Economies of Scale and Scope in Hospitals

Michael Freeman
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom mef35@cam.ac.uk
Nicos Savva
London Business School, Regent's Park, London NW1 4SA, United Kingdom nsavva@london.edu
Stefan Scholtes
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom s.scholtes@jbs.cam.ac.uk

General hospitals across the world are becoming larger (i.e. admitting larger volumes of patients each year) and more complex (i.e. offering more complex portfolios of services to patients with diverse levels of acuity). Although prior work has shown that increased volume is positively associated with patient outcomes, it is less clear how volume interacts with organizational complexity to affect costs across service lines and acuity levels. This paper investigates this relationship using panel data for 14 service lines comprising both elective and emergency admissions across 130 hospitals in England over a period of nine years. Although we find significant economies of scale for both elective and emergency admissions, we also find evidence of *negative* economies of scope across the two admission types, with increased elective volume at a hospital being associated with an *increase* in the cost of emergency care. Furthermore, for emergency admissions we find evidence of economies of scope across service lines – increased emergency activity in one service line is associated with lower costs of emergency care in other service lines. By contrast, we find no evidence of such economies of scope across service lines for elective admissions. Our findings have implications for individual hospitals and for the organization of regional hospital systems. Specifically, at the hospital level our findings suggest that growth strategies that target elective patients may have unintended negative productivity implications for emergency services. At the regional level, our findings offer support for the reorganization of regional hospital systems toward general hospitals that focus on the provision of emergency care across a full range of services, complemented by high-volume clinics that focus on elective services in a single service line.

Key words: healthcare; productivity; economies of scale; economies of scope; complexity; econometrics
History: June 27, 2016 **Preliminary version – Please do not cite or circulate**

1. Introduction

Scale is an important determinant of productivity and a recurrent theme in the operations management and economics literature. Although scale is generally associated with higher productivity (Panzar and Willig 1977), scholars have pointed out that the productivity gains of increased output have to be traded off against the potential productivity losses caused by the increased heterogeneity of that output (Penrose 1959, Schoar 2002). The tension between benefits of scale and potential

disbenefits of scope is of particular concern in the hospital industry (Argote 1982, Clark and Huckman 2012). General hospitals provide a large and diverse range of services and use a wide array of technologies and expertise. From both a strategic and operational perspective, this diversity is surprising. At the strategic level, it is at odds with the focus principle (Skinner 1974), and at the process level, it impedes improvement techniques that are based on the reduction-of-variation principle (Hopp and Spearman 2004). Recent studies have discussed the negative impact of the extensive scope of hospital services on service quality measures (e.g. mortality in the hospital context (Kuntz et al. 2015)); however, perhaps due to the lack of data, there has been little research on the productivity (i.e. cost) implications. This paper uses a novel dataset to provide empirical evidence of the trade-off between scale and scope in the context of the hospital industry.

We focus on two particularly important sources of heterogeneity in hospital services: (i) *admission-type* heterogeneity (i.e. elective or emergency care) and (ii) *service-line* heterogeneity. Admission-type heterogeneity is the result of collocating the treatment of elective and emergency patients within the same hospital. Elective care is often surgical, ranging from simple day cases (e.g. hernia repairs) and short stays (e.g. joint replacements) to complex, long-stay operations (e.g. open-heart surgery). Elective care allows doctors to plan ahead for the services they will deliver in the hospital. Emergency care has a very different dynamic as emergency patients exhibit symptoms that need to be diagnosed and treated under significant time pressure (RCS/DH 2010, AHRQ 2014); there is no a priori treatment plan and the eventual treatment sequence emerges as a consequence of decisions made "on the spot" as the hospital service progresses. Service-line heterogeneity is the result of a number of clinical specializations being collocated within the same hospital and a consequence of the historical evolution of hospital care. The development of biological science and technological innovation over the past century has necessitated medical and surgical specialization and led to the creation of distinct service lines, typically structured around specific body parts (e.g. eye, heart), systems (e.g. nervous system, respiratory system), or diseases (e.g. cancer, metabolic diseases). These service lines share some resources required for patient care (e.g. diagnostic equipment), while other resources are service line-specific (e.g. specialist physicians).

From a cost perspective, arguments can be put forward both in favor of and against the co-location of multiple service lines that treat both elective and emergency patients. On the one hand, pooling spreads fixed costs across a larger customer base (Moore 1959) and can make investing in more productive assets or process structures economical (Argote 2013). Pooling these services may also confer statistical economies of scale by reducing relative arrival variability (Dijk and Sluis 2004), allowing firms to hold less capacity. Finally, pooling provides more opportunities for

organizations to learn and accumulate experience (Pisano et al. 2001). On the other hand, it is known that pooling benefits diminish with the degree of dissimilarity between the pooled activities (Joustra et al. 2010, Schilling et al. 2003, Staats and Gino 2012), and these benefits may also be offset by the increase in organizational complexity that comes with mixing customers with different service needs (Argote 1982, Kuntz et al. 2015). The prioritization of patients by acuity level can also complicate the pooling–productivity relationship. Elective patients, who are typically less severely ill, may be regarded as a variability buffer for emergency patients as resources that are booked for an elective care episode (e.g. operating theatre time) may be redeployed if an emergency patient needs urgent intervention (RCS/DH 2007). Although this may reduce the costs associated with unpredictable variability in emergency activity (as less surplus capacity is required to guarantee that emergency patients are served promptly), it may well lead to a cost increase in elective care as specialist resources (e.g. surgeons) booked for the cancelled elective activity become idle as a consequence. In summary, whether or not the advantages of pooling counteract the disadvantages of increased organizational complexity and buffering in the hospital context is an open question and one that this paper seeks to answer empirically.

Our empirical study is based on annual average cost data related to nearly 105 million hospital admissions for over 2,000 conditions treated in 130 acute care hospitals in England over a period of nine years. Since the data is longitudinal and comprised of multiple service lines across multiple hospitals, we estimate the volume effects of interest with within- and between-random-effects multilevel modeling that allows us to model the variation in volume over time and between different service lines/hospitals explicitly (Mundlak 1978, Gelman and Hill 2007). We find that the more elective patients a hospital treats within a service line, the lower the cost of these patients (with each doubling in volume reducing costs by 5.7%). Similarly, if the number of emergency patients in a service line increases, the cost of these patients decreases (with a doubling in volume resulting in a 12.1% reduction in costs). We then focus on volume spillover effects between admission types as well as spillovers between service lines within admission types. For electives, scale economies appear to be isolated to the specific service line and admission type: An increase in the volume of emergency patients within a focal service line, or of patients of any admission type across other service lines, has no significant effect on the cost of elective patients in the focal service line. By contrast, for emergency patients, we find that an increase in volume coming from emergency patients in different service lines has a positive impact on productivity (with each doubling in emergency volume reducing costs by 9.9%), while the volume of elective patients (coming from the focal or other service lines) has a detrimental effect on productivity (with a 2.0% increase in costs if elective

patients within the service line are doubled and a 13.5% cost increase if elective patients from all other service lines are doubled).

These findings have important practical implications at both the hospital and regional level. At the hospital level, they suggest that elective care growth strategies – which are often pursued by hospitals to improve overall productivity because elective care has greater standardization potential and, therefore, productivity gains are deemed easier to achieve – may actually lead to a drop in productivity overall because of the unintended negative spillover effect on emergency service productivity. To demonstrate this, we perform a counterfactual analysis based on a large hospital in the City of London and show that a 20% increase in hospital admissions across both admission types leads to a cost saving of 1.2%; however, increasing elective admissions alone by the same number of patients leads to a 2.5% reduction in elective costs but increases emergency costs by 6.8%, leading to a total cost *increase* of 3.1%. Surprisingly, a targeted emergency growth strategy, much less favored by hospital managers due to the complexity of emergency care, is estimated to lead to a cost saving of 6.4% in emergency services without having a significant negative effect on elective care productivity, resulting in a total cost saving of 4%.

At the regional level, our results suggest that redesigning how hospitals are organized could lead to an aggregate reduction in the cost of providing care. A further counterfactual analysis shows that if pairs of hospitals in the London area worked together and redistributed elective service lines so that only one of two hospitals provided any particular service, then the cost of elective treatments could be 4.2% lower without a substantial change in the hospitals' total admissions volumes. Furthermore, our work also presents an additional argument for separating elective patients out of general hospitals. Such patients are better treated in specialized, elective-only treatment centers organized along a single service line. Physicians and health management researchers have repeatedly called for such reorganization (ASGBI 2007, RCS/DH 2007, Christensen et al. 2009, Bohmer 2009, Hopp and Lovejoy 2012, Monitor 2015), and there is evidence to suggest that this would offer quality benefits across the system (RCS/DH 2007, Kuntz et al. 2015). Our findings complement these studies by providing evidence that such a reorganization would also result in productivity gains. Extending our counterfactual analysis, we estimate, for example, that if London were to operate stand-alone elective treatment centers focused on single service lines only, then elective costs could potentially be reduced by 15.3%.

2. Prior Work on Scale Effects in Hospitals

The empirical literature examining economies of scale in hospitals is quite extensive. Although the majority of studies find evidence of the existence of economies of scale, their degree and moderating

circumstances remain subjects of debate (Aletras 1997, Posnett 2002). From an empirical perspective, identifying the magnitude of scale economies is challenging as estimations may be confounded by unmeasured inter-hospital variation in quality, patient mix and severity, cost accounting and reporting procedures, or the degree of utilization of existing capacity (Dranove 1998, Posnett 2002, Kristensen et al. 2008). The study of scale economies also poses theoretical challenges because economies of scale may arise through several causal mechanisms (Dranove 1998), including the spreading of fixed costs (Moore 1959), learning and innovation (Pisano et al. 2001), and new and better utilization of capacity (Hopp and Lovejoy 2012, Argote 2013). This causal complexity suggests that the degree to which scale affects productivity depends on the organizational level at which an analysis takes place. Most studies investigate scale economies at either the level of the hospital as a whole (e.g. Marini and Miraldo 2009) or the level of a particular patient condition (e.g. Gaughan et al. 2012). However, the insights into scale effects that we can expect by studying either level in isolation have their limitations. Scale at the hospital level is often a consequence of the pooling of heterogenous services, and estimated effects may underestimate the economies achievable by pooling highly complementary activities (Dijk and Sluis 2004, Joustra et al. 2010, Vanberkel et al. 2012); studies at the condition level fail to account for spillover effects among related patient lines (Schilling et al. 2003). The level of analysis also matters greatly for practical reasons. If, on the one hand, economies of scale are present primarily at the condition level, with little spillover to other conditions, then this would support calls for greater specialization, with patients being referred to specialist hospitals that act as focused factories (Skinner 1974) that perform with greater efficiency and foster innovation better (Greenwald et al. 2006, Porter and Teisberg 2006). If, on the other hand, economies of scale are achieved by providing care at high volumes regardless of the patient mix, then this would support the call for small general hospitals to be closed and activity to be pooled in large, comprehensive regional general hospitals (West 1998). Our work contributes to this literature by examining economies of scale across multiple levels of analysis, namely service lines and admission type (emergency vs. elective). The evidence we provide supports a simultaneous approach of consolidating emergency services in larger general hospitals and separating out elective services into high-volume focused factories.

