental concentration of service tasks provides communication benefits among employees due to collocation (Gray et al. 2015) and reduced team dispersion (Bardhan et al. 2013). Both the search for the best solution and the interdependencies between the service tasks required to execute the solution are likely to be more complex and require more collaboration for complex patients. As such, these patients are most likely to benefit from high segment concentration.

# 4. Empirical study

## 4.1. Setting

We use patient-level discharge records from German hospitals to test our hypotheses. This setting has two important advantages. First, as mentioned in Section 3, the hospital industry employs an internationally accepted disease classification system – the WHO's ICD system – which is coded in discharge records and allows us to define meaningful industry-wide patient segments based on the reason for a patient's hospitalization. Second, in contrast to the US or UK, German hospitals have a nationally standardized clinical department classification and their patient discharge records include standardized codes for the departments patients are admitted to during their hospital stay. This allows us to measure the degree of segment concentration in a hospital.

The regulatory framework and organizational structure of German hospitals is fairly homogeneous, beginning with the top management team, which is composed of a commercial director, medical director and nursing director with clearly defined roles and responsibilities for the hospital as a whole. Inpatient services are managed at the level of specialist departments, led by a powerful clinical director – the *Chefarzt* – who is the superior of all doctors in that department, is responsible for safety and clinical outcomes for all patients in that department, and has full budgetary responsibility for that department. Patients access hospitals either through the emergency department or are referred directly to a specialist department by a primary care doctor or independent practicing specialist. Health insurance is compulsory in Germany – about 90% of the population is insured via public sickness funds and 10% through private insurance – and access to hospitals and hospital reimbursement for inpatient services is independent of a patient's insurance type.

### 4.2. Segmentation

The WHO's ICD represents the culmination of over a century of global efforts to classify diseases and causes of death for statistical morbidity and mortality studies. These codes serve as "*the international standard diagnostic classification for all general epidemiological and many health management purposes*" (Gersenovic 1995, p. 172). ICD-10, the tenth and most current revision of the classification scheme, contains several thousand codes for diseases, signs, and symptoms and other circumstances that may cause a patient to be admitted to a hospital. Importantly, these codes are organized into a hierarchy. At the highest level there are 22 ICD chapters, which group diseases by the body's main biological systems, such as "diseases of the respiratory system," or by general disease pathways, such as "infectious and parasitic diseases." Each chapter is divided into ICD blocks, of which there are 263 in total. These blocks group more specific disease areas within chapters, such as "influenza and pneumonia" within the chapter "diseases of the respiratory system." Each ICD block is subdivided into three-digit codes, e.g. J13, "pneumonia due to streptococcus pneumoniae," totaling more than 2,000 conditions that patients may present with in hospitals, and is further subdivided into four- and five-digit codes. Each level in the ICD hierarchy groups patients into segments with related clinical needs. We chose the 263 ICD blocks as the disease segments for our study. While a lower ICD hierarchy level would provide a higher degree of homogeneity of the clinical needs of patients in a segment, discussions with clinicians suggested that while coding at the level of ICD blocks is generally accepted as reliable, coding accuracy is likely to deteriorate rapidly at a more granular coding level.

### 4.3. Dependent variable and data sample

The unit of observation in our study is a patient episode in a hospital, from admission to discharge. In line with other service quality studies in hospitals, we use in-hospital mortality as a binary

indicator of clinical quality (Clark and Huckman 2012, Cram et al. 2005, KC and Terwiesch 2011). Specifically, following Kuntz et al. (2015), we report results on patient mortality during the seven days following hospital admission, which comprises the critical period of most hospital stays. The results are qualitatively robust over other fixed observation windows and for total in-hospital mortality.

The data for this study are derived from three sources. First, we employ a database of standardized administrative discharge records for all patients discharged from 95 German hospitals over a fixed observation period, totaling 942,296 patient episodes. For 66 hospitals, the observation period spans 12 months, either from January 1, 2004 to December 31, 2004 or January 1, 2005 to December 31, 2005. For the remaining 29 hospitals, discharge records are available for 24 months from January 1, 2004 to December 31, 2005. The data contain a wealth of clinical and operational information, which we use to define the independent variables and patient-level control structure for our models. Second, we use a database of Hospital Quality Reports for 2006 (Gemeinsamer Bundesausschuss 2016). This database extends beyond our sample hospitals, covering 82% of all German general hospitals in 2006, and therefore also provides information about out-of-sample hospitals, which we use to construct instrumental variables (IVs) that allow us to address concerns about endogenous hospital selection. Third, to allow us to control for socioeconomic factors we use data from the Federal Office for Building and Regional Planning (Bundesamt für Bauwesen und Raumordnung 2012), which we match to discharge records using the patient's postal code.

We exclude patients with missing or invalid postal code information, which we need to construct instrumental variables. In order to increase the homogeneity of the sample, we further exclude potential outlier hospitals and clinical departments – small hospitals with fewer than 1,000 discharges per year and small clinical departments with fewer than 50 discharges per year – as well as hospitals with no emergency admissions (e.g. rehabilitation clinics) and all pediatric services. Finally, to increase the homogeneity of the sample for a mortality study, we exclude disease segments with very low mortality risk. We report results for the sample excluding disease segments with a seven-day in-hospital mortality rate below 1%. The results are qualitatively robust over a range of cut-off values. After these exclusions, the final sample comprises 265,133 patients across 60 hospitals and 39 disease segments. Table 1 in Chapter 1 of the supplementary material (Anonymized 2016) shows the chosen segments and provides descriptive statistics for this sample.

### 4.4. Independent variables: Segment volume, segment focus, and segment concentration

A hospital's segment volume is its annualized number of admissions of patients in that segment. Since typical numbers of hospital admissions vary considerably between segments, a model coefficient of the raw volume variable would not capture the average effect of volume variation across

segments; what is a low hospital volume in one segment may be beyond the maximal hospital volume in another. In a first step, we therefore standardize segment volumes by computing the z-scores for each segment across the hospitals that admit patients of the segment.

The medical literature provides robust evidence that the relationship between volume and mortality is nonlinear and best approximated by a threshold model, where increased volume reduces mortality only up to a point, after which there is no additional benefit. A comprehensive recent study concludes that "*Admission to higher-volume hospitals was associated with a reduction in mortality for acute myocardial infarction, heart failure, and pneumonia, although there was a volume threshold above which an increased condition-specific hospital volume was no longer significantly associated with reduced mortality*" (Ross et al. 2010, p. 1110). Such threshold phenomena can be naturally modeled through the use of linear splines (see e.g. Kuntz et al. 2015). In our case, however, the operationalization of a spline model is complicated because we are interested in moderation effects with patient complexity. The spline knots (the thresholds) as well as the slopes of the spline pieces are likely to differ by segment as well as by patient complexity, making the resulting model specification cumbersome and estimated effects difficult to compare between complexity levels. We therefore opt for a simpler approach and dichotomize the volume variable by categorizing hospitals into two groups for each segment, using a median split: We rank hospitals by volume of patients in the segment and categorize the bottom half as low-volume hospitals and the remaining hospitals as high-volume hospitals for the segment.

While dichotomization is often criticized for the associated loss of information and statistical power, we believe that its benefits outweigh the drawbacks in our context. First, the literature that is critical of dichotomization is largely based on simulation studies that compare the performance between linear and dichotomized models in the case when the linear model is the correct specification of the conditional mean function. In our case, however, the correct model is nonlinear and the linear model itself would be misspecified. An example in the supplementary material (Anonymized 2016, Section 4) illustrates that dichotomization may well be preferable in this situation. Second, MacCallum et al. (2002), while arguing against dichotomization in general, confirm that dichotomization is justified when *"there is a large number of observations at the most extreme score on the distribution. (...) Such a distribution indicates the presence of two groups."* (MacCallum et al. 2002, p.38). Several of the volume distributions of the segments in our study have that property, as illustrated in Figure 1. Hospitals with very low volumes at the extreme left of the distribution are likely to have hospital volumes below the afore-mentioned mortality-volume threshold, while hospitals in the right tail of the distribution are likely to be above that threshold. This leads to a natural underlying categorization which supports our choice to dichotomize the volume variable. Finally, dichotomization of the independent variables has the additional advantage

that it allows us to calculate instrumental variables based on the patient's proximity to a hospital that is categorized as high-volume (high-focus, high-concentration) for their segment (McClellan et al. 1994). In light of these arguments, we present the results for a specification with dichotomized independent variables. We report the results of a linear specification with the continuous z-scores in the supplementary material (Anonymized 2016).

