Gatekeepers at Work: An Empirical Analysis of a Maternity Unit

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

We use a detailed operational and clinical dataset from a maternity hospital to investigate how workload affects decisions in gatekeeper-provider systems, where the servers act as gatekeepers to specialists but may also attempt to serve customers themselves, albeit with a probability of success that is decreasing in the complexity of the customer's needs. We study the effect of workload during a service episode on gatekeepers' service configuration decisions and the rate at which gatekeepers refer patients to a specialist. The data shows that gatekeeper-providers (midwives in our context) make substantial use of two levers to manage their workload (measured as patients per midwife): They ration resource-intensive discretionary services (epidural analgesia) for customers with non-complex needs (mothers with spontaneous onset of labor) and, at the same time, increase the rate of specialist referral (physician-led delivery) for customers with complex needs (mothers with pharmacologically induced labor). The workload effect in the study unit is surprisingly large and comparable in size to those for leading clinical risk factors: When workload increases from two standard deviations below to two standard deviations above the mean, low-complexity patients are 28.1% less likely to receive an epidural, leading to a cost reduction of 8.6%, while high-complexity patients are 11.8% more likely to be referred for a physician-led delivery, leading to a cost increase of 2.4%. These observations are consistent with overtreatment at both high and low workload levels, albeit for different types of patients, and highlight the importance of workload smoothing in the context of gatekeeper-provider systems.

Key words: Gatekeeper Systems; Workload Management; Health care : Hospitals; Service Operations; Econometrics

History: June 17, 2015

1. Introduction

In many service settings (e.g. healthcare, call centers, maintenance and restaurants) customers interact with a server (e.g. nurse, telephonist, engineer or waiter) who acts as a gatekeeper, i.e. decides whether to refer the customer to a specialist (e.g. doctor, service manager or sommelier), and who may also attempt to provide a service to the customer herself, albeit with a probability of success that is decreasing in the complexity of the customer's needs (Shumsky and Pinker 2003). In deciding whether to refer or self-serve the customer, the gatekeeper-provider (GP) makes a

trade-off between the desire to protect specialists' valuable time and the cost of failing to resolve the customer's problem herself. This implies a system-optimal referral rate that depends on (i) the cost of failure to solve the customer's problem and (ii) the distribution of the complexity of the customer's needs. The GP referral problem has been studied analytically in the operations management and economics literature, with emphasis on healthcare applications (for more details, see §2). A central assumption in this literature is that both the referral rate and the type of service offered by the GP are independent of the system load, i.e. a GP under load-induced pressure refers at the same rate and provides the same type of service as a GP who does not face such pressure. There is, however, extensive evidence to suggest that worker behavior is not immune to changes in conditions in the work environment (Boudreau et al. 2003). Despite this evidence, there is limited empirical research into how the work environment affects GP behavior and, in particular, whether referral rates and service configuration decisions are indeed independent of workload. This paper aims to fill this gap.

An example of such a service setting – and the motivation for this paper – is the delivery unit (DU) of a UK maternity hospital, where births take place. The GP in this case is the midwife, who, as the primary carer assigned to the delivering mother, makes decisions (together with the patient) about aspects of the delivery process (e.g. delivery and pain management methods) and whether to refer to a specialist physician for further interventions (e.g. instrumental delivery or emergency cesarean section (C-section)). Due to cost-cutting efforts over the past few years, maternity wards across the UK have experienced an increase in workload, i.e. periods where the number of mothers delivering is greater than the number of midwives present have become more frequent (Clover 2010). Through its influence on GP behavior, this increase in workload is believed to have given rise to fundamental changes in the types of deliveries performed and, as a consequence, in patient outcomes. This work uses detailed data over five years (16,316 births) and appropriate econometric models to investigate whether this is indeed the case.

Building on existing theory, we hypothesize that both referral and service configuration decisions are affected by GP workload. In particular, at high workloads we expect that GPs will be more likely to refer patients to an expert and that those customers served by GPs will receive less resource-intensive service configurations. Both of these actions help GPs reduce their workload. Whether a customer is referred or served directly, albeit at a less resource-intensive level, is determined by the complexity of their needs: Complex cases are more likely to be referred, while less complex cases are more likely to receive rationed services. Indeed, after we control for the non-random assignment of patients to interventions using appropriate econometric models and instrumental variables, we find strong support for all of these hypotheses in the context of the DU: As the patient–midwife ratio increases from two standard deviations below to two standard deviations above the mean

(0.40 to 1.88 patients per midwife), low-complexity patients, defined as those who present with the spontaneous onset of labor, are approximately 28.1% less likely to receive a resource-intensive pain management intervention, while high-complexity patients, defined as those whose labor was induced in the hospital prior to their arrival at the DU, are approximately 11.8% more likely to be referred to a specialist for an interventional delivery. Interestingly, the magnitude of the effect of workload on pain management methods and interventions is comparable to that of leading clinical factors, e.g. maternal diabetes or length of gestation period.