A stream of literature complementary to studies of scale economies investigates how volume and focus affect the quality of patient care in hospitals. In their studies on performance in cardiothoracic surgery, Pisano et al. (2001) show that as surgeons perform more procedures they accumulate experience and become faster, while Huckman and Pisano (2006) find that this is also associated with a reduction in mortality, although this effect is firm-specific, and KC and Staats (2012) identify that

the reduction in mortality associated with learning is greater if surgeons perform a larger volume of focal tasks rather than similar but related tasks (see also Ramdas et al. (2014)). The degree to which the volume–outcome relationship is moderated by task similarity has also been studied in the focus literature. Clark and Huckman (2012) find that cardiovascular patients experience better clinical outcomes when a hospital specializes in cardiovascular care but also that there are positive spillovers for these patients if the hospital provides complementary ancillary services as well. This finding is complemented by a number of studies outside the healthcare context, with Schilling et al. (2003) showing that there are learning benefits associated with performing both repeated and related tasks but not with unrelated activities (see also Boh et al. 2007, Narayanan et al. 2009, Staats and Gino 2012). Our work differs from these studies in its focus on productivity (i.e. the cost of providing care) rather than quality (e.g. patient mortality) as well as in its investigation of the productivity spillover effects of volume between different service lines and admission types. The question as to the existence of productivity spillovers associated with volume is not answered in the extant literature and is highly relevant for the current debate on business model innovation in regional hospital systems (ASGBI 2007, RCS/DH 2007, Christensen et al. 2009, Bohmer 2009, Hopp and Lovejoy 2012, Monitor 2015, Kuntz et al. 2015).

Finally, we note that our work is related to a large and growing stream of empirical operations management literature that examines the impact of organizational workload on operational performance and patient outcomes in hospital care. Recent examples include KC and Terwiesch (2009), KC and Terwiesch (2012), Kim et al. (2014), Green et al. (2013), Powell et al. (2012), Chan et al. (2016), Kuntz et al. (2014), and Batt and Terwiesch (2016), among others. In contrast to this literature, which exploits short-term temporal variation in workload, our work focuses on the more long-term impact of volume on hospital costs. Our estimation models exploit variation across hospitals and service lines after controlling for changes in utilization over time.

3. Hypothesis Development

In this section we discuss the general mechanisms behind volume–productivity effects that are relevant for the hospital context. Although most of the literature suggests that treating more patients of the same admission type within a service line (e.g. elective patients with a cardiac condition) should allow hospitals to deliver care at a lower cost for these patients, the productivity spillover effects for other admission types and service lines are less clear and the extant literature offers competing arguments both for (e.g. spreading fixed costs) and against (e.g. increased organizational complexity) pooling. Table 1 provides an overview of these mechanisms and the hypothesized aggregate effects, which we discuss in more detail in this section. The aim of this study is to measure the aggregate effect of volume on costs rather the individual effect of each mechanism.

Table 1 Hypothesized effects of volume on productivity

Effect on...	of an increase in...	from the...	Hypothesized effect due to... (P-F)	(P-S)	(P-L)	(B)	(C)	Aggregate hypothesis
Elective productivity	Elective vol.	Focal SL	↑↑	–	↑↑	–	–	↑↑
		Other SLs	↑	–	–	–	↓	?
	Emergency vol.	Focal SL	↑	–	↑	↓↓	↓	?
		Other SLs	↑	–	–	↓	↓↓	?
Emergency productivity	Elective vol.	Focal SL	↑	–	↑	↑↑	↓↓	?
		Other SLs	↑	–	–	↑	↓	?
	Emergency vol.	Focal SL	↑↑	↑↑	↑↑	–	–	↑↑
		Other SLs	↑	↑	↑	–	↓	?

'SL' is the abbreviation of 'service line'; (P-F) denotes spreading fixed costs, different assets and processes; (P-S) denotes statistical economies of scale; (P-L) denotes learning and experience; (B) denotes buffering; (C) denotes organizational complexity; ↑ denotes a positive effect; ↑↑ denotes a strongly positive effect; ↓ denotes a negative effect; ↓↓ denotes a strongly negative effect; – denotes no effect; ? denotes an ambiguous overall effect.

3.1. Economies of Pooling

Treating more patients may lead to productivity gains in three important ways: fixed-cost amortization, statistical economies of scale, and learning effects. We discuss each in turn and then explain how we expect them to affect productivity across admission types and service lines.

3.1.1. Fixed-cost Amortization Hospitals are largely fixed-cost operations; they maintain a collection of assets to satisfy current and projected future demand and invest in new and improved organizational capabilities and physical infrastructure in order to improve service quality and reduce costs (Wedig et al. 1989). Hospitals that treat more patients will be able to spread their fixed costs across a wider activity base, thereby reducing the average cost per patient. In fact, not only are assets better utilized in higher-volume organizations but the better returns on investment make it more likely that productivity-improving assets will be economical in the first place; such assets are therefore more likely to be found in larger hospitals. For example, studies consistently find that larger hospitals are more likely to adopt innovative health information technology than smaller hospitals (Wilson and Carey 2004). Therefore, larger hospitals have more flexibility in choosing their asset configuration and in organizing their resources, e.g. through the division of labor and specialization (Staats and Gino 2012, Argote 2013). Examples of asset and process improvements that are affordable at scale include more effective medical equipment, technology and facilities, more experienced or specialized doctors and surgeons, and clearer and better delineated care pathways (Best et al. 2015). These improved asset and process structures allow the corresponding activities to be performed more effectively and efficiently, which is expected to result in lower costs and shorter hospital stays (Porter 1979).

3.1.2. Statistical Economies of Scale In addition to being able to spread fixed costs more widely, larger hospitals gain from statistical economies of pooling. It is well known that the pooling of separate queues made up of homogenous customers in a single queue staffed by the same servers can reduce average waiting times (Hopp and Lovejoy 2012, p.513). Furthermore, as higher operating volumes reduce the coefficient of variation of patient arrivals, service systems can achieve the same service level with less surplus capacity. This statistical pooling effect is especially relevant in the hospital context, where outcomes can be highly contingent on patients being seen in a timely manner (see e.g. AHRQ 2014, Chan et al. 2016). Therefore, safety concerns often necessitate high levels of staffing and, consequently, high labor costs – which are estimated to constitute more than half of hospital expenses (Guerin-Calvert 2011, Hurst and Williams 2012).

3.1.3. Learning Effects The third mechanism by which volume affects productivity is learning. At higher volumes there are more opportunities for individuals and organizations to learn, and there is evidence that with additional accumulated experience individuals and organizations become more productive and effective at completing tasks (Pisano et al. 2001, Nembhard and Tucker 2011, Argote 2013). Quality improvements have also been attributed to organizational learning at high volumes (Li and Rajagopalan 1998, KC and Staats 2012, Ramdas et al. 2014). The medical literature complements the management literature and provides strong evidence of a positive association between volume and clinical outcomes across a variety of clinical conditions and surgical procedures (Begg et al. 1998, Birkmeyer et al. 2002). The idioms "practice makes perfect" and "learning by doing" capture the drivers of these effects: Providers that see a high volume of similar patients gain experience and become more effective in applying a given standard of care and, at the same time, are more innovative and develop new routines for improving service delivery (Porter and Teisberg 2006, Christensen et al. 2009). The improvements in service quality and effectiveness expected as a consequence of learning and experience from higher volumes should thus impact positively on productivity and reduce costs.

3.1.4. Complementarity as a Moderator of the Volume–Productivity Relationship Although there are clear benefits associated with pooling, the extent to which there are spillovers from treating more patients of different admission types or in different service lines on the productivity of treating patients of a specific admission type in a specific service line depends on the degree of complementarity, i.e. the extent to which capacity can be reassigned and learning benefits transferred across heterogenous patient groups. An investment that is beneficial for some patients (e.g. equipment to speed up emergency diagnosis or a specialist consultant hired to perform complex orthopedic surgery) may not be obviously beneficial for other, dissimilar patients, or at least

not to the same degree. Thus, we hypothesize that the amortization effect – the advantage of being able to spread fixed costs and afford improved assets and/or process structures – is strongest if volume increases within the same service line and for patients of the same admission type, with reduced (but still positive) effects for increases in the volume of the other admission type or of other service lines. This is summarized in Column P–F of Table 1.

Turning to statistical economies of scale, these will be particularly beneficial for emergency patients, whose arrivals are random and service times are highly variable, rather than for elective patients, whose services are scheduled in advance and the resources for which can, to some extent, be scheduled to match demand. Statistical economies of pooling are also known to be contingent on the degree of homogeneity between the customers who are pooled. The more heterogenous the service requirements, the more that there is to gain from serving customers in dedicated queues (Rothkopf and Rech 1987, Dijk and Sluis 2004, Joustra et al. 2010, Vanberkel et al. 2012). Song et al. (2015), for example, find in the context of the emergency department (ED) that dedicated, single-doctor queues can in fact improve performance compared to non-pooled systems. Since emergency patients with different conditions may differ in their service needs, we hypothesize that the effect of statistical economies of scale on productivity is stronger within a service line than across service lines. This is summarized in Column P–S of Table 1.

There is also evidence that the volume benefits of learning and experience are not isolated to a narrow scope of activities. Schilling et al. (2003) show that there are positive learning spillovers when teams perform tasks that are different but related to a focal task (see also Boh et al. 2007, Narayanan et al. 2009, Huckman and Staats 2011, Staats and Gino 2012, Clark et al. 2013). Although there are clearly overlaps in the service requirements of elective and emergency patients (e.g. in observations, tests, and treatment), there are also many differences in their needs (e.g. fast diagnosis is critical for emergency patients, while elective patients arrive with their care plan already determined); as such, not all accumulated knowledge is transferable across these patient types. Similarly, while initial uncertainty in the diagnosis and preferred treatment plan of emergency patients means that they may benefit from being treated in a hospital that handles a higher volume and wider variety of emergency activities (e.g. diversity of experience may speed up accurate diagnosis and lead to better patient routing (Kuntz et al. 2015)), it is not obvious whether the same is true for elective patients, who are routed to the correct provider on arrival and are less likely to interact with other parts of the service. Therefore, we hypothesize that productivity improvements from learning and experience are greater when treating a high volume of patients from the same service line and of the same admission type. We also hypothesize that there are some learning

spillovers that come from treating a higher volume of elective and emergency patients together within the same service line and that emergency productivity improves when a hospital acquires experience from treating a high volume of emergency patients in general (regardless of the service line). These learning-related hypotheses are summarized in Column P–L of Table 1.

3.2. Buffer Effects of Prioritization

In addition to the benefits listed above, pooling heterogenous customer types may have other benefits, such as increasing the availability of workload-management strategies that can be used to limit the impact of congestion-related deterioration in system performance. Freeman et al. (2016), for example, find that workers employ different levers to manage their workload depending on the complexity of customer needs and that a change in the mix of patients in a unit (e.g. the number of elective patients relative to emergency patients) can affect the flexibility of the system in responding to an increase in workload. KC and Terwiesch (2012) find also that demand pressures caused by the arrival of critical patients can cause patients who are relatively less unwell to be discharged earlier from the intensive care unit (ICU).

In our context, these benefits are most likely to be realized by emergency patients treated alongside a high volume of elective patients, where resources intended for elective care can often be redeployed to more time-sensitive emergency cases at short notice (e.g. by canceling elective procedures). Since delays in treating emergency patients are known to cause complications and higher mortality rates (see e.g. Jestin et al. 2005, RCS/DH 2010, NCEPOD 2010), it is possible to consider the elective patient pool as a buffer that can be exploited to speed up assessment and access to treatment for emergency patients during periods of high demand. Therefore, through this buffering effect, an increased volume of elective patients while emergency volume remains constant should improve the productivity of emergency care. Since this buffering effect is likely to be stronger the more flexibly resources can be redirected, we hypothesize that emergency productivity in a focal service line increases most in the volume of elective activity in the same service line but also benefits from higher elective patients volumes in other service lines.

Although prioritization protocols may create buffers that are beneficial for the higher-priority admissions stream – typically emergency patients – the opposite is likely to be the case for lower-priority electives. In fact, an elective patient treated alongside a high volume of emergency cases may be at increased risk of having her service disrupted and/or delayed (RCS/DH 2007). In a single-hospital study in the UK, Sanjay et al. (2007) determined that approximately 10% of elective surgery cancelations were caused by emergency cases filling elective theater slots, while in a study of German district hospitals, Schuster et al. (2011) found that elective surgery cancelations were higher

in larger hospitals, with nearly twice the rate of cancelations due to emergency prioritization in large hospitals than in small and medium-sized hospitals. Cancelations are costly, wasting expensive resources (e.g. surgical beds and doctors' time) and potentially harmful for patients (Gillen et al. 2009, Argo et al. 2009). This leads us to hypothesize that elective productivity is worst affected when there is a higher volume of emergencies in the same service line, with a weaker effect of emergency volume increases in other service lines. These hypotheses are summarized in Column B in Table 1.