We stress that dichotomization only allows us to gain insights into the differences between low-volume and high-volume hospitals within a segment and that our results should not be used to assess the effects of small volume changes. In the motivational context of our study, however, this is not a severe limitation. Any reorganization of general hospitals into value adding process clinics for routine patients and solution shops for complex patients in a patient segment will naturally require large changes of segment volumes in hospitals.

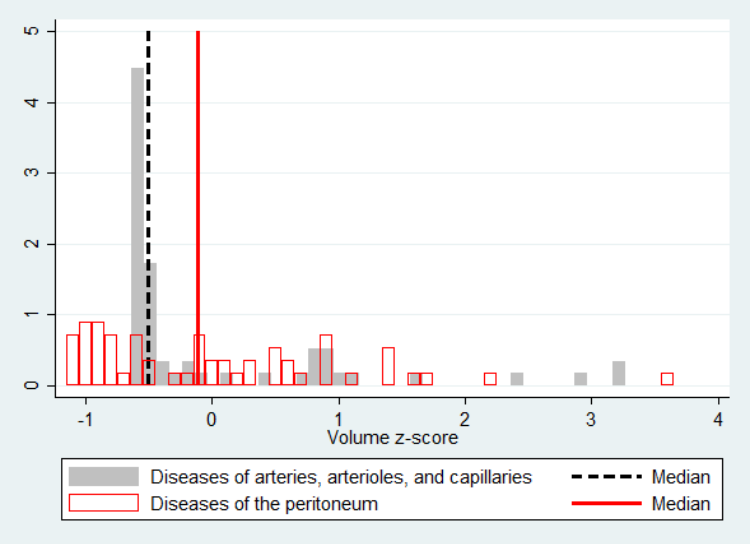


**Figure 1 Distribution of hospital volume (z-score) for two selected segments**

For similar reasons, and to stay consistent, we dichotomize the other two independent variables (focus and segment concentration) as well. Specifically, following the literature (Clark and Huckman 2012, KC and Terwiesch 2011, McDermott and Stock 2011), we first calculate a hospital's degree of focus on a fixed segment as its *relative* volume in that segment, measured as the number of admissions in the segment as a percentage of total hospital admissions over the observation period and then convert this continuous variable into a binary variable via a median dichotomization analogously to the volume variable. For the hospital's degree of segment concentration, we first identify the hospital's *default department* for a given segment $s$ as the department that admits the largest proportion of the hospital's patients in segment $s$. We then calculate the hospital's degree of concentration for segment $s$ as the proportion of segment $s$ patients routed to the hospital's

default department for segment $s$, and finally convert this continuous variable into a binary variable via a median split across hospitals $h$ that treat patients in segment $s$. We refer to Section 3 of the supplementary material Anonymized (2016) for more details on the calculation of the three dichotomous independent variables.

Note that the binary independent variables $V_{ish}, F_{ish}, C_{ish}$ for volume, focus and segment concentration, respectively, vary only at the level of segments-in-hospitals $(s, h)$ and are constant for all patients $i$ within a fixed segment in a fixed hospital. We interpret these binary variables as treatment variables for patients and are interested in the average effect of these treatments on a patient's mortality risk.

### 4.5. Moderator variable

Our hypotheses posit differing volume-outcome, focus-outcome and concentration-outcome effects for routine and complex patients within a disease segment. We therefore group the patients within a disease segment into three categories: routine, benchmark, and complex patients. As explained in Section 3, we use the admission type – emergency or elective (i.e. planned) – and the patient's comorbidity burden to classify patients by their degree of complexity. Following the medical literature, we use the list of so-called Elixhauser comorbidities, which are known to be associated with elevated mortality risk (Elixhauser et al. 1998) and available at patient-level in the discharge records. We classify a patient as routine if she has a planned admission and no Elixhauser comorbidities and complex if she is an emergency admission with multiple Elixhauser comorbidities. We present the results for the case where we require at least three comorbidities for a patient to be classified as complex; the results are robust to other cut-offs choices. The remaining patients are either elective patients with Elixhauser comorbidities or emergency patients with at most two Elixhauser comorbidities and are referred to as "benchmark" patients. We use the binary variables $PR_{ish}$ and $PC_{ish}$ to indicate that patient $i$ in segment $s$ and hospital $h$ is a routine or complex patient, respectively.

While all routine patients are electives (comprising 66% of all elective patients in our sample) and all complex patients are emergencies (comprising 38% of all emergency patients), the admission status is well balanced among the benchmark patients in our sample (with 51% emergency patients).

### 4.6. Instrumental variables

We cannot assume that patients are randomly assigned to hospitals. As such, any beneficial effect of volume, focus or segment concentration on mortality may be the result of selection bias of patients with an unobserved lower mortality risk (e.g. Gaynor et al. 2005, KC and Terwiesch 2011). To test and correct for such endogeneity, instrumental variable (IV) methods are required. IVs are significant predictors of a patient's hospital choice (e.g. a high-volume or low-volume hospital for

her disease segment) but are uncorrelated with the error term in the probit mortality model. We use two types of IVs for each of the three endogenous variables, i.e. a total of six IVs.

Our first type of IV is the differential distance (DD) variable used in McClellan et al. (1994). In the case of volume, the IV is calculated as the difference between a patient's distance to the nearest high-volume hospital for her patient segment $s$ and the distance to the nearest hospital that treats patients in segment $s$, independently of the segment volume in the hospital. The variable is zero if the nearest hospital is a high-volume hospital and, otherwise, is a measure of the in for the patient of choosing a more distant high-volume hospital over her nearest hospital. While our main database provides information on the patient's place of residence, the data cover only a sample of hospitals and not the total hospital population in Germany. We use information from the Hospital Quality Reports database (Section 4.3) to calculate this differential distance and refer the interested reader to Chapter 4 of the supplementary material (Anonymized 2016) for details of this calculation. The second instrument is a set of binary variables, based on the idea that a patient's propensity to be admitted to a high-volume hospital is the higher the more high-volume hospitals there are in the vicinity of her place of residence (KC and Terwiesch 2011). We operationalize this idea with a set of $K$ binary variables $D_{ik}$ ($k \in \{1, \ldots, K\}$), where $D_{ik} = 1$ if the $k$-th nearest hospital to patient $i$ is a high-volume hospital for patient $i$'s segment and zero otherwise. The results are robust with respect to the number of binary variables $K$ that comprise the instrument, and we report results for $K = 5$. Both sets of IVs for focus and concentration are calculated analogously, with high-volume hospitals replaced by high-focus and high-concentration hospitals, respectively. Note that our IVs are based on convenience of access, i.e. a patient is more likely to be admitted to a high volume hospital if there is a high volume hospital for her segment close to her residency. Differences in access to high hospitals, however, are also likely correlated with regional factors. Patients in rural regions, for example, are less likely to have access to high volume hospitals and may also have, on average, worse health outcomes and therefore higher mortality risks when admitted to hospitals. We control for such differences in our model through regional and socio-economic controls. We refer the reader to Tables 3-5 in the supplementary material (Anonymized 2016) for further details and some descriptive evidence that supports the validity of the exclusion restriction for the IVs.

### 4.7. Control variables.

We control for patient- and hospital-level factors in our econometric models. At the patient level, we control for the patient's disease segment via segment fixed effects, age, age squared and gender, the presence of each of the 31 Elixhauser comorbidities, the patient's admission day of the week and admission month (January 2004 to December 2005), and socioeconomic factors (population density, employment rate, gross domestic product per capita, number of residents per physician) for the patient's county and federal state of residence (Cram et al. 2005). At the hospital level, we control for heterogeneity between the hospitals by means of hospital fixed effects.

### 4.8. Econometric models

**4.8.1. Cluster-specific probit models.** We estimate models at the patient level to allow us to control for patient-specific characteristics. However, in doing so, we have to account for the hierarchy in our data, with patients clustered within segments in hospitals. To this end, we make use of a cluster-specific probit model as explained in Wooldridge (2010, Chapter 15). In this section, we first introduce the standard model and then its endogeneity-corrected version.