From a theoretical perspective, this work identifies two distinct workload buffers available to GPs: (i) specialists, whose time GPs should protect but who eventually receive a larger proportion of customers as GP workload increases and (ii) service configuration, which, instead of being dictated by customer needs and preferences, is rationed as GP workload increases. Interestingly, there seems to be a dichotomy in the use of these buffers: Customers tend to be either referred or rationed more aggressively as workload increases but not both. Which of these buffers is used at high workload appears to depend on patient characteristics: Less complex cases have their service configuration rationed – in our setting we observe a reduction in resource-intensive pain management interventions – and complex cases are referred more frequently to a specialist provider – in our setting this translates into more physician-led deliveries. This last observation suggests that workload shifts the trade-off between protecting specialists' valuable time and the cost of failing to resolve customers' problems in the direction of the latter: Specialists' time becomes less valuable to the GP when she is busy. These empirical observations suggest that the modeling literature on the GP problem, which ignores the presence of both of these endogenous buffers, needs updating. For example, conclusions regarding (i) commonly used staffing rules (e.g. Borst et al. 2004), (ii) economic contracts used to outsource GP activities (e.g. Lee et al. 2012), and (iii) the use of GPs to induce downstream specialist competition (e.g. Brekke et al. 2007) all assume state-independent service/referral rates and may no longer be valid in the presence of state-dependent referral rates.

From a practical perspective, this work provides a methodological framework that helps us understand how GPs adapt their behavior as workload increases, which has implications for costs and staffing policies. More specifically, the rationing of discretionary components of service for patients with less complex needs leads to a cost reduction: In this case, the rationing-related reduction in the cost of treating patients at two standard deviations below to two standard deviations above the average workload is approximately £200 per patient (or -8.6%). Surprisingly, the rationing effect does not necessarily lead to uniformly worse outcomes for these patients. In fact, non-complex patients experience a reduction in maternal post-birth length of stay (LOS), a key measure of quality of care. By contrast, the workload-induced increase in referrals, which only affects high-complexity patients, increases intervention rates and the associated costs: In this case, there is weak

statistical evidence of an increase of approximately £56 per patient (or +2.4%). This suggests that the overall impact of workload on measures such as throughput or cost may be context specific: In environments where cases are more likely to be complex, periods of high workload may be more costly due to the increased number of referrals, while the converse may be true in environments with a large proportion of low-complexity cases due to the rationing effect. Furthermore, the methodology developed in the paper can be used to predict how changes in GP staffing levels, through the impact on GP workload, affect outcomes and costs. In the context of the DU that we study, our work suggests that adding more GPs would, on average, lead to a net increase in the cost of patient treatment. For example, if the hospital increased staffing to achieve the target of not more than one patient per midwife for 95% of deliveries (NAO 2013), this would add approximately £0.8m to the annual staffing bill and, as a result, epidural analgesia, physician-led deliveries and post-birth LOS would change by +7.8%, +2.6%, and +1.6%, respectively, for low-complexity patients and by +0.8%, -2.4%, and -0.3%, respectively for high-complexity patients. This workload-induced change would increase the cost of treating low-complexity patients by £53K p.a. (or 6.3% on top of the added staffing cost), which would be partially offset by a reduction in the cost of treating high-complexity patients by £37K p.a. (4.5% of the added staffing cost).

In the more general healthcare context, our findings on the impact of workload on GP behavior may also have some bearing on the unnecessary care phenomenon. Unnecessary care, which by some estimates is as high as 30% (Smith et al. 2012), is defined as the dispensing of diagnostic or treatment services that provide no demonstrable benefits to patients. Our results are consistent with patients being overtreated at high workloads (through increased specialist referrals) and at low workloads (through the increased provision of discretionary services). Therefore, operational interventions that smooth out GP workload (e.g. flexible staffing plans) have the potential to reduce unnecessary care. For example, if the hospital in question moved to a perfectly flexible staffing plan, it could achieve the target of not more than one patient per midwife for 95% of deliveries at a cost of only £0.3m p.a. and, equally importantly, reduce the higher rates of epidural analgesia, physician-led deliveries and post-birth LOS and their associated costs by 20–30%.