3.3. Organizational Complexity

Although there may be benefits associated with pooling heterogeneous customers, doing so can also complicate service delivery when operationally effective process designs and service delivery modes for the different customer types are misaligned. Christensen et al. (2009) argue that the organizational complexity of modern general hospitals stems largely from their attempt to serve two fundamentally different types of patients within a single organization: those that require the delivery of well-specified value-adding activities and those that arrive at the hospital with poorly diagnosed symptoms and require a search for the best solution in a "solution shop" environment. Elective and emergency patient activity map naturally onto the two fundamentally different activities. For elective patients, the emphasis of the service is on solution *execution*, i.e. the carrying out of planned, often-routine procedures that, following Argote (1982), are best executed in a service setting oriented toward programmed mechanisms of coordination, such as formalized rules and standardized plans and schedules (March and Simon 1958, Thompson 1967). For emergency patients, on the other hand, as their needs are often not known on arrival, the service is often a trial-and-error process, with an iterative process of solution search, treatment execution, and analysis of treatment outcomes. Argote (1982) shows that these patients, for whom there is greater process uncertainty at the service outset, are better served in a service environment oriented toward non-programmed coordination mechanisms, such as interdisciplinary team meetings, which allow greater autonomy and flexibility for individuals and teams to work toward searching for appropriate actions. This basic tension between programmed coordination, where the organization specifies activities in advance and manages compliance, and non-programmed coordination, where the organization leverages its members' autonomy to work out appropriate activities "on the spot," makes the coexistence of the two basic coordination modes in the same unit or firm challenging and potentially ineffective (see also Kuntz et al. 2015). A higher volume of patients who differ in their service requirements may reinforce this tension as it is likely that the balance of misalignment in delivery modes changes in favor of the patient type with the increased volume, rendering the treatment of

the other patient type less efficient. We therefore hypothesize that any increase in the volume of patients of a different admission type and/or from a different service line should negatively affect service productivity for the remaining patients as a result of this increase in organizational complexity. This is summarized in Table 1 by downward arrows in all rows of Column C except for those of the same admission type and same service line.

3.4. Aggregate Effect

With the exception of the impact of volume on productivity within a service line and admission type, the aggregation of the effects described above and summarized in the final column of Table 1 is ambiguous. The resolution of these tensions has important implications, and hence, the main objective of this work is to estimate the aggregate effect stemming from these competing mechanisms empirically. A preview of our empirical findings can be found by referring to Table 5 in §5.

4. Description of the Data, Variables Definitions and Econometric Models

Our primary data set consists of annual costing and inpatient activity data for the nine financial years from 2006/07 to 2014/15 for all acute hospital trusts operated by the National Health Service (NHS) in England over that time period. Acute NHS hospital trusts provide secondary and tertiary healthcare in facilities that range from small district hospitals to large teaching hospitals. Services include EDs, inpatient and outpatient medicine and surgery, and specialist medical services. We focus our attention on admitted patient care and exclude outpatient activity and ED visits that do not result in hospital admission. In total, our data comprises aggregate annual information for nearly 105 million patient admissions to 130 acute hospital trusts. Some trusts operate more than one hospital and a number of trusts were merged and hospitals closed during the observation period. For consistency, we only retain those trust-years in the sample that correspond to the longest period for which the number of hospital sites operated by a trust remained unchanged.

For regulatory purposes, each NHS hospital trust is mandated to complete an annual return of so-called reference costs, reporting the trust's activity for each patient condition treated over the preceding year. Patient conditions are defined using so-called healthcare resource groups (HRGs), which are the UK equivalent of the diagnosis-related groups (DRGs) used by Medicare in the US. HRGs are designed so that patients within an HRG are clinically similar and require a relatively homogeneous bundle of resources for their treatment (Fetter 1991). Each patient episode is assigned to a unique HRG using a semi-automated process based on information provided in the discharge notes, including standardized ICD-10 medical diagnosis codes, OPCS procedure codes,

and contextual information such as patient age and gender and the existence of any complications or comorbidities (see e.g. DH 2013). The costs incurred by the hospital each year are allocated to specific HRGs, with each hospital reporting the average cost of treating patients within each HRG, the average length of stay (LOS) of these patients, and the volume of patients treated from each HRG.

These cost submissions are used by the UK Department of Health to determine the price (also known as the "tariff") to be paid to hospitals for each discharged patient in an HRG in the following financial year. While the specifics are complex, the main principle is to reimburse hospitals at a rate that is close to the national average cost of providing treatment for each specific HRG patient. The intention behind this benchmarking approach is to generate cost reduction incentives (see Shleifer 1985, Savva et al. 2016). Since the reported costs are critical for hospital reimbursement, it is of paramount importance that they are reliable and comparable across hospitals. To ensure that this is the case, hospitals are issued with extensive guidelines on how to allocate direct, indirect, and overhead costs to different HRGs (e.g. HFMA 2016) and the UK Department of Health commissions regular independent audits. In 2010, halfway through our observation period, the UK Audit Commission, a statutory corporation that performs regular audits of public bodies in the UK, performed a comprehensive audit of the data accuracy of seven years of NHS reference cost submissions (UKAC 2011). The report concluded that "most trusts' reference costs submissions were accurate in total." Nevertheless, the report also noted that "the accuracy of individual unit costs varied and, in some cases, was poor." We address this point in our definition of service lines.

Definition of a Service Line. Although each individual HRG can be thought of as a distinct service line, we have chosen to define service lines at a coarser level for two reasons. First, HRG codes are updated annually and have become more granular over time; the number of HRG codes in our data increases every year, from 1,147 in 2006/07 to 2,432 in 2014/15, leading to a total of 4,744 unique HRG codes in our data. To account for this change in coding over time, we map these 4,744 codes to a set of 496 HRG roots – using a publicly available data source intended for this purpose (HSCIC 2015) – which group similar HRGs together. These are then combined into 15 clinically meaningful core HRG chapters that correspond to the major body systems, e.g. nervous or respiratory system, or to particular medical specialties, e.g. obstetrics or cardiac conditions. Although two identical patients in different years may be assigned different HRG codes or, to a lesser extent, different HRG roots, it is unlikely that they would be allocated to different HRG chapters. The HRG chapters therefore provide time-consistent clusters of patients with related conditions, which we define as the service lines.

The second reason for choosing this higher level of aggregation has to do with concerns about the reliability of cost allocations at the individual HRG level. Cost allocation conventions for specific HRG codes *within* HRG chapters can vary significantly between hospitals, but any such deviations within chapters average out when aggregated to the chapter level, leading to considerably more consistent cost allocations at the HRG chapter level. This was confirmed by a former director of costing at the UK Healthcare Financial Management Association, the main advisory body for the financial governance of hospitals in the UK. To further alleviate concerns about the reliability of cost accounting, we corroborate the results of the costing analysis with a LOS analysis; LOS does not suffer from accounting errors (as patient admission and discharge dates are easy to capture) and is highly correlated with hospital costs.

We note that a similar aggregation approach to that described above has been adopted in related empirical research (e.g. Greenwald et al. 2006, Clark 2012, Clark and Huckman 2012). A list of the service lines (i.e. HRG chapters) included in this study appears in the caption of Figure 2.

Definition of Admission Type. For every HRG, costs, volume, and LOS are reported separately for three patient admission types: (1) day cases, (2) elective inpatients, and (3) emergency (non-elective) inpatients. In contrast to emergency admissions, elective inpatient and day-patient admissions are scheduled in advance, with the former including at least one overnight stay in a hospital bed. We merge day cases and elective inpatients since they are often substitutable (and an increasing number of planned procedures can be performed as either), leaving two admission types: electives (El) and emergencies (Em).

Note that elective and emergency patients can be assigned to the same service line (HRG chapter) but, importantly for our analysis, the costs, LOS and activity data are reported separately for each admission type. Finally, note that due to a coding convention that makes it difficult to distinguish between elective and emergency patients, we have removed the service line for obstetric services from the sample.

Unit of Analysis. The final unit of analysis is the admission-type–service-line–trust-year, for which we have a sample of 15,354 observations for each of the two admission types across 14 service lines and 1,097 trust-years, structured as a four-level (non-nested) panel. To simplify the analysis, we investigate each admission type (emergency or elective) separately, reducing the panel to three levels. The full sample of 15,354 observations is used for the analysis of emergency admissions, while 15 observations are dropped – leaving 15,339 – from the analysis of elective admissions due to no patients being observed in the corresponding service lines in these trust-years.

4.1. Dependent Variables

The main dependent variables in this study are the average costs per patient for emergency and elective hospital admissions. As discussed above, we complement this analysis with an additional measure, the average LOS per patient for the two admission types. For the purposes of our study we adjust the average cost and LOS per admission-type–service-line–trust-year for: (i) regional cost variation, (ii) case-mix variation, i.e. the mix of individual HRGs within a service line (HRG chapter), and (iii) any variation due to temporal shocks to costs that are common within an admission-type–service-line across all hospital trusts.

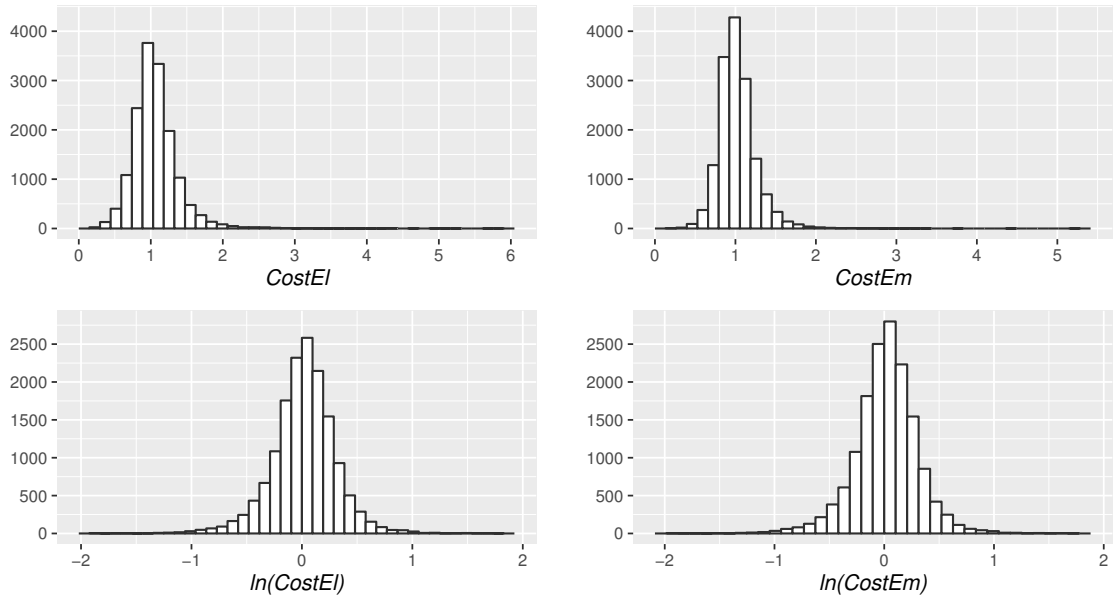
We adjust for regional differences as costs may vary due to local factors outside the hospital trusts' control, e.g. regional differences in the cost of wages, land, and buildings. We do this by adjusting the reported average costs per patient using a government-produced market forces factor (MFF) designed for this purpose (Monitor 2013). The MFF is a scalar unique to each hospital trust and year that is used to weight its costs based on the level of unavoidable spending faced relative to other trusts. Specifically, the regionally adjusted cost for a patient of admission type $p \in \{El, Em\}$ assigned to HRG code c in hospital trust h and year t is equal to $\frac{\text{cost}_{thcp}}{m_{th}}$, where cost_{thcp} are the costs reported in the data and m_{th} is the MFF of trust h in year t.