In common with the standard probit model, the cluster-specific probit model assumes that death within seven days of admission for patient $i$ in segment $s$ and hospital $h$, indicated by a binary variable $D_{ish}$, is linked to an unobserved health index $D^*_{ish}$ for that patient. This index is modeled as a linear function of three sets of variables:

- The independent variables of interest are the dichotomous variables $V_{ish}$, $F_{ish}$ and $C_{ish}$, indicating that patient $i$'s hospital $h$ is a high-volume, high-focus or high-concentration hospital for her segment $s$.
- The moderators of interest are the dichotomous variables $PR_{ish}$ and $PC_{ish}$, indicating that patient $i$ in segment $s$ and hospital $h$ is a routine or complex patient, respectively (note that there is a third category of benchmark patients which serves as an omitted category).
- The vector $X_{ish}$ contains the control variables for patient $i$ in segment $s$ and hospital $h$, including the segment and hospital fixed effects.

In contrast to the standard model, the cluster-specific probit model makes the assumption that the error term is composed of an unobserved cluster-specific effect $\nu_{sh}$ for the segment-within-hospital $(s,h)$ and an idiosyncratic error term $\epsilon_{ish}$ for patient $i$. Including the interactions that are required to test our hypotheses, the model therefore takes the latent variable form

$$\begin{aligned} D^*_{ish} &= \alpha + V_{ish}\beta_V + F_{ish}\beta_F + C_{ish}\beta_C + PR_{ish}\beta_{PR} + PC_{ish}\beta_{PC} + V_{ish}PR_{ish}\beta_{VPR} + V_{ish}PC_{ish}\beta_{VPC} + \\ &\quad F_{ish}PR_{ish}\beta_{FPR} + F_{ish}PC_{ish}\beta_{FPC} + C_{ish}PR_{ish}\beta_{CPR} + C_{ish}PC_{ish}\beta_{CPC} + X_{ish}\beta_X + \nu_{sh} + \epsilon_{ish} \quad (1) \\ D_{ish} &= 1[D^*_{ish} > 0], \end{aligned}$$

where $1_{[.]}$ is the indicator function and $D_{ish}$ indicates death of patient $i$ in segment $s$ and hospital $h$. Assuming, as in the standard probit model, that the idiosyncratic error terms $\epsilon_{ish}$ are sampled independently from a standard normal distribution, i.e. $\epsilon_{ish} \mid V_{ish}, F_{ish}, C_{ish}, PR_{ish}, PC_{ish}, X_{ish}, \nu_{sh} \sim N(0,1)$, turns (1) into the probability model

$$\begin{aligned} P(D_{ish} = 1 | V_{ish}, F_{ish}, C_{ish}, PR_{ish}, PC_{ish}, X_{ish}, \nu_{sh}) = \\ \Phi(\alpha + V_{ish}\beta_V + F_{ish}\beta_F + C_{ish}\beta_C + PR_{ish}\beta_{PR} + PC_{ish}\beta_{PC} + V_{ish}PR_{ish}\beta_{VPR} + \quad (2) \\ V_{ish}PC_{ish}\beta_{VPC} + F_{ish}PR_{ish}\beta_{FPR} + C_{ish}PC_{ish}\beta_{CPC} + X_{ish}\beta_X + \nu_{sh}), \end{aligned}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution. We cannot estimate the cluster-specific effects $\nu_{sh}$ because the independent variables of interest $V_{ish}$, $F_{ish}$, and $C_{ish}$ vary only at the cluster level, making it impossible to identify $\beta_V$, $\beta_F$, $\beta_C$ together with $\nu_{sh}$. Instead, following Wooldridge (2010, p. 612–613), we assume that $\nu_{sh}$ is independent of $V_{ish}, F_{ish}, C_{ish}, PR_{ish}, PC_{ish}, X_{ish}, \epsilon_{ish}$ and is sampled from a normal distribution with mean zero and variance $\tau^2$. With this assumption, $\nu_{sh} + \epsilon_{ish}$ is independent of the covariates and normally distributed with zero mean and variance $\sigma^2 = \tau^2 + 1$ and the latent variable model (1) turns into the probability model

$$
\begin{aligned}
P(D_{ish} = 1 | V_{ish}, F_{ish}, C_{ish}, PR_{ish}, PC_{ish}, X_{ish}) = \qquad (3)\\
P(\nu_{sh} + \epsilon_{ish} > -\alpha - V_{ish}\beta_V - F_{ish}\beta_F - C_{ish}\beta_C - PR_{ish}\beta_{PR} - PC_{ish}\beta_{PC} - V_{ish}PR_{ish}\beta_{VPR} -\\
V_{ish}PC_{ish}\beta_{VPC} - F_{ish}PR_{ish}\beta_{FPR} - F_{ish}PC_{ish}\beta_{FPC} - C_{ish}PR_{ish}\beta_{CPR} - C_{ish}PC_{ish}\beta_{CPC} - X_{ish}\beta_X)\\
= \Phi(\tilde{\alpha} + V_{ish}\tilde{\beta}_V + F_{ish}\tilde{\beta}_F + C_{ish}\tilde{\beta}_C + PR_{ish}\tilde{\beta}_{PR} + PC_{ish}\tilde{\beta}_{PC} + V_{ish}PR_{ish}\tilde{\beta}_{VPR} + V_{ish}PC_{ish}\tilde{\beta}_{VPC} +\\
F_{ish}PR_{ish}\tilde{\beta}_{FPR} + F_{ish}PC_{ish}\tilde{\beta}_{FPC} + C_{ish}PR_{ish}\tilde{\beta}_{CPR} + C_{ish}PC_{ish}\tilde{\beta}_{CPC} + X_{ish}\tilde{\beta}_X),
\end{aligned}
$$

with scaled versions $\tilde{\alpha} = \frac{\alpha}{\sigma}, \tilde{\beta}_V = \frac{\beta_V}{\sigma}, \ldots, \tilde{\beta}_X = \frac{\beta_X}{\sigma}$ of the parameters (Wooldridge 2010, pp. 583-584). Standard probit estimation provides consistent estimates of the parameters $\tilde{\alpha}, \tilde{\beta}_V, \ldots, \tilde{\beta}_X$, and standard errors clustering at the level of segments in hospitals $(s, h)$ corrects for the dependence of the error terms due to the composite structure $\nu_{sh} + \epsilon_{ish}$ in the population model (1). While we cannot recover the parameters $\alpha, \beta_V, \ldots, \beta_X$ of model (2) because we cannot estimate $\sigma$, the fact that $0 < \frac{1}{\sigma} = \frac{1}{\sqrt{1+\tau^2}} < 1$ implies that the probit inference on the sign of $\tilde{\alpha}, \tilde{\beta}_V, \ldots, \tilde{\beta}_X$ remains valid for $\alpha, \beta_V \ldots, \beta_X$. In addition, average partial effect estimates based on the scaled parameters will be consistent (Wooldridge 2010, p. 584).

**4.8.2. Endogeneity-controlled models.** As we have seen, statistical inference and average partial effect estimations for the cluster-specific probit model are possible with a standard probit procedure. However, studies on the quality effects of volume and focus in hospitals have rightly raised concerns about endogenous hospital selection (e.g. Cram et al. 2005, Gaynor et al. 2005, KC and Terwiesch 2011). To alleviate endogeneity concerns for the independent variables of interest, we estimate a recursive equation system (Roodman 2011). The equation system consists of three probit selection equations that model whether the patient is assigned to a high- or low-volume hospital for their disease segment (selection equation 1), whether the patient is assigned to a high- or low-focus hospital for their disease segment (selection equation 2), or whether the patient is assigned to a high- or low segment concentration hospital for their disease segment (selection equation 3). As a fourth component, the system contains an outcome equation that estimates the

effect of this assignment on the patient's mortality risk, with recursive regressors $V_{ish}$, $F_{ish}$, and $C_{ish}$. Unobserved factors that may affect both hospital selection and outcomes are included by way of the correlation matrix of the multivariate error term of the four equations, which is an additional parameter of the model. A likelihood ratio test of zero correlation between the selection equations and the outcome equation can then be used as a Hausman endogeneity test (Knapp and Seaks 1998). Formally, the model is specified as follows:

$$\begin{aligned}
V^*_{ish} &= \xi^V + PR_{ish}\eta^V_{PR} + PC_{ish}\eta^V_{PC} + Y_{ish}\eta^V_X + Z^V_{ish}\eta^V_Z + \epsilon^V_{ish},\\
V_{ish} &= 1[V^*_{ish} > 0]\\
\\
F^*_{ish} &= \xi^F + PR_{ish}\eta^F_{PR} + PC_{ish}\eta^F_{PC} + Y_{ish}\eta^F_X + Z^F_{ish}\eta^F_Z + \epsilon^F_{ish},\\
F_{ish} &= 1[F^*_{ish} > 0]\\
\\
C^*_{ish} &= \xi^C + PR_{ish}\eta^C_{PR} + PC_{ish}\eta^C_{PC} + Y_{ish}\eta^C_X + Z^C_{ish}\eta^C_Z + \epsilon^C_{ish},\\
C_{ish} &= 1[C^*_{ish} > 0]\\
\\
D^*_{ish} &= \psi + V_{ish}\gamma_V + F_{ish}\gamma_F + C_{ish}\gamma_C + PR_{ish}\gamma_{PR} + PC_{ish}\gamma_{PC} + V_{ish}PR_{ish}\gamma_{VPR} + V_{ish}PC_{ish}\gamma_{VPC} +\\
&\quad F_{ish}PR_{ish}\gamma_{FPR} + F_{ish}PC_{ish}\gamma_{FPC} + C_{ish}PR_{ish}\gamma_{CPR} + C_{ish}PC_{ish}\gamma_{CPC} + X_{ish}\gamma_X + \epsilon^D_{ish}\\
D_{ish} &= 1[D^*_{ish} > 0],
\end{aligned} \tag{4}$$

where the errors $(\epsilon^V_{ish}, \epsilon^F_{ish}, \epsilon^C_{ish}, \epsilon^D_{ish})$ are sampled from a multivariate standard normal distribution

$$\begin{pmatrix} \epsilon^V_{ish} \\ \epsilon^F_{ish} \\ \epsilon^C_{ish} \\ \epsilon^D_{ish} \end{pmatrix} \sim N\left[ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{VF} & \rho_{VC} & \rho_{VD} \\ \rho_{VF} & 1 & \rho_{FC} & \rho_{FD} \\ \rho_{VC} & \rho_{FC} & 1 & \rho_{CD} \\ \rho_{VD} & \rho_{FD} & \rho_{CD} & 1 \end{pmatrix} \right]. \tag{5}$$

and the vector $Y_{ish}$ denotes all control variables affecting the patient's hospital choice (i.e. the hospital fixed effects are excluded).

Following the convention in recursive simultaneous equation models, we improve the robustness of the estimations by including instruments $Z_{ish}$ in the selection equations. Since our equation system is recursive and fully observed (i.e. the endogeneous variables appear on the right-hand side of the outcome equation only as observed and not in their latent form) we can use the full information maximum likelihood estimator implemented in the user-written STATA command *cmp* (Roodman 2011).

In our context, maximum likelihood estimation of a simultaneous equations model is more suitable than two-stage procedures for several reasons. First, our hypotheses require us to identify the coefficients of interaction terms, which are identified in our models whenever the coefficients of the direct effects in the system without interaction terms are identified (Wooldridge 2010, p.266 and p.596). We therefore do not need additional instrumental variables for the identification of interaction coefficients. Second, the approach allows us to incorporate all three endogenous variables simultaneously in the mortality equation. Specifically, patient selection of a hospital on a specific

variable does not only lead to a selection bias of the coefficient of this variable but also of the coefficient of correlated independent endogeneous variables. For example, if patients who are sicker in unobserved ways prefer higher volume hospitals for their segment and segment volume and segment focus are positively correlated, then sicker patients become more likely in highly focused hospitals, not because they chose focus hospitals but because focus is positively correlated with volume. The simultaneous equations model controls for unobserved factors that affect the correlations between the three endogenous variables via the error correlation coefficients $\rho_{VF}, \rho_{VC}$ and $\rho_{FC}$. Lastly, in the case of binary dependent variables multivariate probit models are generally preferable to two-stage procedures because the latter do not provide consistent estimators of the parameters of the treatment effects. Estimation by simultaneous equation probit models outperforms two-stage methods and is more robust to deviations from multivariate normality assumptions (Bhattacharya et al. 2006).

## 5. Results

Table 1 reports the estimation results of a single-equation probit model and of the endogeneity controlled simultaneous equations model (4), using the *cmp* command in STATA 14. The upper panel reports coefficient estimates for the binary independent variables, indicating that a patient was admitted to a high-volume (Vol), high-focus (Foc) or high-concentration (Con) hospital for her disease segment, as well as coefficients for their interactions with routine (PR) and complex (PC) patient types. The direct effects of the patient types and the control variables described in Section 4.7 are included in all models but are not reported. The second panel reports the coefficients of the IVs in the selection equations for the model (4), with $DD_V, DD_F$ and $DD_C$ referring to the three differential distance instruments, one for each selection equation, and $Dk_V, Dk_F$ and $Dk_C$ referring to the three sets of binary variables indicating that the patient's k-th nearest hospital is a high-volume, high-focus or high-concentration hospital, respectively, for her disease segment (see Section 4.6). The third panel reports the estimated correlation coefficient of the error terms in (4). The final three panels report the test statistic and significance for the effects of volume, focus and concentration for routine and complex patients, respectively, as well as the differential effects.

We find clear evidence for endogeneity. There is a significant unobserved negative correlation between mortality and volume ($\rho_{VD} = -0.109, p < 0.001$), suggesting that patients in high volume hospitals are less severely ill on unobserved factors. We also find a weakly significant positive correlation between mortality and concentration ($\rho_{CD} = 0.068, p < 0.1$). Moreover, all correlations between the endogenous independent variables are significantly different from zero. Note that for each selection equation at least one type of instrumental variable is highly significant. Since we only need one instrument per equation, the lack of significance of the other type of instrument is not a

**Table 1 Probit and simultaneous equations models**

| Mortality equation | Probit | Simultaneous equations model | | |
|---|---|---|---|---|
| Vol | −0.106** | 0.035 | | |
| | (0.034) | (0.056) | | |
| Vol * PR | 0.011 | 0.016 | | |
| | (0.041) | (0.041) | | |
| Vol * PC | 0.129** | 0.127** | | |
| | (0.041) | (0.041) | | |
| Foc | −0.057* | −0.007 | | |
| | (0.023) | (0.053) | | |
| Foc * PR | −0.196*** | −0.189*** | | |
| | (0.037) | (0.037) | | |
| Foc * PC | 0.038 | 0.038 | | |
| | (0.033) | (0.033) | | |
| Con | −0.056* | −0.144* | | |
| | (0.026) | (0.072) | | |
| Con * PR | 0.094* | 0.089* | | |
| | (0.039) | (0.039) | | |
| Con * PC | −0.071* | −0.068* | | |
| | (0.034) | (0.034) | | |
| Selection equations (IVs) | | | Vol | Foc | Con |
| $DD_V, DD_F, DD_C$ | | | −0.023*** | −0.003 | −0.002 |
| | | | (0.005) | (0.006) | (0.005) |
| $D1_V, D1_F, D1_C$ | | | 0.460*** | 0.603*** | 0.323*** |
| | | | (0.080) | (0.096) | (0.085) |
| $D2_V, D2_F, D2_C$ | | | 0.127* | 0.225** | 0.144* |
| | | | (0.063) | (0.074) | (0.064) |
| $D3_V, D3_F, D3_C$ | | | 0.134* | 0.115 | 0.165** |
| | | | (0.057) | (0.077) | (0.061) |
| $D4_V, D4_F, D4_C$ | | | −0.061 | −0.069 | 0.147* |
| | | | (0.053) | (0.064) | (0.067) |
| $D5_V, D5_F, D5_C$ | | | 0.020 | −0.024 | −0.018 |
| | | | (0.053) | (0.074) | (0.060) |
| Error correlations | | | | | |
| $\rho_{VD}, \rho_{FD}, \rho_{CD}$ | | | −0.109*** | −0.041 | 0.068+ |
| | | | (0.031) | (0.030) | (0.042) |
| $\rho_{VF}, \rho_{VC}$ | | | | 0.462*** | −0.407*** |
| | | | | (0.049) | (0.051) |
| $\rho_{FC}$ | | | | | 0.136* |
| | | | | | (0.059) |
| Total effect routine patients | | | | | |
| Vol | −0.094* | 0.051 | | | |
| | (0.042) | (0.061) | | | |
| Foc | −0.253*** | −0.196** | | | |
| | (0.033) | (0.062) | | | |
| Con | 0.038 | −0.055 | | | |
| | (0.035) | (0.077) | | | |
| Total effect complex patients | | | | | |
| Vol | 0.024 | 0.162* | | | |
| | (0.043) | (0.063) | | | |
| Foc | −0.018 | 0.032 | | | |
| | (0.031) | (0.058) | | | |
| Con | −0.127*** | −0.212** | | | |
| | (0.035) | (0.079) | | | |
| Effect differences | | | | | |
| Δ Vol (PR, PC) | −0.118* | −0.111* | | | |
| | (0.052) | (0.051) | | | |
| Δ Foc (PR, PC) | −0.234*** | −0.228*** | | | |
| | (0.043) | (0.043) | | | |
| Δ Con (PR, PC) | 0.165*** | 0.158*** | | | |
| | (0.048) | (0.048) | | | |
| Observations | 265,133 | 265,133 | | | |
| Segments-in-hospitals | 2,067 | 2,067 | | | |