We note that our study focuses on a specific setting: the DU of a maternity hospital. This setting is important in its own right as childbirth is the most common cause of hospital admission and accounts for 2.8% of all healthcare expenditure in the UK (NAO 2013) and approximately 1.4% of expenditure, or $40B p.a., in the US.[1] Nevertheless, this might raise concerns regarding the generalizability of our findings to other service settings (e.g. call centers or equipment maintenance). We concur but note that the nature of the setting, which is governed by strict professional guidelines

[1] Authors' calculation, based on 2012 US figures: $9,775 average cost per birth (Rosenthal 2013), 4M babies born (Hamilton and Sutton 2013) and healthcare expenditure of $2.8T (Martin et al. 2014).

involving specialized and highly trained personnel, makes it less and not more likely to see workload-induced deviations in GP behavior than in less formal/consequential settings staffed by a less well-trained work force. We also note that all staff in the setting we study are salaried employees whose remuneration is not linked to performance. Our results are therefore not confounded by economic incentives, which may alter decision-making. As such, care should be taken when generalizing our findings to contexts where GP decisions are subject to individual economic incentives.

2. Literature Review

Our research relates to two strands of literature: (i) research that explores the gatekeeper paradigm for service delivery and (ii) econometric investigations on the effect of workload on system performance.

The two-tier system, where the first tier acts as a gatekeeper for the second tier, has been studied extensively in healthcare economics and operations management. In the former, this paradigm has been employed to model the relationship between patients and primary care physicians (PCPs), who act as gatekeepers for specialized care. PCPs serve to protect specialists' resources but are subject to informational frictions (González 2010, Mariñoso and Jelovac 2003, Malcomson 2004, Brekke et al. 2007). This research tries to identify conditions under which the gatekeeper system is preferable to one without a gatekeeper and to design contracts that shape PCP incentives to minimize the impact of asymmetric information. In fact, in order to focus on informational frictions, this work abstracts away the detailed flow dynamics that are inevitably present in such a service setting. By contrast, work on the gatekeeper model in the operations management literature focuses explicitly on such service dynamics. The first analysis of the two-tier system in operations management was the modeling work of Shumsky and Pinker (2003), who derive the optimal referral rate given deterministic customer inter-arrival and service times and propose incentive structures that induce system optimal gatekeeping behavior in a principal–agent setting. Hasija et al. (2005) extend these results to a stochastic system, while Lee et al. (2012) use the same framework to explore the problem from an outsourcing perspective, where one or both tiers are outsourced to a profit-maximizing third-party vendor.

For tractability purposes, the gatekeeper literature makes two assumptions: (i) gatekeeper referral rates and (ii) the types of service offered to customers by the gatekeeper are independent of system load. Either of these assumptions has been relaxed in single-tier models, where the server is either a gatekeeper that routes the customer without providing any part of the service or the server performs no gatekeeping function. For example, Alizamir et al. (2013) relax the first assumption by developing a dynamic model to study how system congestion affects the number of investigations a gatekeeper performs before deciding whether to refer a customer to a specialist. The paper shows

that in this setting the gatekeeper usually compromises diagnostic accuracy and therefore makes errors in the referral decision in order to increase the speed at which customers are processed. This work, however, focuses on gatekeepers' triage decisions and does not explicitly consider the possibility that gatekeepers may attempt to serve customers themselves. By contrast, there is a complementary stream of literature that focuses explicitly on the service dimension, which it models as being endogenous to workload. Hopp et al. (2007) present a model that shows that the service configuration decision may be affected by workload, i.e. discretionary aspects of the service may be removed. Debo et al. (2008) show that revenue-maximizing severs may find it optimal to reduce service rates at low workloads, a result further explored in Anand et al. (2011) and Kostami and Rajagopalan (2013). Similarly, Paç and Veeraraghavan (2015) show that expert servers, who have an informational advantage over their customers, have an incentive to overtreat and that congestion moderates this tendency. This stream of work, however, focuses on a single-tier model and cannot therefore analyze whether workload affects referral processes. Our work contributes by presenting an integrated empirical validation of these two workload-independence assumptions in the two-tier gatekeeper context. As we show, referral and service configuration decisions jointly act as buffers for workload variability, albeit for different types of customers. Furthermore, we show that these systematic deviations from what is typically assumed have a material impact on managerial decisions such as staffing.