A hospital's average cost per patient in service line C is the average of its costs across individual HRG codes $c \in C$ weighted by the relative volume of patients in $c \in C$. Since these relative volumes vary across hospitals, and hospitals that treat a larger proportion of patients from high-cost HRGs within a service line are likely to also have higher average costs per patient for that service line, the raw average costs are inappropriate for comparing the productivity of hospitals. To avoid such case-mix confounding, instead of adopting a hospital's observed relative volume of patients with each HRG c in service line C as weights, we weight instead using the relative volume of patients in this HRG code across the entire sample of hospitals. Specifically, hospital h's average cost $\mathbf{Cost}_{thCp}$ for patients of admission type $p \in \{El, Em\}$ in service line C and year t is calculated as

$$\mathbf{Cost}_{thCp} = \sum_{c \in C_{thp}} \alpha_{tcp} \frac{\text{cost}_{thcp}}{m_{th}}, \quad \text{with weights} \;\; \alpha_{tcp} = \frac{n_{tcp}}{\sum_{c \in C} n_{tcp}}, \tag{1}$$

where n_{tcp} is the total number of patients of admission type p with HRG c in year t across all hospital trusts and C_{thp} is the subset of HRGs c in service line C for patients of admission type p that are observed in trust h in year t. We perform a similar weighting procedure to calculate the case-mix-adjusted average LOS. Any differences in average costs or LOS between hospital trusts and/or service lines that are not accounted for by the regional and case-mix adjustment methods are captured through an appropriate control structure in the econometric models.

Figure 1 Histograms of cost: Average cost ratios (top) for elective (left) and emergency (right) admissions and for the natural logarithm of the respective ratios (bottom).



We further adjust costs for potential temporal shocks to the cost of treating patients of a specific admission-type–service-line that are common across all hospital trusts. This adjustment aims to reduce variability in costs due to macroeconomic factors, such as inflation, or changes in guidance or regulation that are common to all hospital trusts and that may render specific service lines more (or less) costly. We make this adjustment by dividing $\mathbf{Cost}_{thCp}$ by the system-wide expected average costs, which are calculated by replacing the costs at hospital trust h in equation (1) with the average cost calculated across a time-invariant set of reference trusts, T_h. The reference trusts for hospital trust h are those, excluding trust h (so that the relationship between costs and expected costs is not endogenous), that are present in the analysis sample in each year that hospital trust h is in the sample. This ensures that the costs for each hospital trust are always compared with the same set of reference trusts and will therefore not be affected by changes in the set of trusts in the analysis sample. To see how this works, suppose that inflation causes costs to increase by 3% in all hospitals. Expected costs would then also increase by 3%, and controlling for this would therefore remove the inflationary effect. Moreover, if costs are, say, 20% higher in service line A than in service line B, on average, then the expected costs will also be 20% higher in service line A than in service line B and will be captured with this adjustment. A similar adjustment is made for LOS.

In summary, differentiating between elective and emergency admissions, we obtain the four dependent variables: $CostEl$ and $CostEm$, the regionally, case-mix-, and temporally adjusted

Table 2 Descriptive statistics and correlation table

		Descriptive statistics				Correlation table						
	Variable	Mean	SD	Min	Max	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Avg. elect. cost (£1,000)	$CostEl$	1.20	0.54	0.10	6.88	0.45*	0.47*	0.25*	0.04*	0.05*	0.13*	0.13*
(2) Avg. emerg. cost (£1,000)	$CostEm$	1.47	0.64	0.15	6.71		0.42*	0.74*	0.06*	-0.14*	0.11*	0.06*
(3) Avg. elect. LOS (days)	$LOSEl$	1.50	0.46	1.00	8.04			0.51*	-0.06*	0.09*	0.04*	-0.03*
(4) Avg. emerg. LOS (days)	$LOSEm$	3.73	1.66	1.00	16.96				-0.10*	-0.16*	-0.01	-0.09*
(5) Elect. service vol. (1,000 patients)	$nElS$	2.64	3.01	0.00	23.34					0.38*	0.31*	0.34*
(6) Emerg. service vol. (1,000 patients)	$nEmS$	2.99	3.00	0.01	23.57						0.37*	0.39*
(7) Elect. hospital vol. (1,000 patients)	$nElH$	38.21	19.70	6.68	117.94							0.82*
(8) Emerg. hospital vol. (1,000 patients)	$nEmH$	51.54	25.12	8.30	168.63							

All correlation coefficients significant with $^{*}p < 0.001$, otherwise $p > 0.10$.

average costs per elective and emergency patient, respectively, and $LOSEl$ and $LOSEm$, the average case-mix- and temporally adjusted LOS for elective and emergency patients, respectively. The distribution of the cost variables is shown in Figure 1. We note that other specifications of the dependent variables, with different (or indeed no) adjustments, lead to qualitatively similar results.

4.2. Independent Variables

To investigate the hypotheses outlined in Table 1 we use four measures of volume: the volume of (i) elective, $nElS$, and (ii) emergency, $nEmS$, activity within a service line (the focal service line) and the volume of (iii) elective, $nElH$, and (iv) emergency, $nEmH$, activity from all service lines *other than* the focal service line. Volume refers to the total number of patient admissions per annum.

Summary statistics for costs, LOS, and service line and hospital volume, reported separately for the elective and emergency patient segments, appear in Table 2.

4.3. Econometric Specification

We organize the data in two distinct panels: one for emergency and one for elective patients. Each observation within a panel belongs to two (non-nested) levels: the service line and the hospital trust. Time is a third level. In this section, we review a series of panel-based models that can be used to identify the impact of volume on costs. These are the pooled regression, fixed-effect (FE) and random-effect (RE) regressions, and within–between volume decomposition methodology in the multilevel modeling (MLM) literature. We present the models for the costs of elective patients; the equivalent models for emergency costs or for LOS can be formulated by replacing the dependent variables accordingly.

The simplest model that can be used to examine the impact of volume on costs is pooled regression (Hsiao 2015). In this model, the average cost per patient in year t at hospital trust h and in service line C is linked to the volume variables according to the following equation:

$$ln(CostEl_{thC}) = \alpha_0 + \beta P_t + \beta_1^h ln(nElH_{thC}) + \beta_1^s ln(nElS_{thC})$$

$$+ \beta_2^h ln(nEmH_{thC}) + \beta_2^s ln(nEmS_{thC}) + \epsilon_{thC}\,, \tag{2}$$

where the variables P_t are a set of time fixed effects (FEs) that control for temporal variation that is common to all service-lines–hospital-trusts. Note that in the model formulation above we take the logarithmic transformation of both the dependent and volume-related independent variables in order to symmetrize the residuals (see Figure 1) and improve the interpretability of the estimated coefficients.

The pooled-regression model assumes that there is no systematic variation across service lines or hospital trusts and that errors are independent and identically distributed (iid). While the second of these assumptions can be somewhat ameliorated by using clustered errors (e.g. by clustering on hospital trusts), the first cannot be easily addressed with this model. Instead, we could estimate a FE model where, in addition to the time FEs in equation (2), we add hospital trust, P_h, and service line, P_C, FEs or, alternatively, service-line–hospital FEs, P_{hC}:

$$\begin{aligned} ln(CostEl_{thC}) = \alpha_0 + \beta^t P_t + \beta^h P_h + \beta^C P_C + \beta_1^h ln(nElH_{thC}) + \beta_1^s ln(nElS_{thC}) \\ + \beta_2^h ln(nEmH_{thC}) + \beta_2^s ln(nEmS_{thC}) + \epsilon_{thC}\,, \end{aligned} \tag{3}$$

$$\begin{aligned} ln(CostEl_{thC}) = \alpha_0 + \beta^t P_t + \beta^{hC} P_{hC} + \beta_1^h ln(nElH_{thC}) + \beta_1^s ln(nElS_{thC}) \\ + \beta_2^h ln(nEmH_{thC}) + \beta_2^s ln(nEmS_{thC}) + \epsilon_{thC}\,. \end{aligned} \tag{4}$$

The first FE model allows for a different intercept for each hospital trust and service line, thus correcting for any variation that affects all service lines within a hospital trust and all hospital trusts within a service line. However, it fails to account for any variation that is service-line–hospital-trust specific (e.g. a specific service line being cost efficient in an otherwise inefficient hospital trust). The second FE model does account for such variation by estimating a different intercept for each category (e.g. service-line–hospital-trust). This, however, comes at a cost. If there are relatively few observations within each category, as is the case in our panel where each category has at most nine observation years, then the standard errors of the FE estimates are likely to be large, leading to unreliable estimates and data overfitting – the large number of parameters that need to be estimated absorbs the variation in the data. This is the opposite of the problem with the pooled regression model, which ignores all variation between categories and therefore underfits the data (see Gelman and Hill 2007). Furthermore, as we expand on later, this model, which relies on the variation in costs and volume within a hospital (across time), is more likely to capture the effect of the higher utilization of existing assets rather than the impact that an increase in volume

may have on costs through different asset configurations. The latter is more likely to be present in between-hospital variation and is arguably of more interest.

An alternative approach is the RE model, which offers the flexibility to trade off these two disadvantages and provide a better fit for the data. This is done by explicitly modeling the variation between the different categories j as a draw from a normal distribution (Bafumi and Gelman 2007):

$$ln(CostEl_i) = \alpha_{(thC)[i]} + \beta_1^h ln(nElH_i) + \beta_1^s ln(nElS_i) + \beta_2^h ln(nEmH_i) + \beta_2^s ln(nEmS_i) + \epsilon_i\,, \quad (5)$$

where the intercept is given by

$$\alpha_{(thC)[i]} = \lambda Teach_i + \boldsymbol{\gamma}^T Region_i + \alpha_{(t)[i]} + \alpha_{(h)[i]} + \alpha_{(C)[i]} + \alpha_{(th)[i]} + \alpha_{(tC)[i]} + \alpha_{(hC)[i]}\,. \quad (6)$$

Using the notation recommended in Gelman and Hill (2007), the index $(thC)[i]$ denotes the time, t, hospital trust, h, and service line, C, corresponding to observation i, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncratic error term. It is the specification of the varying intercept, $\alpha_{(thC)[i]}$, that makes this model more flexible than traditional FE or pooled regression techniques. The terms $\alpha_{(x)[i]}$, where $(x)[i]$ takes values $(t)[i]$, $(h)[i]$, $(C)[i]$, $(th)[i]$, and $(tC)[i]$, denote the time, hospital trust, service line, time–hospital-trust, and time–service-line random effects (REs), respectively, which are all assumed to be Normal random variables with a standard deviation to be estimated.[1] Furthermore, in this model we can also add higher-level variables (i.e. variables that would have been collinear with the FEs included in the model given in (4)), which may reduce the (unexplained) variability in the random error. We have added two such variables in the model above: $Teach_i$, which is a binary variable taking the value 1 if the hospital trust corresponding to observation i has teaching status and 0 otherwise, and $Region_i$, which indicates which of the 10 UK regions (so-called "strategic health authorities") the hospital belongs to.

The RE model above assumes that the random intercepts are not correlated with the independent variables (i.e. the volume). If this assumption is violated (e.g. if there are unobservable factors such as "management quality" that make a hospital more likely to have both high cost realization and high volume), then the estimated coefficients would suffer from heterogeneity bias. Furthermore, the errors would be unreliable (Hsiao 2015). For this reason, FE models are sometimes preferable to the RE model. A more elegant solution to this problem is to explicitly model (and therefore correct) for the dependence of the random intercept and the independent variable using the Mundlak (1978) within–between formulation in the MLM literature.