Standard errors clustered on segments-in-hospitals; controls included as per section 4.7.
*** p<0.001, ** p<0.01, * p<0.05 + p<0.10.

concern. Overall, there is clear evidence for endogeneity and the strength of our instruments. We therefore use the results of the mortality equation of the simultaneous equations model to evaluate the evidence for our hypotheses. We include the probit estimations in Table 1 to allow readers to evaluate the selection bias.

**Hypothesis 1** posits a positive effect of volume on service quality (i.e. a negative effect of volume on mortality) for routine patients and a deterioration of this effect for complex patients. This hypothesis is only partially supported by the simultaneous equations model in Table 1. We find no significant effect of volume on mortality for routine patients ($0.035+0.016=0.051, p>0.1$). However, we find evidence that the volume–mortality effect differs between routine and complex patients ($0.016-0.127=-0.111, p<0.05$) and that there is a significant *negative* effect of volume for complex patients ($0.035+0.127=0.162, p<0.05$) in our data. Note that the standard probit model supported our hypothesis fully, estimating a beneficial effect of volume on mortality for routine patients ($-0.106+0.011=-0.094, p<0.05$), no significant effect for complex patients ($-0.106+0.129=0.024, p>0.1$) and a significant differential effect between routine and complex patients ($0.011-0.129=-0.118, p<0.05$). However, the simultaneous equations model shows that this evidence is spurious and can be explained by a selection effect. The significant negative correlation coefficient in the volume selection equation ($\rho_{VD}=-0.109, p<0.001$) provides evidence that the patient pool in high volume hospitals has a lower overall mortality risk. This observation is consistent with "cherry-picking" by larger hospitals (KC and Staats 2012). Once this selection effect has been taken into account, there is no longer a difference between high and low volume hospitals for benchmark patients (coeff = 0.035, n.s.). Interestingly, the interaction coefficients are very similar between the probit and simultaneous equations models (0.011 to 0.016 and $0.129, p<0.01$ to $0.127, p<0.01$), i.e., the data provides no evidence that the volume selection effect differs by patient complexity.

Note that the signs of the correlations of the error terms of the selection equations are consistent with the correlations in the raw data (Table 2 in Anonymized 2016). Specifically, the positive correlation $\rho_{VF}$ is expected because higher volume is positively correlated with higher relative volume. The negative correlation between volume and concentration is likely due to the fact that higher segment volume allows for subspecialization in the segment and these subspecialists may be distributed across different departments in the hospital, leading to more fragmented departmental routing patterns for patients in the segment. We do not observe the subspecialty distribution in our data but it is captured in the estimated correlation coefficient.

**Hypothesis 2** posits a positive effect of focus on service quality for routine patients and a deterioration of this effect for complex patients. This hypothesis is fully supported by our estimation results. We find a significant effect of focus on mortality for routine patients ($-0.007-0.189=$

$-0.196, p < 0.01$). We also find that the focus–mortality effect differs significantly between routine and complex patients ($-0.038 - 0.189 = -0.228, p < 0.001$). Focus does not have a significant effect on mortality for the complex patients in our data ($-0.007 + 0.038 = 0.032, p > 0.1$).

**Hypothesis 3** posits a positive effect of segment concentration on service quality for complex patients and a deterioration of this effect for routine patients. This hypothesis is fully supported by the estimation results. We find a significant effect of segment concentration on mortality for complex patients ($-0.144 - 0.068 = -0.212, p < 0.01$). We also find that the concentration–mortality effect differs significantly between routine and complex patients ($0.089 + 0.068 = 0.158, p < 0.001$). Segment concentration does not have a significant effect on mortality for the routine patients in our data ($-0.144 + 0.089 = -0.055, p > 0.1$).

Note that our hypotheses juxtapose the effects of the three independent variables for routine patients against the most complex patients. The mid-range of complexity is covered by the reference category of benchmark patients and one may ask whether the moderating effect of complexity on the three independent variables is strictly monotone over the three complexity categories. While there is not always a significant difference between either routine and benchmark or benchmark and complex patients, the data provides some evidence for monotonicity. Specifically, whenever the interaction terms in the simultaneous equations model in Table 1 are significant, the sign supports monotonicity in the expected direction:

- $Vol * PC$ is significantly positive ($0.127, p < 0.01$), i.e., any beneficial effect of volume is worse for complex patients than for benchmark patients;
- $Foc * PR$ is significantly negative ($-0.189, p < 0.001$), i.e., any beneficial effect of focus is more beneficial for routine patients than for benchmark patients;
- $Con * PR$ is significantly positive ($0.089, p < 0.05$) and $Con * PC$ is significantly negative ($-0.068, p < 0.05$), i.e., segment concentration is significantly more beneficial for benchmark patients than for routine patients and significantly more beneficial for complex patients than for benchmark patients.

Table 2 reports risk-adjusted mortality rates by patient type and independent variable, based on the mortality equation of the simultaneous equations model in Table 1. These mortality rates are obtained by calculating the model-predicted probabilities of death for each patient in the sample in the two counterfactual worlds, where either all sample patients are admitted to low-volume (low-focus, low-concentration) hospitals or all are admitted to high-volume (high-focus, high-concentration) hospitals for their segment. The difference between these two estimates, averaged over the sample, is the average partial effect (APE). We report APEs only for significant differences and as percentage point changes (pp). All statistically significant effects are also clinically highly significant.

**Table 2 Estimated risk-adjusted seven-day in-hospital mortality rates**

| | Volume-mortality | | |
|---|---|---|---|
| | Low | High | APE |
| Routine | 1.39% | 1.55% | n/s |
| Benchmark | 3.35% | 3.59% | n/s |
| Complex | 4.58% | 6.08% | 1.50 pp |
| | Focus-mortality | | |
| | Low | High | APE |
| Routine | 2.09% | 1.36% | -0.73 pp |
| Benchmark | 3.57% | 3.52% | n/s |
| Complex | 5.56% | 5.88% | n/s |
| | Concentration-mortality | | |
| | Low | High | APE |
| Routine | 1.59% | 1.41% | n/s |
| Benchmark | 3.91% | 2.98% | -0.93 pp |
| Complex | 6.49% | 4.50% | -1.99 pp |

Predictions based upon the outcome equation of the simultaneous equation model in Table 1.

## 6. Robustness and limitations

In view of the multiple modelling choices we had to make, we conducted the following robustness checks which we report on in detail in the supplementary material (Anonymized 2016). For each robustness check we estimated two models (a single equation probit model and a simultaneous equations model to control for endogeneity) analogously to Table 1.

1. We use several arbitrary thresholds in the definition of our variables and exclusion criteria for the sample and therefore checked the robustness of our findings to changes in these thresholds (Anonymized (2016), Chapter 8, Tables 16-18).

2. We estimated different models with varying sets of instrumental variables in the three selection equations (Anonymized (2016), Chapter 5, Tables 7-10), including a parsimonious model with only one IV per selection equation (Anonymized (2016), Chapter 5, Table 10) .

3. We varied the measure of patient complexity in three ways, (i) by using two and four instead of three Elixhauser comorbidities as the cut-off for complex patients (Anonymized (2016), Chapter 9, Tables 19 and 20), (ii) by using an alternative medical complexity measure based on information in the German DRG codes that is analogous to complications and comorbidity information in the US Medicare DRGs (Anonymized (2016), Chapter 9, Table 21), and (iii) by using a complexity measure based on whether the comorbidity is likely to be known at the time of admission (Anonymized (2016), Chapter 9, Table 22).