Our work also contributes to the growing body of literature that empirically examines how human behavior deviates from that assumed by classic operations management models (see Boudreau et al. (2003) and Bendoly et al. (2006) for excellent summaries of the literature). For example Schultz et al. (1998) performed a series of laboratory experiments to show that worker behavior, and worker productivity in operations management settings in particular, is affected by environmental factors such as individual and system workload. Our work belongs to a more recent stream of literature that aims to confirm and expand on experimental findings by using observational data from different service environments (e.g. Huckman et al. 2009, Staats and Gino 2012, Kesavan et al. 2014, Ramdas et al. 2014).

The stream of literature that is closest to our work investigates how workload affects important aspects of individual or system performance. Data availability and the importance of the setting mean that many of these studies focus on healthcare. KC and Terwiesch (2009) use operational data from patient transport services and cardiothoracic surgery to show that workers respond to an increase in workload in the short term by reducing service times. By contrast, Berry-Jaeker and Tucker (2013) show that in the context of inpatient care, very high workload can prolong service times and increase patient LOS. In the emergency care context, Batt and Terwiesch (2012) show that simultaneous speed-up and slow-down mechanisms come into play as workload changes,

with task reduction being counterbalanced by a general slowdown in common treatment processes. In addition to service times, researchers have also studied the relationship between workload and other operational, financial, and service quality metrics. For example, Kuntz et al. (2014) show that elevated workload beyond a safety tipping point is associated with higher patient mortality. Powell et al. (2012) find a reduction in hospital revenue per patient as discharging physician workload increases, and Green et al. (2013) show that nurse absenteeism rates are linked to anticipated workload.[2] Aside from healthcare, Tan and Netessine (2014) find a non-linear effect between the number of diners assigned to waiting staff and staff sales performance in the context of a restaurant chain: Sales initially increase with load as staff become more motivated but ultimately decline as staff place more emphasis on speed. With the last two papers we share an emphasis on the implications of the endogenous response to workload on staffing decisions. More specifically, as in Green et al. (2013), we show that increasing staffing levels may generate a cost saving that (partially) offsets the cost of extra staff (in their case, higher staffing is associated with reduced absenteeism, while in our case, it is associated with a reduction in referrals for complex patients). However, as in Tan and Netessine (2014), higher staffing may also compromise aspects of system performance (in their case, this is associated with lower motivation to cross-sell and up-sell, while in our case, extra staff are associated with an increase in discretionary interventions for non-complex patients). Our study deviates from previous work as (i) we focus on the impact of workload on a two-tier GP system, (ii) we examine two distinct buffers, referral and service configuration, which a GP can use to absorb workload variability, and (iii) we examine how customer characteristics, and complexity in particular, interact with workload.

Finally, our work is also related to Kim et al. (2014) and KC and Terwiesch (2012), who study decisions to admit emergency department (ED) patients to the intensive-care unit (ICU). In the language of the two-tier gatekeeper model, the ED represents the first-tier GP system and the ICU, the second-tier expert system. At higher levels of ICU occupancy the former study finds that the chance of ICU admission is reduced, while the latter identifies an increased chance of being discharged early. Together these papers indicate that the workload of the second-tier expert system affects patient routing decisions and that this has an adverse effect on patient outcomes, as re-routed patients are more likely to require costly readmission to the ICU. In contrast to these papers, our work focuses on the impact of workload at the level of the first-tier GP system as well as the implications this has for customer experience and GP staffing.

[2] These workload studies in the operations literature are complemented by studies in the medical literature, see review by Kane et al. (2007) and more recently Needleman et al. (2011).

3. Clinical Setting

The setting for this study is the DU in the maternity department of a large UK teaching hospital. The DU is the primary location for childbirth and immediate post-natal care and is made up of standard delivery rooms, clinical rooms for high-complexity patients, obstetric theaters, and a recovery bay. The unit is part of a larger maternity department, which also contains an antenatal unit to provide care prior to the onset of labor for patients with problematic pregnancies, a midwifery-led birthing unit, where very-low-risk mothers can give birth in a more natural environment and without physician oversight, a post-natal unit to care for mothers and babies in the period post-birth but before discharge, and a neonatal unit, which specializes in additional care for babies. We study this setting because (i) it is a significant and indispensable part of any healthcare system (as mentioned, childbirth costs $40B p.a. in the US alone), (ii) the job description of the main service provider – the midwife – closely matches that of the GP we want to study, and (iii) the variable and unpredictable nature of arrivals makes midwife workload highly variable (see §5).