[1] In the reported models, instead of REs we estimated FEs for time and service line as the number of categories (nine years and 14 service lines) makes the RE estimation qualitatively similar to that for FE (Gelman and Hill 2007). The results are similar if we leave these as REs instead.

Although within–between MLMs are frequently used in other fields, they are less common in the operations management literature, despite their numerous advantages (Bell and Jones 2015). The within–between MLM is similar to the RE model described above with two differences: The effect of volume is decomposed into (1) the within-hospital trust variation, by subtracting from the observed volume in each year the hospital trust average volume calculated across all years (i.e. each volume measure is mean centered around the hospital trust average) in the cost regression equation, and (2) the between-hospital trust variation, which is represented by the hospital trust average,[2] which enters the random intercept equation:

$$\begin{aligned} ln(CostEl_i) = \alpha_{(thC)[i]} &+ \beta_1^h \left[ln(nElH_i) - \overline{ln(nElH_i)}\right] + \beta_1^s \left[ln(nElS_i) - \overline{ln(nElS_i)}\right] \\ &+ \beta_2^h \left[ln(nEmH_i) - \overline{ln(nEmH_i)}\right] + \beta_2^s \left[ln(nEmS_i) - \overline{ln(nEmS_i)}\right] + \epsilon_i \,, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \alpha_{(thC)[i]} = \beta_1^b \overline{ln(nElH_i)} + \beta_2^b \overline{ln(nElS_i)} + \beta_3^b \overline{ln(nEmH_i)} + \beta_4^b \overline{ln(nEmS_i)} \\ + \lambda Teach_i + \boldsymbol{\gamma}^T Region_i + \alpha_{(t)[i]} + \alpha_{(h)[i]} + \alpha_{(C)[i]} + \alpha_{(th)[i]} + \alpha_{(tC)[i]} + \alpha_{(hC)[i]} \,. \end{aligned} \tag{8}$$

As in the RE model given in equation (6), the errors terms ϵ_i and $\alpha_{(x)[i]}$ in equation (8) are assumed to be normally distributed with a variance that is to be estimated. This formulation has a number of advantages, including the fact that correct standard errors are automatically estimated without resorting to error clustering (Bell and Jones 2015). Moreover, this model corrects for the potential correlation between the random intercept and the independent (volume) variables by including the average of these variables explicitly in the model (Mundlak 1978).

An additional advantage of this formulation for our study is that the within-effect coefficients β_i^h (which are similar to the FE coefficients in the model given in (4) that exploit variation in volume within a service-line–hospital-trust) and the between-effect coefficients β_i^b (that exploit variation in volume between service-line–trusts) are likely to capture two distinct effects of volume on costs. First, an increase in volume may allow a hospital to improve the utilization of existing assets, leading to productivity gains (i.e. lower costs) that may come at the expense of worse access, longer waiting

[2] To demonstrate how the hospital trust average volume is calculated, suppose for an observation i the corresponding hospital trust is given by h, the service line by C, and the year by t, with I_{hC} as the set of all observations j that share the same service line C and hospital trust h as observation i. Then, $ln(nElS_i)$ is the logged volume of elective admissions to hospital trust h in service line C in year t and $\overline{ln(nElS_i)} = \frac{\sum_{i \in I_{hC}} ln(nElS_i)}{|I_{hC}|}$ is the average volume in service line C at hospital trust h over the observation period, where $|I_{hC}|$ denotes the cardinality of set I_{hC}. Thus, $\left[ln(nElS_i) - \overline{ln(nElS_i)}\right]$ is the within-effect, i.e. the difference between the observed elective volume in year t and the average volume over all years, and $\overline{ln(nElS_i)}$ is the between-effect, i.e. the average volume of elective activity observed in the specified service line at the chosen hospital trust.

times, and, as shown empirically, worse patient outcomes (see e.g. Kuntz et al. 2014). Second, with higher volume, larger hospitals may be able to afford *different* assets and process structures and deploy them in ways that small hospitals cannot emulate economically (Posnett 2002). The latter volume effect is our primary effect of interest as the additional flexibility and/or better learning effects that volume affords in the configuration of a hospital's asset structure may allow the hospital to be more efficient (i.e. leading to lower costs) as well as more effective (i.e. leading to better patient outcomes) – a win–win situation. The formulation above helps distinguish between the two effects of volume as the within-coefficients β_i^h are more likely to capture the effect of an increase in volume through the better utilization of existing assets, while the between-coefficients β_i^b capture the effect of volume that is, to a large extent, due to different asset configurations. A particularly fortunate feature of our dataset that helps bolster this claim is that the asset configuration of UK hospitals is likely to have remained relatively stable during the observation period: In the wake of the economic crisis, the national government decided essentially to freeze the NHS budget in real terms, despite continuously increasing demand pressure (NAO 2011, HMT 2015, NT 2016), making it very difficult for hospitals to find the capital to invest in significant changes to asset structures. Instead, they had to sweat their existing assets. Therefore, a productivity improvement *within* a specific hospital or service line over the observation period can be largely attributed to the increased utilization of existing capacity. Once this longitudinal effect within hospitals and service lines is controlled for, the residual volume effect between hospitals is more likely to reflect the volume–productivity effects of interest.

Finally, in the models for the electives (emergencies) we also control for the relative proportion of elective (emergency) services offered in a hospital in a particular year within the focal service. This is calculated as:

$$\mathbf{Prop}_{thCp} = \sum_{c \in C_{thp}} \alpha_{tcp}\,, \tag{9}$$

where α_{tcp} is given in equation (1) and C_{thp} is the subset of HRGs c in service line C for patients of admission type p observed in hospital trust h in year t. This control accounts for the possibility that in addition to volume, costs may depend on the variety of treatments available (see also the discussion on endogenous service line formation in §6.2). While not the focus of this study, more detail on this control – together with estimated coefficients, robustness tests on, and model estimations without its inclusion – are provided in §5 of the supplementary material.

An estimation of the models presented above is made in R (version 3.2.3) using the `lmer()` function of the `lme4` package, with model parameters calculated using restricted maximum likelihood estimation (Bates et al. 2015).

5. Results

We present the estimation of within–between RE (MLM) regressions for costs and LOS in Table 3. For reference, the results of the more commonly employed pooled, FE, and RE regressions (i.e. without within–between decomposition) using the models (2), (3), and (5), respectively, are provided in Table 4. While we do not comment on them directly, we note that the estimated coefficients, especially of the RE models (in which volume is not decomposed to within and between variation), are consistent with those in Table 3.

The upper two panels of Table 3 report coefficient estimates and standard errors of the within-effects and between-effects of the MLM regressions for cost and LOS. Since the dependent and independent variables have been log-transformed, the coefficients can be interpreted as the approximate percentage change in average costs/LOS for a doubling of the respective volumes. The third panel ("control structure") of Table 3 reports which factors are included as FE and the estimated standard deviations (σ_z) of the RE. The lower panel ("model fit") reports the marginal R^2, which describes the proportion of variance explained by non-random factors (e.g. the volume variables) alone, and the conditional R^2, which describes the proportion of variance explained by both the non-random and random factors. (The marginal and conditional R^2 are calculated using the method in Johnson 2014.)

We focus on the interpretation of the between-effect coefficients, which, as explained above, are likely to capture the effect resulting from the increased flexibility that higher volume provides in terms of economically viable asset structures and configurations as opposed to the cost-effect of utilizing existing assets better. We note that the latter effect, which is captured by the within-effect coefficients, is largely as one would expect: As assets become more utilized, the average costs and LOS within a service line decrease, with a much smaller (or no) effect across service lines.

The main results, based on the between-effect coefficients, are also summarized qualitatively in Table 5, which complements the summary of hypothesized effects provided in Table 1 in §3. Note that the reported coefficients of volume effects within a service line and in other service lines (e.g. Elect. vol. (focal SL) vs. Elect. vol. (other SLs)) are not directly comparable because the aggregate volume of patients from the other service lines will typically be much larger than the volume of patients from the focal service line. Therefore, doubling the former constitutes much less of a change in the aggregate volume of the hospital than doubling the latter. To improve comparability, we calculate and report the marginal effects at the mean of a 1,000-patient increase for each of the four volume types in the final column (Column 6) of Table 5. This is achieved by converting the absolute change in volume to a percentage increase in volume at the mean, where the means are

Table 3 Model parameter estimates – RE models using within–between volume decomposition

	Costs		LOS	
	Elective	Emergency	Elective	Emergency
Within-effects				
Elect. vol. (focal SL)	-0.117^{***}	0.007	-0.062^{***}	0.001
	(0.007)	(0.004)	(0.003)	(0.003)
Emerg. vol. (focal SL)	-0.011	-0.162^{***}	$0.010^{\dagger}$	-0.083^{***}
	(0.012)	(0.008)	(0.006)	(0.005)
Elect. vol. (other SLs)	-0.129^{***}	0.070^{*}	-0.039^{*}	0.114^{***}
	(0.031)	(0.029)	(0.016)	(0.027)
Emerg. vol. (other SLs)	0.049	-0.162^{***}	$0.027^{\dagger}$	-0.117^{***}
	(0.032)	(0.031)	(0.016)	(0.029)
Between-effects				
Elect. vol. (focal SL)	-0.057^{***}	0.020^{***}	-0.022^{***}	0.006^{*}
	(0.009)	(0.004)	(0.004)	(0.003)
Emerg. vol. (focal SL)	0.007	-0.121^{***}	0.021^{***}	-0.081^{***}
	(0.014)	(0.011)	(0.006)	(0.008)
Elect. vol. (other SLs)	0.043	0.135^{***}	-0.005	0.069^{**}
	(0.030)	(0.028)	(0.014)	(0.026)
Emerg. vol. (other SLs)	-0.008	-0.099^{**}	0.015	$-0.052^{\dagger}$
	(0.034)	(0.031)	(0.015)	(0.028)
Control structure				
Year	Y	Y	Y	Y
Service line	Y	Y	Y	Y
Hospital	0.072	0.076	0.031	0.073
Hospital:Service line	0.153	0.094	0.065	0.070
Hospital:Year	0.034	0.016	0.020	0.011
Service line:Year	0.080	0.088	0.041	0.092
Residual std. error	0.206	0.138	0.105	0.087
Model fit				
Observations	15,339	15,354	15,339	15,354
Marginal R^2	0.096	0.179	0.107	0.100
Conditional R^2	0.513	0.623	0.463	0.739
Bayesian inf. crit.	481.0	-11,166.4	-20,650.8	-23,870.5

$^{\dagger}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; coefficient estimates and standard errors reported for the within- and between-effects; inclusion of FE in the control structure indicated by "Y," inclusion of RE indicated by the reporting of its estimated standard deviation; marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

as given in Table 2. For example, a 1,000-patient increase in elective volume for the focal service line is equivalent to a $\frac{1}{2.64} \approx 38\%$ increase in volume at the mean.

We first discuss the effect of a change in volume on costs/LOS for elective patients within and across service lines, followed by the equivalent effect for emergency patients.

5.1. Volume Effects for Elective Patients

A doubling in the volume of elective patients in a service line is estimated to reduce the costs for these patients by 5.7% (p-value < 0.001) and the LOS by 2.2% (p-value < 0.001) on average. This result, which is based on the first between-effects row (in Columns 2 and 4) in Table 3, provides strong support for the hypothesis that economies of scale exist for elective patients within a service line (Row 1 of Table 1).