4. To increase homogeneity, we estimated models for two subsamples, (i) for the 94,622 patients with diseases of the circulatory system (Anonymized (2016), Chapter 10, Table 23)) and (ii) for

a subsample of 62,470 patients with six conditions for which quality of care is known to affect mortality significantly (for details see Kuntz et al. 2015) (Anonymized (2016), Chapter 10, Table 24). Although, as argued in Section 4.4, dichotomization of the independent variables is preferable in our context, we re-estimated the models with continuous independent variables as a robustness check. While the results support the results of the dichotomous model, the continuous model restricts the instrumentation that we can use and the estimation results lead to concerns about weak instruments. This is an empirical reason why we prefer the dichotomized model for our data. We refer to Anonymized (2016), Chapter 11, Table 25 for details.

5. We estimated four alternative model specifications,

(i) models with standard errors clustered at the hospital level (Anonymized (2016), Chapter 12, Table 26), (ii) models with random segment-within-hospital effects (Anonymized (2016), Chapter 12, Table 27), (iii) a discrete survival analysis with discharge as competing risk (Kuntz et al. 2015) (Anonymized (2016), Chapter 12, Table 28), (iv) a linear probability model. The linear probability model offered a poor fit to the data as it predicted negative mortality probabilities for over 20% of the sample patients. Horrace and Oaxaca (2006) show that OLS estimates of linear probability models are biased and inconsistent when there is a positive probability that the data generating process produces covariates that lead to predictions outside the unit interval and that the bias grows with the proportion of predictions that fall outside this interval. We therefore consider this model as unreliable in our context and do not present the results in Anonymized (2016).

The coefficient signs of all 17 tables of alternative specifications presented in Anonymized (2016) are the same as in Table 1; 14 models show qualitatively identical or stronger significance patters, while three models lend somewhat weaker support to our hypotheses (Tables 17, 24, 28). Two of the latter three models provide similar coefficient estimates but are based on smaller samples (Tables 17 and 24), indicating a lack of statistical power. In summary, the robustness checks support the result of our main model presented in Table 1.

### 6.1. Limitations

Naturally, this study has limitations beyond the issues raised in the robustness checks. Specifically, there are several sources of estimation bias in our context, including the non-randomness of the hospital sample, the hierarchical nature of the data, patient selection effects and measurement errors. While we have attempted to address these sources in our model specification, we acknowledge that there are residual concerns and limitations.

First, our hospital sample is not a random sample from the German hospital industry and is somewhat biased towards larger hospitals, with an average of 378 beds compared to a national average of 243 beds, and against for-profit hospitals, which account for 15.2% of the sample hospitals

compared to a German average of 27.8% (Statistisches Bundesamt 2008). In addition, the sample contains two large clusters in two federal states, with the remaining hospitals scattered across the country. In terms of teaching hospital status, the sample's proportion of 30% is close to the national proportion of 31%. The non-representativeness of the hospital sample is a limitation of the study and we caution against the generalizability of our findings to smaller hospitals in Germany and beyond the German hospital context. We have addressed the non-random nature of the hospital sample by including hospital fixed effects to control in an aggregate way for structural differences between hospitals that affect mortality. The use of fixed effects, however, may cause incidental parameter bias and random effects models are often proposed as better alternatives. In our context, a fixed effects model is preferable to a random effects model because our hospital sample cannot be assumed to be randomly drawn from the hospital population. In fact, Greene (2004) shows in a simulation study that the bias due to random effects may be larger than the bias caused by the incidental parameter problem. While the results of this study suggest that the fixed effects model is appropriate for our data (60 hospitals and 39 disease segments), incidental parameter bias cannot be entirely ruled out, albeit it is likely to be small (Greene 2004).

Second, to address the hierarchical nature of our data, with patients within segments-of-hospitals, we have estimated a cluster-specific probit model (see Section 4.8.1). This model provides a consistent estimation of average partial effects but is predicated on the assumption that the cluster-specific effects ($\nu_{sh}$ in (1)) are normally distributed and independent of the other covariates in the model (Wooldridge 2010, pp. 610-611). While this is a standard assumption, we acknowledge that the assumption is quite strong, even after controlling for hospital and segment fixed effects.

Third, we believe patient selection and other causality effects, such as reverse causality, to be the most likely source of bias. We addressed this directly in the model specification, with one outcome and three selection equations. While such recursive multi-equation probit models are generally identified without instrumental variables (Wilde 2000), the use of instruments makes the estimation more robust to deviations from the multi-normality assumptions. We have used variants of two established instruments for each selection equation and are confident that these address selection bias appropriately. However, the cross-sectional nature of our data limits the strengths of the evidence of a general causality claim beyond our setting. The effect that we identify could be specific to the context of the period during which our data were collected and its environmental and regulatory regimes. We cannot generalize beyond this context with our data.

Fourth, while some of the robustness checks have addressed issues of measurement error, residual concerns remain. Specifically, our complexity measures, based on aggregating comorbidity information in discharge records, are coarse and generic. They are therefore likely to suffer from measurement errors, which may attenuate our estimates. It would be desirable to work with segment-specific complexity measures that account for the differential impact of different comorbidities on

the service process of specific diseases. The development of such measures, however, is a significant undertaking that would require disease-specific clinical input; this goes beyond the scope of this study.

Finally, our estimated volume, focus and concentration effects are averaged across patient types (routine, benchmark, complex) in the segment. When we apply these estimates to assess the effect of separating out routine patients into value-adding process clinics, we make the assumption that a reduction of the volume of routine patients in a segment has the same effect as a reduction of segment volume with the mix of patient types kept constant. While an extension of our model could be applied to evaluate this assumption by estimating models at the level of patient-type volumes, focus and concentration in the segments, we have refrained from this level of granularity in the interest of keeping the paper focused and legible. Further studies are needed to address this specific question.

## 7. Managerial implications and conclusions

Christensen et al. (2009) argue that the coexistence of two fundamentally misaligned operational models lies at the heart of the unsustainable increase in healthcare costs in general hospitals. On the one hand, general hospitals operate "solution shops", where doctors practice "intuitive medicine" for patients with complex and ill-diagnosed conditions for whom no standard diagnosis and treatment path is available. On the other hand, general hospitals operate value-adding processes, where clinicians apply rules-based medicine to patients who are well-diagnosed and have clear treatment plans. The trial-and-error approach that is required for the first type of patient is at odds with the "getting it right first time" paradigm that one expects from an effective value-adding process. Edmondson (2012) explains the different teaming approaches that are required by routine and complex operations. Christensen et al. (2009) propose a radical change of general hospitals, separating these two types of activities, so that solution shops and value-adding process clinics can concentrate on optimizing their respective operations. Using data from 60 German hospitals, this paper studies the effect of such a reorganization on service quality, specifically mortality. The results support Christensen et al. (2009)'s proposal in three ways.

First, as a hospital's routine patients are separated out for treatment in organizationally separate value-adding process clinics, the hospitals' volume in the remaining solution shop shrinks. It is therefore important to understand how the volume–outcome relationship affects the different types of patients – routine patients in value-adding process clinics and more complex patients in solution shops. Our data suggests a significant *negative volume–outcome* effect for complex patients: Complex patients who are treated in hospitals with a high volume in their patient segment have, on average, a higher mortality risk after controlling for hospital selection. Although this observation

contradicts much of the literature, which typically stipulates a positive volume-outcome effect, we are not the first to show a negative effect (e.g. Horwitz et al. 2015, for readmissions). We argue theoretically that effective care for complex patients relies heavily on informal and relational coordination between practitioners and that such coordination becomes increasingly ineffective with scale. Our data suggests that the reduction in hospital scale that would be caused by separating out routine patients into value-adding process clinics will not harm but may in fact benefit the remaining more complex patients in the solution shops. Note, however, that our results are based on a coarse comparison of mortality rates for two groups of hospitals – high-volume and low-volume hospitals – for each disease segment. It is quite possible that the volume effect is nonlinear and that hospitals that fall below a minimal volume threshold perform worse for both routine and complex patients. If so, the separation for such very low volume segments might not bring quality benefits. Showing the existence and estimating the magnitude of such minimal segment volumes for the separation of routine patients needs further research.