The DU deals essentially with two types of patients: scheduled and unscheduled. Scheduled patients, who make up 15.0% of all deliveries, are those admitted for an elective C-section. Elective C-sections are performed in an operating theater attached to the DU by a dedicated team of specialists. For these patients, the date of delivery is pre-booked and the care pathway is locked-in in advance. The remaining deliveries, which take place in the DU itself, are the main focus of our study. Of these patients, 65.7% arrive at the DU directly from home following the spontaneous onset of labor, while the remaining 34.3% are induced at the hospital prior to transfer to the DU. Induction involves one or more of the following procedures (Reed 2011): preparing the cervix with a vaginally administered drug (prostaglandins), artificial rupture of membranes (also known as "breaking the waters"), and inducing contractions of the uterus with a synthetic hormone (oxytocin). Induction is most commonly performed when the pregnancy is overdue, although other factors, such as maternal health, may indicate induction. While induced mothers have their inductions scheduled, they are still considered as unscheduled arrivals at the DU owing to the significant and unpredictable time lag between the commencement of induction and the level of labor progression required for admission to the DU.

The staff working in the DU are, as all hospital staff in the UK, National Health Service (NHS) employees and receive a fixed salary, i.e. their remuneration is not linked to performance or results. The unit in question is staffed by three types of employees:

1. Midwives, who are specialist nurses that have completed a three-year full-time midwifery course. All nursing staff in the study unit are licensed midwives. There are typically eight or nine midwives on duty at any time, although the DU tries to add staff when the number of patients exceeds the number of midwives present.

2. Obstetricians, who are medical doctors. They monitor and treat high-risk women during pregnancy and are available in the DU to perform high-risk births, including C-sections. Senior obstetricians (referred to as consultant obstetricians in the UK) are also involved in the training of junior doctors. Junior obstetricians are present in the unit at all times, while senior doctors are present during working hours (8 a.m. to 6 p.m.) and are on call out of hours.

3. Obstetric anesthesiologists, who are specialists responsible for pain management and anesthesia in the DU and/or DU operating theaters. There is always one anesthesiologist on duty in the DU. When scheduled obstetric activities take place (e.g. elective cesareans), a second is present. There is also an additional anesthesiologist on call.

While the number of midwives on duty is carefully recorded and monitored, the number of doctors and anesthesiologists present is less transparent.

When a patient is admitted to the DU, she is assigned a primary midwife, who is responsible for the well-being of mother and baby throughout labor and childbirth. Once assigned to a patient, the midwife must attend the patient regularly in order to observe the frequency of contractions, monitor fetal and maternal heart rate, record temperature and blood pressure, determine whether a doctor needs to intervene, and perform other related activities. For an uncomplicated birth, the midwife will also perform the delivery, carry out an initial examination of the baby, and provide immediate post-natal care for the mother.

Depending on individual cases, there is a range of interventions that can be used in the DU. The most common of these are epidural analgesia, instrumental delivery, and emergency C-section. All of these interventions are carried out by anesthesiologists and/or physicians. Epidural analgesia is usually administered to improve the patient experience when less invasive pain management methods provide insufficient pain relief. It involves the injection of painkilling drugs into the lower back, which aims to block the nerves and reduce or eliminate labor pain. This form of intervention is typically administered no later than one hour before delivery and must be administered by an anesthesiologist, who assesses suitability based on the progress of labor and any presence of contraindications. The procedure normally takes place within 30 minutes of being requested and takes approximately 20 minutes to perform. Post-provision, a midwife must be with the patient continuously for at least 30 minutes and regularly thereafter in order to take blood pressure and monitor the baby's heart rate to ensure that no complications arise (OAA 2013). The need for specialist doctors and post-procedure supervision makes epidurals highly resource intensive. From a clinical perspective, epidurals can also have disadvantages, such as reducing maternal blood pressure (which may affect the flow of oxygen to the baby), the potential for drugs to cross the placenta (which can affect the baby's breathing and cause drowsiness), slower labor, and increased risk of further interventions (Anim-Somuah et al. 2011).