Table 4 Model parameter estimates – without using within–between volume decomposition

	Elective costs			Emergency costs			Elective LOS			Emergency LOS		
	P	F	R	P	F	R	P	F	R	P	F	R
Main effects												
Elect. vol. (focal SL)	-0.019^{***}	-0.083^{***}	-0.096^{***}	0.018^{***}	0.013^{**}	0.015^{***}	-0.020^{***}	-0.037^{***}	-0.046^{***}	0.005^{***}	0.006	0.004^{*}
	(0.003)	(0.011)	(0.005)	(0.001)	(0.005)	(0.003)	(0.001)	(0.005)	(0.003)	(0.001)	(0.004)	(0.002)
Emerg. vol. (focal SL)	−0.003	0.006	0.003	-0.025^{***}	-0.134^{***}	-0.144^{***}	0.015^{***}	0.021^{**}	0.019^{***}	-0.015^{***}	-0.090^{***}	-0.081^{***}
	(0.002)	(0.014)	(0.009)	(0.002)	(0.013)	(0.007)	(0.001)	(0.007)	(0.004)	(0.001)	(0.010)	(0.004)
Elect. vol. (other SLs)	0.001	-0.175^{***}	−0.025	0.165^{***}	0.044	0.133^{***}	0.003	-0.061^{**}	−0.014	0.098^{***}	0.129^{***}	0.099^{***}
	(0.008)	(0.036)	(0.020)	(0.006)	(0.038)	(0.019)	(0.004)	(0.019)	(0.010)	(0.005)	(0.031)	(0.017)
Emerg. vol. (other SLs)	0.049^{***}	0.037	0.067^{**}	-0.186^{***}	-0.213^{***}	-0.098^{***}	0.023^{***}	0.020	0.034^{**}	-0.133^{***}	-0.165^{***}	-0.086^{***}
	(0.009)	(0.042)	(0.023)	(0.007)	(0.050)	(0.021)	(0.005)	(0.020)	(0.011)	(0.005)	(0.041)	(0.019)
Model fit												
Observations	15,339	15,339	15,339	15,354	15,354	15,354	15,339	15,339	15,339	15,354	15,354	15,354
Adjusted R^2	0.024	0.177	–	0.106	0.321	–	0.044	0.175	–	0.051	0.320	–
Marginal R^2	–	–	0.099	–	–	0.181	–	–	0.122	–	–	0.109
Conditional R^2	–	–	0.519	–	–	0.625	–	–	0.469	–	–	0.742
Bayesian inf. crit.	5,544.0	4,338.1	471.1	−3,953.9	−6,768.0	−11,200.2	−16,712.1	−17,549.3	−20,653.1	−11,769.0	−15,471.5	−23,930.2

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; "P", "F" and "R" columns refer to the pooled, FE, and RE regression models, respectively; coefficient estimates and standard errors (clustered by hospital for the fixed columns) reported for the main effects; in the RE models, marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

An increase in emergency volume in the same service line does not have a significant effect on the cost of providing treatment for the elective patients in the service line. The LOS of elective patients, however, increases by an estimated 2.1% (p-value < 0.001) each time the emergency volume in the service line doubles. These results can be seen in the second between-effects row (in Columns 2 and 4) in Table 3. The effect of emergency volume on elective LOS can perhaps be explained by emergency patients having priority and, therefore, elective services becoming more likely to be disrupted when there are more emergency patients in the service line. Nevertheless, the fact that the cost of elective patients does not change as the emergency volume in the service line increases suggests that this increase in LOS is offset by other benefits associated with increased volume (e.g. fixed-cost pooling).

Turning to the effect of an increase in volume from other service lines, we find that this does not have a significant impact on the costs or LOS of elective patients in the focal service line. This is the case irrespective of whether the increase comes from elective or emergency patients in other service lines. This can be seen from the third and fourth between-effects rows (in Columns 2 and 4) in Table 3. This result suggests that for elective patients any benefits associated with higher volume across service lines are offset by the additional complexity of treating heterogenous patients.

5.2. Volume Effects for Emergency Patients

Similar to the effect of volume on elective patients, we find that an increase in emergency volume in a service line reduces the costs and LOS for these patients significantly and by an estimated 12.1% (p-value < 0.001) and 8.1% (p-value < 0.001), respectively, on average. This result, which can be seen in the second row (in Columns 3 and 5) of Table 3, provides strong support for the

Table 5 Hypothesized and estimated direction of volume effects on productivity

Effect on...	of an increase in...	from the...	Aggregate hypothesis	Estimated effect	Approximate effect size on costs[(1)]
Elective productivity	Elective vol.	Focal SL	↑	↑↑***	−2.2%
		Other SLs	?	–	–
	Emergency vol.	Focal SL	?	–	–
		Other SLs	?	–	–
Emergency productivity	Elective vol.	Focal SL	?	↓***	+0.8%
		Other SLs	?	↓***	+0.4%
	Emergency vol.	Focal SL	↑	↑↑***	−4.0%
		Other SLs	?	↑***	−0.2%

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; [(1)]effect on costs is approximated by adding 1,000 patients (from the service line(s) and admission type in the corresponding row) to the mean volume level given in Table 2; ↑ denotes a positive effect; ↑↑ denotes a strongly positive effect; ↓ denotes a negative effect; ↓↓ denotes a strongly negative effect; – denotes no effect; ? denotes an ambiguous overall effect.

hypothesis that economies of scale exist for emergency patients within a service line (Row 7 of Table 1). Furthermore, and in contrast to the results reported above for elective patients, there are significant economies of scope in emergency care: On average, a doubling of the emergency volume in *other service lines* is estimated to reduce the costs of emergency patients in a focal service line by 9.9% (p-value < 0.001) and to reduce the average LOS for these patients by 5.2% (p-value $= 0.064$). This can be seen in the fourth row (in Columns 3 and 5) of Table 3. This result shows that the shared components of emergency services, such as an ED, generate substantial positive economies of scale not only within but also across service lines.

Turning to the effect of elective volume on emergency patients, we find that both the cost of providing care to emergency patients and their LOS increase as the volume of elective patients within and across the service line increases. This can be seen in the first between-effects row (in Columns 3 and 5) in Table 3 where, on average, costs increase by 2.0% (p-value < 0.001) and LOS increases by 0.6% (p-value $= 0.031$) when elective volume within a service line is doubled. Similarly, the third between-effects row (in Columns 3 and 5) in Table 3 shows that a change in the elective volume in other service lines has a negative effect on emergency productivity: When the elective volume in other service lines doubles, the cost of emergency patients in the focal service line increases by an estimated 13.5% (p-value < 0.001) and LOS increases by an estimated 6.9% (p-value $= 0.007$). Importantly, these results suggest that the increased complexity associated with treating elective patients in parallel with emergency patients completely offsets any volume-related benefits of such collocation for emergency patients.

6. Robustness

To confirm the robustness of the results presented in §5 we extend the model to allow the volume effects to vary by service line, discuss potential reverse causality, and describe the findings from a number of other model specifications. More details on these additional analyses is presented in the supplementary material.

6.1. Random slopes

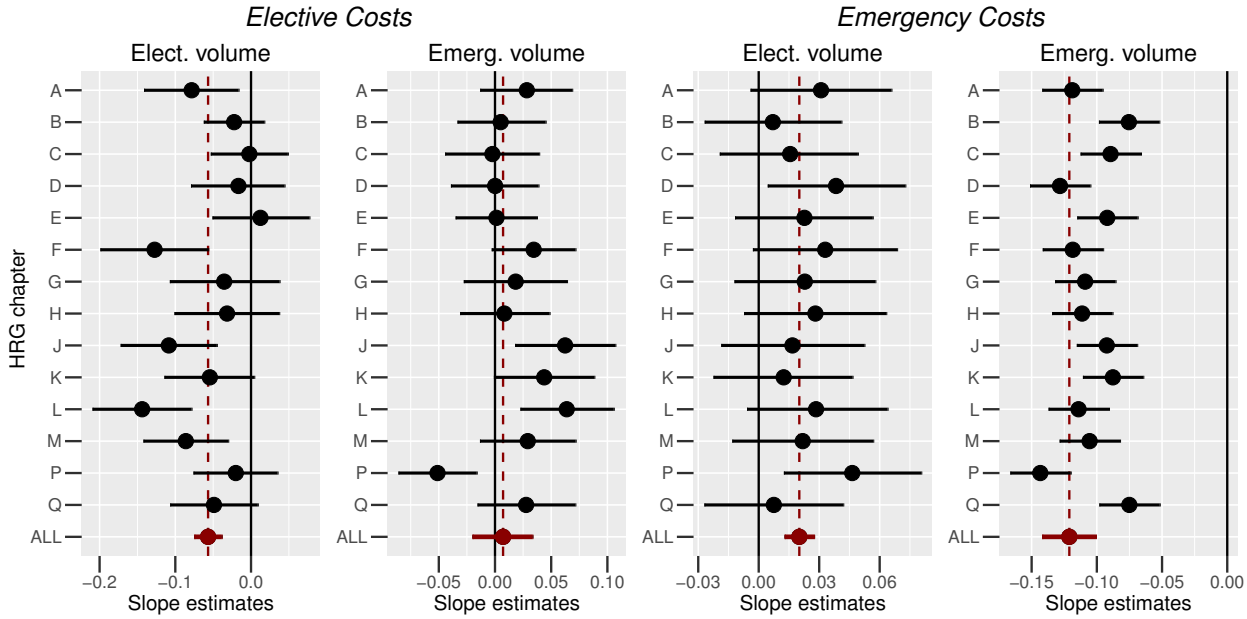
In the models presented in the previous section we estimated the average impact of volume on costs and LOS across different service lines, implicitly assuming that this impact of volume was homogeneous across the different service lines. We can relax this assumption and allow for heterogeneous slopes for each of the service lines by estimating a model where, in addition to random intercepts, we *also* allow for random slopes. (We discuss and present results here for random slope estimates for Elect. vol. (focal SL) and Emerg. vol. (focal SL), with results for volume in other service lines being similar and reported in §2 in the supplementary material.) To do this we replace the coefficients β_2^b and β_4^b in equation (7) with random, service line-dependent slopes $\beta^b_{2,(C)[i]}$ and $\beta^b_{4,(C)[i]}$, respectively. It is typical in the MLM literature to allow the random slopes to be correlated with the service line-specific intercepts. To achieve this we need to also replace the service-line FE in equation (8) – which we use in place of RE (see footnote 1) – with the RE, $\alpha_{(C)[i]}$. We then model $(\alpha_{(C)[i]}, \beta^b_{2,(C)[i]}, \beta^b_{4,(C)[i]})$ using a trivariate normal distribution to allow for correlation between the RE (see Gelman and Hill 2007, for details). This is a more flexible approach than adding interaction terms between the service lines and volume effects of interest, although the interpretation is similar.

Figure 2 shows the random slope estimates of the between-effects in the cost models, together with bootstrapped 95% confidence intervals (using 10,000 simulations from the posterior distribution). The separate slopes that are derived for each service line give an estimate of the service line-specific between-effects of elective and emergency volume on cost. These can be compared with the combined slope estimates from §5, which are also plotted (as "ALL") in Figure 2. Comparing these, it can be seen that the directions of the service line-specific effects are consistent with the combined estimates, with 95% confidence intervals overlapping in nearly all cases. Due to limited data, the confidence intervals are wide for these service line-dependent random slopes, and, as such, in presenting the main results we instead report the aggregate effects across service lines.

6.2. Reverse Causality

In this paper we have argued that higher volumes confer a productivity advantage. However, the direction of causality is not obvious: It could be argued instead that the positive relationship

Figure 2 Random slope coefficient estimates for the effect of volume in the focal service line on costs, reported by service line (black) and combined (red), with bootstrapped 95% confidence intervals.



Note. A – nervous system; B – eyes and periorbita; C – mouth, head, neck, and ears; D – respiratory system; E – cardiac surgery and primary cardiac conditions; F – digestive system; G – hepatobiliary and pancreatic system; H – musculoskeletal system; J – skin, breast and burns; K – endocrine and metabolic system; L – urinary tract and male reproductive system; M – female reproductive system; P – diseases of childhood and neonates; Q – vascular system.

identified between volume and productivity is actually the result of more productive hospitals being referred a higher volume of patients or patients self-selecting these hospitals. Since April 2008, all patients in England have been able to choose to be treated at any NHS-funded hospital (Dixon et al. 2010). As such, it is possible for patients to choose to be treated in more productive hospitals. This may especially be the case in larger cities, such as London, where a high density of hospitals allows patients to more readily access care from multiple providers. While we cannot rule this out empirically with the available data, there are several arguments that suggest that this is not the case.