Second, our models include a hospital's segment total volume and relative volume (focus) simultaneously and allow these variables to compete to explain mortality. These two variables are positively correlated by nature. We find that for routine patients, focus is the dominant driver of service quality, while volume becomes non-significant when considered together with focus. This suggests that value-adding process clinics should be set up as *focused factories* (Skinner 1974) that specialize in specific disease segments for routine patients. It also suggests that these clinics do not have to be organized at a regional level to achieve high volume, as sometimes proposed. They may be set up at a local level, in proximity to but organizationally separated from the general hospital that will concentrate on solution shop activities for emergency and complex elective patients. This may allow some resource sharing and, importantly, would facilitate referral between these two types of organizations if and when appropriate.

Third, we investigate a new variable that measures the extent to which a hospital routes newly arriving patients in a segment into the same department. We find that a high degree of segment concentration is associated with reduced mortality for the complex patients in that segment, while the data do not show a significant mortality effect for routine patients. This observation suggests that after separating routine patients into focused, value-adding process clinics, the remaining downsized general hospitals should adapt disease-based departmental routing strategies to ensure that all patients in a disease segment are admitted to one clinical department that takes responsibility for that entire segment.

In order to offer an estimate of the potential magnitude of the quality effect of the proposed reorganization of general hospitals, we conduct a counterfactual analysis for our sample, using the endogeneity corrected outcome equation of the simultaneous equations model in Table 1 to

predict mortality effects. We perform the counterfactual analysis in two steps, a separation and a concentration step at the level of each hospital.

First, we split each hospital into two organizationally separate main divisions, a solution-shop division for non-routine patient care and a value-adding process division for routine patient care. We then further subdivide the value-adding process division into a set of segment-focused value-adding process clinics. Note that each segment is now split between the two main divisions and therefore the patient volume in the two divisions is smaller than in the original general hospital. For the counterfactual prediction we therefore re-calculate the segment volume dummy at the divisional level. For example, if the annualized volume of *routine patients* in segment $s$ in hospital $h$ falls below this median threshold for segment $s$ as calculated in Section 4.4, then we reset the volume indicator $V_{ish}$ to zero for all routine patients in segment $s$ and hospital $h$. We do the same for the focus dummy variable for non-routine patients. For routine patients, however, we set the focus variable equal to 1 because these patients are now treated in segment-focused factories. We perform a first set of counterfactual predictions with these changes of the volume and focus variables alone, without changing the concentration variable. In a second step, we increase the concentration variable for the non-routine patients in the solution shops from the original 31% of patients stepwise to 60%.

Table 3 and Figure 2 summarize the results of this reorganization for several segment concentration levels. Pulling the first lever, i.e. separating routine patients and non-routine patients and delivering care for the former in focused value-adding process clinics, reduces mortality for routine patients by 13.43% (95% CI [6.87%; 18.95%]; Number needed to treat to avoid one death (NNT)=491, 95% CI [331; 946]). The relative effect of this first stage of the re-organization is small and statistically insignificant for the non-routine patients in the solution shop divisions, whose mortality rate decreases by 1.21% (95% CI [−2.49%; 5.43%]) for benchmark patients and by 2.57% (95% CI [−3.44%; 2.49%]) for complex patients after reorganization. However, Figure 2 shows that the second lever, increasing segment concentration, has a large effect on mortality of the non-routine patients. When segment concentration increases from the original 31% to 45% of hospital patients, mortality reduces by 4.44% for benchmark patients (95% CI [0.27%, 8.44%]; NNT=637, 95% CI [329; 10,820]) and by 7.03% for complex patients (95% CI [1.29%, 12.43%]; NNT=247, 95% CI [136; 1,381]). If segment concentration is increased to 60% of hospital patients, mortality reduces 7.78% for benchmark patients (95% CI [3.72%, 11.68%], NNT=363, 95% CI [237; 776]) and by 11.67% for complex patients (95% CI [6.13%, 16.86%]; NNT=149, 95% CI [100; 292]).

We acknowledge that this counterfactual analysis is simplistic in nature. Its main purpose is to illustrate that the size of the quality effect that may be achievable with a reorganization of general

**Table 3 Counterfactual analysis: Redesign of general hospitals**

| | Routine patients | Benchmark patients | Complex patients |
|---|---|---|---|
| Before reorganization | | | |
| % patients in high-volume hospitals | 89.88% | 83.70% | 81.91% |
| % patients in high-focus hospitals | 80.89% | 68.51% | 60.33% |
| % patients in high-concentration hospitals | 31.25% | 31.60% | 29.25% |
| predicted seven-day mortality rate | 1.52% | 3.54% | 5.76% |
| After reorganization | | | |
| % patients in high-volume hospitals | 69.01% | 68.89% | 69.61% |
| % patients in high-focus hospitals | 100% | 82.63% | 79.98% |
| % patients in high-concentration hospitals | 31.25% | 31.56% | 29.76% |
| predicted seven-day mortality rate | 1.32% | 3.49% | 5.62% |
| Absolute rate difference | 0.0020 | 0.0004 | 0.0015 |
| % rate difference | 13.43% | 1.21% | 2.57% |
| NNT | 491 | 2,333 | 676 |

NNT: number needed to treat to avoid one death

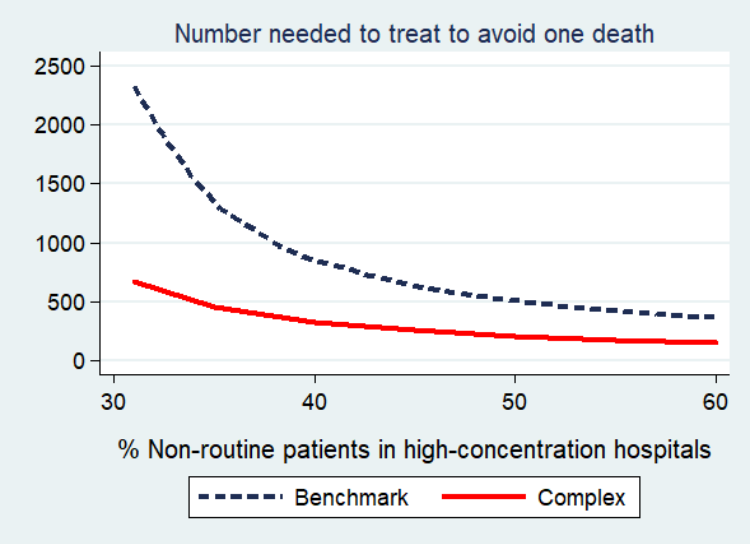


**Figure 2 Number needed to treat depending on segment concentration**

hospitals into organizationally separate solution shops and value-adding process clinics is clinically meaningful. In reality, such a reorganization would pose a number of challenges that are not accounted for in this model. First, economic issues are paramount. General hospitals that provide profitable services for routine patients often use these profit margins to cover losses associated with the care of more complex patients. How would such cross-subsidization be maintained in a split organization? If hospitals receive different reimbursements for the two types of services, to remove the need for cross-subsidization, then economic rather than clinical and operational considerations may determine who they classify as a routine or non-routine patient. Second, a successful implementation of the proposed reorganization will require an effective gatekeeping process. How can

routine patients be identified and who should make the routing decision to value-adding process clinics? These are important questions that need to be addressed by future research. Additional research is also required to test our hypotheses in the context of specific conditions, using condition-specific quality and complexity metrics and allowing for nonlinear threshold effects. The results presented in this paper relate to average effects across patient segments and our data sample is neither large nor granular enough to carry out segment-specific analyses.

## Acknowledgments

The authors gratefully acknowledge the extremely helpful feedback on draft versions of this paper from Michael Freeman, Stelios Kavadias, Christoph Loch, Serguei Netessine, Nicos Savva, and seminar audiences at Insead and IE Madrid as well as senior executives and clinical staff at Cambridge University Hospitals NHS Foundation Trust and the University Hospital Cologne. We thank Christian Rossbach for his help with the data collection, the IMVR for their help with interpreting the data quality reports, and Christine Dentten for her editorial support. Sandra Sülz acknowledges the financial support of the Cologne Graduate School and the German Academic Exchange Service (DAAD).

## References

Alizamir, S, F de Véricourt, P Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Anonymized. 2016. *Separate & Concentrate: Supplementary Material*.

Argote, L. 1982. Input uncertainty and organizational coordination in hospital emergency units. *Administrative Science Quarterly* **27**(3) 420–434.