Instrumental deliveries and/or emergency C-sections are carried out if labor is significantly prolonged or if information becomes available during the progression of labor that elevates the health risk for the mother or baby. The decision to undertake such an obstetric intervention can take place at any point during labor. In an instrumental delivery the baby is delivered vaginally using instruments such as forceps or a vacuum pump. The intervention itself is carried out by an obstetrician, usually in the operating theatre, and takes on average 45 minutes to perform. Emergency C-sections are performed when it becomes clear that the delivery cannot occur vaginally without placing the woman or baby under undue risk. Emergency C-sections are considered major surgeries. They are carried out under regional or, occasionally, local anesthetic and take approximately 1.5 hours to perform. Emergency C-sections carry significant risks for the patient, such as hemorrhage, infection, thrombosis, and an increased risk of complications in subsequent pregnancies as well as prolong post-birth recovery times (Henderson et al. 2001).

After delivery, the mother and baby are monitored in the DU for a short time before being transferred to the post-natal unit, where they recuperate before being discharged. Upon discharge the whole delivery episode is fully costed according to government guidelines using a patient-level information and costing system (DH 2012).

4. Hypothesis Development

A GP service episode consists of two related steps. First, the GP makes an initial diagnosis of the customer's needs and, in consultation with the customer, devises a "service plan", which can be seen as a configuration of tasks to be performed by the GP (in the first instance) to meet the customer's needs. Second, either at the beginning or later in the service episode, when new information might become available, the GP needs to decide, again based on the customer's needs, whether to refer to a specialist, who will then take over and complete the service. Naturally, the decision to refer depends on the complexity of the customer's needs: The GP is less likely to be able to successfully resolve a more complex case, and it is these cases that, all else being equal, are more likely to be referred to a specialist, whose time the GP is tasked with protecting (Shumsky and Pinker 2003).

Since inter-arrival times and service durations in most service settings are stochastic, the GP is subject to time-varying workload, i.e. there are times when there are more customers in the system than GPs. During these high-workload periods, some customers will have to wait for service or, to the extent that parallel processing is possible, will receive only a fraction of the GP's limited capacity. We note that the DU tries to provide one midwife per mother; however, given the highly variable arrival process (see §5) and recognizing the urgent nature of patients' needs, the DU goes into parallel-processing mode, where a single midwife is in charge of more than one delivery simultaneously. Therefore, unless the GP changes the way she serves customers, the mechanics of service

systems suggest that periods of high workload will be associated with delays in customer service (Luo and Zhang 2013, Tan and Netessine 2014). Such delays are associated with poor customer experience, either directly (as customers face costly waiting times (Robinson and Chen 2011)) or indirectly (as customer needs increase if service is delayed (Chan et al. 2015)). Furthermore, excess workload puts pressure on the GPs themselves, as increased workload inevitably generates stress and fatigue (Bendoly et al. 2006). To reduce the adverse impact of excess load, the GP has two natural levers at her disposal: the service configuration decision and the referral decision.

In the following section we discuss the implications of workload for each of these levers in turn. We first frame our discussion in a general service setting and then expound the associated implications for the specific empirical setting of this paper: the DU, where the customer is the expecting mother, the service required is the delivery of the baby, and the midwife assigned to the mother upon arrival at the DU acts as the GP.

4.1. Service configuration decisions

Most types of services have certain components that are indispensable in serving customers' needs. These are the core components of the service, and they cannot be omitted or substituted by other service components without significantly compromising the quality and/or profitability (or even the safety) of the service episode, for which GPs are ultimately responsible. Beyond the core components, some services have additional, more discretionary components (Hopp et al. 2007). Although these non-core components may make a substantial difference to customer experience, they are not directly linked to the primary service outcome and take up GP time and effort. In contrast to core components, discretionary components form a buffer that can be used to protect the core service from the impact of workload variation. When workload increases we therefore expect GPs to use this buffer and ration certain discretionary service components for some customers. This behavior is consistent with the "cutting corners" phenomenon under workload (see Oliva and Sterman (2001)). However, we argue that the corners cut are those that are associated with activities that are not central to the primary service outcomes.

Hypothesis 1. (H1) When workload increases, the likelihood that a GP will include discretionary service components in the service plan decreases.