First, as health services in the UK are free at the point of care, there is little incentive for a patient to select their care provider based on cost. Indeed, such information is not made readily available. However, while patients are unlikely to decide based on cost, it is possible that they will select based on quality. As cost and quality are often correlated, and quality is an unobserved factor that we do not account for in this analysis, this could be driving the results. Information on the quality of hospitals, however, has not been readily available until recently, and it remains challenging for patients to compare treatment for procedures at different hospitals. Moreover, past research has shown that there is little, if any, evidence of patients (or their physicians) exercising such choice (e.g. Gaynor et al. 2004, Gowrisankaran et al. 2006). Patient surveys indicate that hospital choice

is not highly valued by UK patients. In a study at the start of our observation period, patients were asked to rate statements that described aspects of their care by importance. The statement "I have a choice about which hospital I am admitted to" ranked 76 out of 82 statements (Boyd 2007). A more recent study by the King's Fund, an independent UK-based healthcare think tank, reports that only half of patients in England referred to a hospital by their general practitioner in 2008/09 were offered a choice of where their treatment would take place (else being referred to their local provider), and of those who were, nearly 70% selected their local provider. The study also found that most hospital trusts operated in a defined geographical market and only competed for patients "at the boundaries of their catchment areas, where another provider was equidistant" (Dixon et al. 2010). Putting this together, there is little reason to suspect that UK patients choose to go (or are referred) to more productive providers.

Nevertheless, we investigate this further by re-running the analysis using a subset of the data corresponding to those hospital trusts that are geographically more isolated, with a restriction that the nearest trust can be no closer than 20 km away. This has the effect of removing all hospital trusts located in cities and other more densely populated regions and, thus, reducing the number of trust–year observations by 63%, from 1,097 to 406. While this does not completely avoid the problem of selection, the selection effect should be weaker in this subsample (as it is more inconvenient for a patient to attend another provider), especially for emergency patients, who need to be treated quickly. Therefore, if reverse causality were driving our results, then we would expect to find weaker evidence of productivity improvements from pooling similar types of activity when using this sample. The results (available in §6.1 of the supplementary material) show that this is not the case, with coefficient estimates nearly identical in sign and scale.

Another plausible type of reverse causality is selection by hospitals: Certain hospitals may choose to offer a subset of elective and/or emergency services (i.e. treat patients with a subset of conditions/HRGs only), and the choice of which services they offer may well depend on the profitability of these services. We have already partially accounted for this in our models by controlling for hospital–service-line effects as well as for the proportion of services, $\mathbf{Prop}_{thCp}$, offered by a hospital within each service line in each year. Nevertheless, if service lines were formed endogenously in the way described above, then we might expect hospitals that offer fewer services to also be more profitable. In §5 of the supplementary material we show that there is little evidence of endogenous selection for emergency patients. For elective patients, we find that those hospitals that operate at higher volumes are less, not more, selective and offer a greater variety of services. If endogenous service line formation were driving our results, we would therefore expect to find effects in the opposite direction to those we observe.

Together, this evidence suggests that the effects identified are unlikely to be the result of reverse causality and that it is more likely that higher volumes of same-type activity improve productivity, rather than the reverse.

6.3. Other Robustness Checks and Modeling Alternatives

One concern when working with panel data is that errors may be autocorrelated, leading to underestimation of the standard errors of the estimated coefficients when autocorrelation is positive and potentially biasing the estimated coefficients in the within–between formulation (Hsiao 2015). Examining the relationship between the residuals (at time t) and the lagged residuals (at time $t-1$) and performing formal hypothesis testing with the Baltagi–Wu LBI test statistic, we find no evidence of such an effect (refer to §3 of the supplementary material for further details).

Another possibility we consider is that there may be non-linear effects of volume on costs. Although the models we estimate are already non-linear (as they involve the logarithmic transformations of both the dependent and independent variables) and suggest diminishing returns to scale (as the estimated coefficients are all < 1 and > -1), we also estimate models in which we add a squared-volume term for each of the between-effects. We find all estimates to be consistent in both sign and magnitude, with the exception of the effect of same-service-line emergency volume on emergency costs, for which the sign remains the same but the estimated effect size is reduced by $\sim 50\%$ (reported in §4 of the supplementary material).

In addition to the models discussed above, we estimate a number of alternative model specifications that (i) cap costs at the HRG level to reduce the influence of outliers, capping below at 1/5th or 1/10th of and above at 5 or 10 times the system-wide median, (ii) only compare costs and LOS for a subset of HRGs for which treatment in each year is provided in at least 80% of the hospital trusts in the sample, and (iii) constrain the sample to only include those service-line–hospital-trusts with a minimum volume level (e.g. >25% of the system-wide median) in order to reduce the potential influence of outliers. Since some trusts operate multiple hospital sites (with typically one large, main hospital and one or more smaller hospital sites), we also repeat the analysis for the subset of trusts with a single hospital site. The results of these estimations are reported in §6 of the supplementary material and are qualitatively and quantitatively similar to those in §5 of this paper.

7. Managerial and Policy Implications

From a productivity perspective, the prevailing model of the fully comprehensive general hospital is predicated on the assumption that there are economies of scale and scope that come from pooling planned (elective) and unplanned (emergency) patient services and from pooling different

service lines. However, our data and analysis suggest that there are in fact significant diseconomies associated with the pooling of elective and emergency patients. Specifically, an increase in elective services is associated with a significant productivity drop in emergency services. Furthermore, while the collocation of different service lines provides economies of scope for emergency patients, there is no evidence of positive productivity spillovers between service lines for elective patients. These findings have important implications for hospital growth strategies and the configuration of hospital systems, which we explore further in this section through counterfactual analyses.

First, when hospitals consider growth strategies they have to be aware that while increasing elective activity improves the productivity of their elective patients it has a negative impact on emergency activity, not only within the service line that is growing but also for emergency patients in other service lines. To illustrate this, consider the model-predicted effect of different growth strategies for a major London NHS hospital trust, Barts Health, which admitted about 250,000 elective and emergency patients per annum over the final three financial years in our dataset at a total cost of £1bn across the three years. We estimate the impact of increasing total patient admissions by 20% to 300,000 per annum as a result of one of three strategies: (i) a 20% expansion across the board in elective and emergency volume, (ii) a 33% increase in emergency activity only, or (iii) a 55% increase in elective volume only, where growth causes the volume in each service line to increase by the same proportion. Using the modeling results from §5, we estimate that in the first scenario elective costs would fall by 1.1% and emergency costs, by 1.2%, leading to a total cost saving of £12m per annum. The emergency growth strategy would have no effect on elective costs but would reduce emergency costs by 6.4%, leading to a total cost saving of £40m. Finally, the elective growth strategy would reduce elective costs by 2.5% but would have the unintended consequence of a 6.8% *increase* in emergency costs, leading to a total cost *increase* of £31m. The negative spillover across all emergency services quickly erodes the productivity benefits of higher volume in elective services. This finding is surprising and important: The majority of hospitals in the UK are in deficit in the 2015/16 financial year and most chief executives see a growth in elective activity, which is easier to plan and has less variation in costs, as the preferred way of increasing productivity to turn their hospital around. Few hospital managers would consider expanding their emergency activity. Surprisingly, our data suggest that an elective growth strategy can be counterproductive if the hospital has high emergency volume and that such hospitals may actually be better off trying to increase their emergency activity instead.

The second important implication of our findings relates to regional hospital systems. Our results suggest that removing elective volume from general hospitals and instead treating these patients in

regional *focused factories* should improve productivity for both the re-routed elective patients and the emergency patients remaining in the downsized general hospitals. To investigate the possible cost savings at the regional level, we present the results of a counterfactual analysis based on a plausible re-organization of elective services in London. We assume that any two hospital trusts in the city might come to a mutual agreement to redistribute their elective services in such a way that there is no duplication of service lines between the two hospitals. We then estimate the cost implications arising from the increase in elective volume within each service line. In order to minimize the need for additional capacity investment, we match hospital trusts pairwise based on their size, with the match made by pairing trusts that are most similar in terms of their total elective volume. Using the new allocation and estimated coefficients reported in §5, we calculate that for the trust-years in our analysis the total cost of providing elective care would be reduced by 4.1% (from £7.72bn to £7.40bn). Note that the cost savings could potentially be greater if (i) more than two hospital trusts worked together and (ii) the reallocation was based not only on volume but also on costs (so that the increased elective volume would be routed to the cheapest hospital). The implication of this finding is that even simple regional reorganization may result in substantial cost savings.

Our findings also reconcile two seemingly opposing trends: (1) for small general hospitals to be closed or downgraded to urgent care centers and activity moved to larger general hospitals in the proximity and (2) for greater specialization with the opening of specialist hospitals focusing on only particular types of conditions. Interestingly, we show that these trends may not be at odds and that the cost of providing care to different types of patients may be reduced through these different approaches. In particular, the productivity of elective care would benefit if elective patients were treated in specialist hospitals or regional treatment centers focused on specific service lines. We estimate, for example, that if London were to operate 14 such *focused factories* for each of the 14 service lines, then costs could be reduced further from £7.72bn to £6.54bn: a saving of 15.3%. In addition, emergency patients would benefit from being treated in large, general acute hospitals that focus primarily on emergency care and treat a full spectrum of services. Implementing different service delivery modes for planned and unplanned activity could therefore be a highly effective way of increasing the productivity (and quality – see e.g. RCS/DH (2007), Kuntz et al. (2015)) of hospital services in the longer term.

8. Conclusion

We use a unique longitudinal dataset to investigate economies of scale and scope in hospitals. While theory and prior empirical work offer strong support for scale economies within service lines,

which we confirm, there has been little prior work on the spillover effects of volume increases for admission types (elective or emergency) and service lines (e.g. cardiology, urology, etc.). Traditional arguments, based on spreading fixed costs, statistical scale economies, and learning and experience, suggest positive spillover effects, while increased organizational complexity and, specific to healthcare, the fact that elective care gets disrupted by emergency activity when a hospital becomes busy, suggest potential negative spillovers between admission types and service lines. We use our data to explore this theoretical tension and find that there are indeed significant negative cost spillovers between elective and emergency care, with an increase in elective activity rendering emergency activity within the same service line as well as in other service lines considerably less productive. We believe that this is a consequence of the increased organizational emphasis on standardization and programmed coordination mechanisms, which, as argued elsewhere, is at odds with the flexibility and informal coordination required for effective response in emergency care (Argote 1982, Christensen et al. 2009, Kuntz et al. 2015). We also find that while emergency service productivity increases across service lines when volume in a focal service line increases, suggesting significant economies of scope in emergency services, there is no evidence of such economies of scope in elective services.

As with all multi-firm studies based on accounting costs, our analysis has limitations due to the unobserved degree of adherence of individual hospital cost accounting systems to the national guidelines. We believe that our aggregation of the granular HRG codes to which costs are allocated to the coarser level of HRG chapters as our service lines of interest helps alleviate this problem as accounting inaccuracies within service lines average out at the aggregate level and accounting misallocations between service lines are less likely. In addition, we corroborate our findings with an analysis of LOS, which is unaffected by hospital accounting systems but highly correlated with costs, and which confirms our results.