Argote, L. 1993. Group and organizational learning curves: Individual, system and environmental components. *British Journal of Social Psychology* **32** 31–51.

Argote, L. 2013. *Organizational learning: Creating, retaining and transfering knowledge*. Springer Verlag, New York.

Bardhan, I, VV Krishnan, S Lin. 2013. Team dispersion, information technology, and project performance. *Production and Operations Management* **22**(6) 1478–1493.

Beyer, JM, HM Trice. 1979. A reexamination of the relations between size and various components of organizational complexity. *Administrative Science Quarterly* **24**(1) 48–64.

Bhattacharya, J, D Goldman, D McCaffrey. 2006. Estimating probit models with self-selected treatments. *Statistics in medicine* **25**(3) 389–413.

Birkmeyer, JD, AE Siewers, EVA Finlayson, TA Stukel, FL Lucas, I Bastista, H Gilbert Welch, DA Wennberg. 2002. Hospital volume and surgical mortality in the United States. *The New England Journal of Medicine* **246** 1128–1137.

Boh, WF, SA Slaughter, JA Espinosa. 2007. Learning form experience in software development: A multilevel analysis. *Management Science* **53**(8) 1315–1331.

Bundesamt für Bauwesen und Raumordnung. 2012. Indikatoren und Karten zur Raum- und Stadtentwicklung in Deutschland und Europa. ISBN 978-3-87994-933-5.

Christensen, CM, JH Grossmann, H Jason. 2009. *The Innovator's Prescription: A Disruptive Solution for Health Care*. McGraw Hill, New York.

Clark, JR. 2012. Comorbidity and the limitations of volume and focus as organizing principles. *Medical Care Research and Review* **69**(1) 83–102.

Clark, JR, R Huckman. 2012. Broadening focus: Spillovers, complementarities and specialization in the hospital industry. *Management Science* **58**(4) 708–722.

Cram, P, GE Rosenthal, MS Vaughan-Sarrazin. 2005. Cardiac revascularization in specialty and general hospitals. *The New England Journal of Medicine* **352**(14) 1454–1462.

Dunbar, RIM. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* **22**(6) 469–493.

Edmondson, AC. 2012. *Teaming: How organizations learn, innovate and compete in the knowledge economy*. Wiley.

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

Faraj, S, X Xiao. 2006. Coordination in fast-response organizations. *Management Science* **52**(8) 1155–1169.

Freeman, M, N Savva, S Scholtes. 2016. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* URL `http://dx.doi.org/10.1287/mnsc.2016.2512`.

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

Gemeinsamer Bundesausschuss. 2016. Qualitätsberichte der Krankenhäuser. URL `https://www.g-ba.de/institution/themenschwerpunkte/qualitaetssicherung/qualitaetsbericht/?`

Gersenovic, M. 1995. The ICD family of classifications. *Methods of information in medicine* **34**(1-2) 172–175.

Gittell, JH. 2002. Coordinating mechanisms in care provider groups: Relational coordination as a mediator and input uncertainty as a moderator of performance effects. *Management Science* **48**(11) 1408–1426.

Gonçalves, B, N Perra, A Vespignani. 2011. Modeling users' activity on twitter networks: Validation of Dunbar's number. *PloS one* **6**(8) e22656.

Grant, RM. 1996. Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science* **7** 375–387.

Gray, JV, E Siemsen, G Vasudeva. 2015. Colocation still matters: Conformance quality and the interdependence of R&D and manufacturing in the pharmaceutical industry. *Management Science* **61**(11) 2760–2781.

Greene, W. 2004. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal* **7**(1) 98–119.

Hopp, WJ, WS Lovejoy. 2013. *Hospital Operations*. FT Press, Upper Saddle River, New Jersey.

Horrace, WC, RL Oaxaca. 2006. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters* **90** 321–327.

Horwitz, LI, Z Lin, J Herring, S Bernheim, EE Drye, HM Krumholz, JS Ross. 2015. Association of Hospital Volume with Readmission Rates: A Retrospective Cross-sectional Study. *British Medical Journal* 1–9.

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

Huckman, RS, DE Zinner. 2008. Does focus improve operational performance? Lessons from the management of clinical trials. *Strategic Management Journal* **29** 173–193.

KC, DS, BR Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* **14**(4) 618–633.

KC, DS, BR Staats, F Gino. 2013. Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. *Management Science* **59**(11) 2435–2449.

KC, DS, C Terwiesch. 2011. The effects of focus on performance: Evidence from California hospitals. *Management Science* **57**(11) 1897–1912.

Knapp, LG, TG Seaks. 1998. A Hausmann test for a dummy variable in probit. *Applied Econometric Letters* **5** 321–323.

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

Larsson, R, DE Bowen. 1989. Organization and customer: Managing design and coordination of services. *Academy of Management Review* **14**(2) 213–233.

Lee, H-H, EJ Pinker, RA Shumsky. 2012. Outsourcing a two-level service process. *Management Science* **58**(8) 1569–1584.

Lord, CG, L Ross, MR Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* **37**(11) 2098–2109.

MacCallum, RC, S Zhang, KJ Preacher, DD Rucker. 2002. On the practice of dichotomization of quantitative variables. *Psychological methods* **7**(1) 19.

McClellan, M, BJ McNeil, JP Newhouse. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**(11) 859–866.

McDermott, CM, GN Stock. 2011. Focus as emphasis: Conceptual and performance implications for hospitals. *Journal of Operations Management* **29** 616–626.

Mesman, R, GP Westert, BJMM Berden, MJ Faber. 2015. Why do high-volume hospitals achieve better outcomes? A systematic review about intermediate factors in volume-outcome relationships. *Health Policy* **119**(8) 1055–1067.

Mihm, J, CH Loch, D Wilkinson, BA Huberman. 2010. Hierarchical structure and search in complex organizations. *Management Science* **56**(5) 831–848.

Ong, MS, E Coiera. 2011. A systematic review of failures in handoff communication during intra-hospital transfers. *The Joint Commission Journal on Quality and Patient Safety* **37**(6) 274–284.

Pisano, GP, RMJ Bohmer, AC Edmondson. 2001. Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. *Management Science* **47**(6) 752–768.

Rabin, M, JL Schrag. 1999. First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics* **114**(11) 37–82.

Rathnam, S, V Mahajan, AB Whinston. 1995. Facilitating coordination in customer support teams: A framework and its implications for the design of information technology. *Management Science* **41**(12) 1900–1921.

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

Roodman, D. 2011. Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal* **11**(2) 159–206.

Ross, JS, SLT Normand, Y Wang, DT Ko, J Chen, EE Drye, PS Keenan, JH Lichtman, H Bueno, GC Schreiner HM Krumholz. 2010. Hospital volume and 30-day mortality for three common medical conditions. *The New England Journal of Medicine* **362**(12) 1110–1118.

Schilling, MA, P Vidal, RE Ployhart, A Marangoni. 2003. Learning by doing something else: Variation, relatedness, and the learning curve. *Management Science* **49**(1) 39–56.

Shumsky, RA, EJ Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Skinner, W. 1974. The Focused Factory. *Harvard Business Review* **52**(3) 113–121.

Staats, BR, F Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* **58**(6) 1141–1159.

Statistisches Bundesamt. 2008. *Grunddaten der Krankenhäuser*. Fachserie 12 Reihe 6.1.1.

Theokary, C, ZJ Ren. 2011. An empirical study of the relations between hospital volume, teaching status, and service quality. *Production and Operations Management* **20**(3) 303–318.

Thirumalai, S, KK Sinha. 2011. Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences. *Management Science* **57**(2) 376–392.

Tolbert, PS, RH Hall. 2009. *Organizations: Structures, Processes, and Outcomes*. Pearson, New Jersey.

Valentine, MA, AC Edmondson. 2015. Team scaffolds: How mesolevel structures enable role-based coordination in temporary groups. *Organization Science* **26**(2) 405–422.

Van de Ven, AH, AL Delbecq, R Koenig, Jr. 1976. Determinants of coordination modes within organizations. *American Sociological Review* **41** 322–338.

Wilde, J. 2000. Identification of multiple equation probit models with endogenous dummy regressors. *Economic Letters* **69**(3) 309–312.

Wooldridge, JW. 2010. *Econometric analysis of cross section and panel data*. 2nd ed. The MIT Press, Cambridge, Massachusetts.