In the specific context of the DU, the core components of the service provided by the midwife – who takes the role of the GP – are the tasks required to protect the health of the mother and baby. These include following the progress of labor, monitoring the baby's heart rate and providing guidance and support during the final stages of labor and care for the newborn, etc. Components of the service that might be characterized as discretionary are those that are not linked to the health of the mother or baby directly but are more closely associated with the comfort of the patient.

One such component is pain management and, in particular, the provision of epidural analgesia. As discussed in the previous section, this procedure is resource intensive for the midwife because (a) the midwife needs to coordinate with the DU anesthesiologist and prepare the patient for the procedure and (b) the patient's dependence on the midwife increases post-provision (OAA 2013). As a result, any midwife assigned to a patient who has received an epidural is less able to parallel process other delivering mothers. This becomes problematic as the number of patients increases. Therefore, we expect H1 to translate to a reduction in the propensity of epidural analgesia as midwife workload increases.

4.2. Referral decisions

In the absence of congestion, GPs' decisions to refer customers to a specialist should be based on diagnostic evidence about customers' needs. However, as argued by Alizamir et al. (2013), congestion creates the need for GPs to speed up, leading to decisions based on less complete evidence. In a sense, the decision to refer a customer to a specialist becomes an additional lever with which the GP can reduce her workload. In contrast to the service configuration decision, the referral decision involves another service provider besides the GP: the specialist, who needs to be available and willing to take on the customer. If the specialist accepts the referral, the responsibility and a large part of the work required to serve the customer are transferred to that specialist. Therefore, we expect that if the GP is under workload-induced pressure and there is a specialist with spare capacity, the GP will be more likely to refer the customer to the specialist, thus freeing up their own time to tend to the needs of other customers.

Hypothesis 2. (H2) When workload increases, the likelihood that a customer will be referred by the GP to a specialist increases.

In our context, midwives refer mothers for an obstetrician-led birth – either an instrument delivery or an emergency C-section – when information becomes available that renders the service too complex for them to manage safely without physician assistance.[3] Similar to discretionary services, specialists become a buffer that the midwives can use to manage their workload. After controlling for obstetrician workload, we would therefore expect referral rates, and therefore obstetrician-led deliveries, to increase when midwife workload increases.

4.3. The role of customer complexity

Customers in the service system are typically assumed to be heterogeneous in their service needs (Shen and Su 2007). More specifically, some customers will be easy to serve, and the GP will be well placed to do so. Others will exhibit more complex needs that require specialized knowledge

[3] In our setting we bundle together all obstetrician-led deliveries since the decision whether to perform an instrumental delivery or an emergency C-section lies with the physician and not the GP (midwife).

and/or skills that goes beyond the abilities of the GP. Following Shumsky and Pinker (2003), we expect that customers with more complex needs are more likely to be referred to a specialist, who is better suited to resolve their needs. During busy periods, as per H2, the GP will begin to refer patients whose level of complexity may have not justified referral in the absence of excess load. We expect that GPs are more likely on average to refer highly complex patients than low-complexity patients for two reasons. First, high-complexity patients are more likely to benefit from the greater knowledge and skills of a specialist, and GPs may become more aware of their limitations in handling complex cases when under workload pressure. Therefore, workload pressure makes a GP more likely to refer complex cases, which she is uncertain she can handle herself, than less complex cases, which she is more confident in handling. The second reason has to do with the specialist's willingness to take on the customer. If it seems that the referral is without merit, i.e. the case is relatively straightforward, then the specialist may refuse to take on the customer, returning the responsibility to the GP. This is less likely to happen for cases that are complex.

Hypothesis 3. (H3) When workload increases, the increase in specialist referrals is greater for customers with complex needs than for customers with non-complex needs.

Does the degree of complexity of a customer's needs also moderate the rationing response to workload? We believe it does for two reasons. First, it is plausible that a service component that is discretionary (i.e. not critical for service outcomes) in a non-complex case may be less discretionary for a more complex customer, who has, by definition, greater needs. In other words, what is nice-to-have for a customer with basic needs may become a necessity for a customer with complex needs. Second, following the argument preceding H3, the GP has another lever they are more likely to be able use for complex cases: referral to a specialist. Since this lever is less applicable for non-complex customers, rationing becomes a relatively more important workload management method for such cases. In other words, rationing a time-consuming discretionary service component for a customer who is likely to be referred to a specialist will have less of an impact on GP workload than rationing services to non-complex customers, who are more likely to stay with the GP throughout the service episode.

Hypothesis 4. (H4) When workload increases, the reduction in the provision of discretionary service components is more pronounced for customers with complex needs than for customers with non-complex needs.