Our observations have implications for productivity-enhancing growth strategies for hospitals. While many general hospital managers pursue a strategy of increasing the overall productivity of their hospital by expanding more easily manageable elective services, our findings caution them to consider the unintended side effect of increased emergency costs, which may well erode any productivity gains in elective services. Our findings also suggest that general hospitals would be more efficient if they focused on emergency activity, with elective patients being treated instead in high-volume regional *focused factories*. From a productivity perspective, this supports the widely discussed redesign paradigm for regional hospital systems with separate "solutions shop hospitals," focused on unplanned work that requires trial and error and decision-making "on the spot," and "value-adding process clinics" that provide standardized treatments at high volume (Christensen et al. 2009).

References

AHRQ (2014) National healthcare disparities report. Technical Report AHRQ Publication No. 14-0006, Agency for Healthcare Research and Quality.

Aletras V (1997) Part II: The relationship between volume and the scope of activity and hospital costs. *NHS Centre for Reviews and Dissemination* 8(2).

Argo J, Vick C, Graham L, Itani K, Bishop M, Hawn M (2009) Elective surgical case cancellation in the Veterans Health Administration system: Identifying areas for improvement. *The American Journal of Surgery* 198(5):600–606.

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

Argote L (2013) *Organizational learning: Creating, retaining and transferring knowledge* (New York: Springer Science & Business Media), 2nd edition.

ASGBI (2007) Emergency general surgery: The future. Technical report, Association of Surgeons of Great Britain and Ireland, London.

Bafumi J, Gelman A (2007) Fitting multilevel models when predictors and group effects correlate, Paper presented at the Annual meeting of the American Political Science Association, Philadelphia, PA.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1–48.

Batt R, Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science (Forthcoming)* .

Begg C, Cramer L, Hoskins W, Brennan M (1998) Impact of hospital volume on operative mortality for major cancer surgery. *JAMA* 280(20):1747–1751.

Bell A, Jones K (2015) Explaining fixed effects: Random effects modeling of time-series cross- sectional and panel data. *Political Science Research and Methods* 3(1):133–153.

Best T, Sandıkçı B, Eisenstein D, Meltzer D (2015) Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* 17(2):157–176.

Birkmeyer J, Siewers A, Finlayson E, Stukel T, Lucas F, Batista I, Welch H, Wennberg D (2002) Hospital volume and surgical mortality in the United States. *New England Journal of Medicine* 346(15):1128–1137.

Boh W, Slaughter S, Espinosa J (2007) Learning from experience in software development: A multilevel analysis. *Management Science* 53(8):1315–1331.

Bohmer R (2009) *Designing care: Aligning the nature and management of health care* (Boston, MA: Harvard Business School Press).

Boyd J (2007) The 2006 inpatients importance study, Picker Institute Europe, Oxford, UK.

Chan C, Farias V, Escobar G (2016) The impact of delays on service times in the intensive care unit. *Management Science (Forthcoming)* .

Christensen C, Grossman J, Hwang J (2009) *The Innovator's Prescription: A Disruptive Solution for Health Care* (New York: McGraw-Hill).

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

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

Clark J, Huckman R, Staats B (2013) Learning from customers: Individual and organizational effects in outsourced radiological services. *Organization Science* 24(5):1539–1557.

DH (2013) A simple guide to payment by results. Technical report, Department of Health, UK, URL `https://www.gov.uk/government/publications/simple-guide-to-payment-by-results`, Accessed: 2016-03-09.

Dijk N, Sluis E (2004) To pool or not to pool in call centers. *Production and Operations Management* 17(3):296–305.

Dixon A, Appleby J, Robertson R, Burge P, Devlin N, Magee N (2010) *Patient choice: How patients choose and how providers respond* (London, UK: The King's Fund).

Dranove D (1998) Economies of scale in non-revenue producing cost centers: Implications for hospital mergers. *Journal of Health Economics* 17(1):69–83.

Fetter R (1991) Diagnosis related groups: Understanding hospital performance. *Interfaces* 21(1):6–26.

Freeman M, Savva N, Scholtes S (2016) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science (Forthcoming)* .

Gaughan J, Mason A, Street A, Ward P (2012) English hospitals can improve their use of resources: An analysis of costs and length of stay for ten treatments. Technical report, Centre for Health Economics, University of York, London, CHE Research Paper 78.

Gaynor M, Seider H, Vogt W (2004) Volume-outcome and antitrust in U.S. health care markets, Unpublished manuscript, Carnegie Mellon University.

Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical models* (New York, NY, USA: Cambridge University Press).

Gillen S, Catchings K, Edney L, Prescott R, Andrews S (2009) What's all the fuss about? Day-of-surgery cancellations and the role of perianesthesia nurses in prevention. *Journal of PeriAnesthesia Nursing* 24(6):396–398.

Gowrisankaran G, Ho V, Town R (2006) Causality, learning and forgetting in surgery, Unpublished manuscript.

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

Greenwald L, Cromwell J, Adamache W, Bernard S, Drozd E, Root E, Devers K (2006) Specialty versus community hospitals: Referrals, quality, and community benefits. *Health affairs* 25(1):106–118.

Guerin-Calvert M (2011) Assessment of cost trends and price differences for U.S. hospitals. Technical report, Federal Trade Commission, Working paper no. 294.

HFMA (2016) Acute health clinical costing standards. Technical report, Healthcare Financial Management Association, URL `https://www.hfma.org.uk/docs/default-source/our-work/costing/Clinical-Costing-Standards/acute-standards-201617.pdf`, Accessed: 2016-03-09.

HMT (2015) HMT public expenditure statistical analyses (PESA). Technical report, HM Treasury, UK, URL `https://www.gov.uk/government/collections/public-expenditure-statistical-analyses-pesa`, Accessed: 2016-03-12.

Hopp W, Lovejoy W (2012) *Hospital operations: Principles of high efficiency health care* (FT Press).

Hopp W, Spearman M (2004) To pull or not to pull: What is the question? *Manufacturing & Service Operations Management* 6(2):133–148.

HSCIC (2015) HRG4 2014/15 consultation grouper roots, Health & Social Care Information System, Accessed: 2016-03-09.

Hsiao C (2015) *Analysis of panel data* (Cambridge, UK: Cambridge University Press), 3rd edition.

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

Huckman R, Staats B (2011) Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management* 13(3):310–328.

Hurst J, Williams S (2012) Can NHS hospitals do more with less? Technical report, Nuffield Trust, London.

Jestin P, Nilsson J, Heurgren M, Påhlman L, Glimelius B, Gunnarsson U (2005) Emergency surgery for colonic cancer in a defined population. *British Journal of Surgery* 92(1):94–100.

Johnson P (2014) Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution* 5(9):944–946.

Joustra P, Van der Sluis E, Van Dijk N (2010) To pool or not to pool in hospitals: A theoretical and practical comparison for a radiotherapy outpatient department. *Annals of Operations Research* 178(1):77–89.

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

KC D, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Kim S, Chan C, Olivares M, Escobar G (2014) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* 61(1):19–38.

Kristensen T, Olsen K, Kilsmark J, Pedersen K (2008) Economies of scale and optimal size of hospitals: Empirical results for Danish public hospitals, Unpublished manuscript, University of Southern Denmark.

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* 61(4):754–771.

Kuntz L, Scholtes S, Sülz S (2015) Separate and concentrate: Accounting for process uncertainty in the design of general hospitals, Working paper, Cambridge Judge Business School.

Li G, Rajagopalan S (1998) Process improvement, quality, and learning effects. *Management Science* 44(11 Pt 1):1517–1532.

March J, Simon H (1958) *Organizations* (New York: Wiley).

Marini G, Miraldo M (2009) Economies of scale and scope in the English hospital sector, Unpublished manuscript, University of York.

Monitor (2013) A guide to the market forces factor. Technical report, Publication code: IRG 31/13.

Monitor (2015) Improving productivity in elective care. Technical report, URL `https://www.gov.uk/guidance/improving-productivity-in-elective-care`, Accessed: 2016-03-09.

Moore F (1959) Economies of scale: Some statistical evidence. *The Quarterly Journal of Economics* 73(2):232–245.

Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica: Journal of the Econometric Society* 46(1):69–85.

NAO (2011) Delivering efficiency savings in the NHS. Technical report, National Audit Office, UK.

Narayanan S, Balasubramanian S, Swaminathan J (2009) A matter of balance: Specialization, task variety, and individual learning in a software maintenance environment. *Management Science* 55(11):1861–1876.

NCEPOD (2010) An age old problem: A review of the care received by elderly patients undergoing surgery. Technical report, National Confidential Enquiry into Patient Outcome and Death, London.

Nembhard I, Tucker A (2011) Deliberate learning to improve performance in dynamic service settings: Evidence from hospital intensive care units. *Organization Science* 22(4):907–922.

NT (2016) NHS in numbers. Technical report, Nuffield Trust, URL `http://www.nuffieldtrust.org.uk/nhs-numbers-0`, Accessed: 2016-03-12.

Panzar J, Willig R (1977) Economies of scale in multi-output production. *The Quarterly Journal of Economics* 91(3):481–493.

Penrose E (1959) *The Theory of the Growth of the Firm* (New York: John Wiley).

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

Porter M (1979) How competitive forces shape strategy. *Harvard Business Review* 137–145.

Porter M, Teisberg E (2006) *Redefining health care: Creating value-based competition on results* (Boston, MA: Harvard Business Press).

Posnett J (2002) Are bigger hospitals better? McKee M, Healy J, eds., *Hospitals in a changing Europe*, chapter 6, 100–118 (Buckingham: Open University Press).

Powell A, Savin S, Savva N (2012) Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management* 14(4):512–528.

Ramdas K, Saleh K, Stern S, Liu H (2014) New joints more hip? Learning in the use of new components, Working paper, London Business School.

RCS/DH (2007) Separating emergency and elective surgical care: Recommendations for practice. Technical report, The Royal College of Surgeons of England, Department of Health, London.

RCS/DH (2010) The higher risk surgical patient: Towards improved care for a forgotten group. Technical report, The Royal College of Surgeons of England, Department of Health, London.

Rothkopf M, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Operations Research* 35(6):906–909.

Sanjay P, Dodds A, Miller E, Arumugam P, Woodward A (2007) Cancelled elective operations: An observational study from a district general hospital. *Journal of Health Organization and Management* 21(1):54–58.

Savva N, Tezcan T, Yildiz O (2016) Can yardstick competition reduce emergency department waiting times?, Working paper, London Business School.

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

Schoar A (2002) Effects of corporate diversification on productivity. *The Journal of Finance* 57(6):2379–2403.

Schuster M, Neumann C, Neumann K, Braun J, Geldner G, Martin J, Spies C, Bauer M, Group CS (2011) The effect of hospital size and surgical service on case cancellation in elective surgery: Results from a prospective multicenter study. *Anesthesia & Analgesia* 113(3):578–585.

Shleifer A (1985) A theory of yardstick competition. *The RAND Journal of Economics* 319–327.

Skinner W (1974) The focused factory. *Harvard Business Review* 113–121.

Song H, Tucker A, Murrell K (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

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

Thompson J (1967) *Organizations in action: Social science bases of administration* (New York: McGraw-Hill).

UKAC (2011) Improving coding, costing and commissioning: Annual report on the payment by results data assurance programme 2010-11. Technical report, Audit Commission UK, URL `http://collections.europarchive.org/tna/20121205001956/http://audit-commission.gov.uk/SiteCollectionDocuments/Downloads/pbrannualreport2011.pdf`, Accessed: 2016-03-09.

Vanberkel P, Boucherie R, Hans E, Hurink J, Litvak N (2012) Efficiency evaluation for pooling resources in health care. *OR Spectrum* 34(2):371–390.

Wedig G, Hassan M, Sloan F (1989) Hospital investment decisions and the cost of capital. *Journal of Business* 62(4):517–537.

West P (1998) Future hospital services in the NHS: One size fits all? Technical report, The Nuffield Trust, London, Nuffield Occasional Papers, Health Economics Series: Paper No. 6.

Wilson P, Carey K (2004) Nonparametric analysis of returns to scale in the US hospital industry. *Journal of Applied Econometrics* 19(4):505–524.