Together H3 and H4 suggest that there is a divergence in the service experience of patients with complex and non-complex needs as workload increases: The former are more likely to be referred to a specialist while the latter are more likely to experience rationing of discretionary service components.

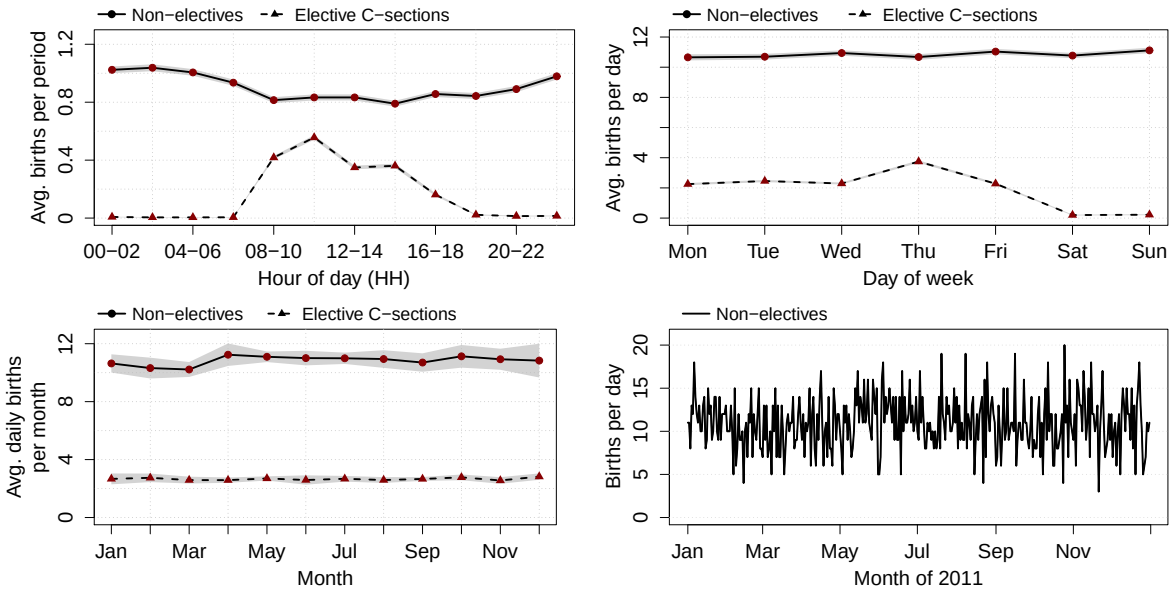
In the specific setting of this study, we need a measure of patient complexity in order to test H3 and H4. Ideally, the level of patient complexity should be readily observable by the GP; for example, it should not be a measure that only becomes available ex-post. We believe a good measure of patient complexity is the type of onset of labor, specifically whether contractions start spontaneously or are pharmacologically induced in the hospital. Women with spontaneous onset of labor tend to have less complex deliveries than induced patients as induction changes the birth process in several ways (Lothian 2006). First, following induction, contractions become stronger and more frequent more quickly and labor will last longer than after spontaneous onset. As a result, the uterine muscle cannot relax as much between contractions, causing stress on the uterus and baby. Second, induced mothers do not benefit from the natural hormonal response to spontaneous contractions, which makes labor more difficult to manage and more painful for the mother. As a consequence, induced mothers will be offered epidural analgesia more readily; in other words, epidural analgesia is less discretionary for these more complex patients. Equally importantly, the mode of labor onset is readily observable by the midwife, and inductions are sufficiently frequent to provide the requisite statistical power. In our context we expect H3 to translate into more complex patients (i.e. those that arrive with pharmacologically induced labor) being more likely to be referred for an obstetrician-led delivery as workload increases vis-à-vis less complex patients (i.e. those that arrive directly from the community after the spontaneous onset of labor). Similarly, we expect H4 to translate into less complex patients being less likely to receive epidural pain relief as workload increases vis-à-vis more complex patients.

5. Data and Variable Description

To investigate the hypotheses laid out in §4 we collaborated closely with the DU of the teaching hospital described in §3 to collect information on all births that occurred in the hospital between April 1, 2008 and March 31, 2013. For each patient we have information on (i) arrival and departure times and time stamps for any transfers between units, (ii) pregnancy-related diagnoses, classified according to the WHO's International Classification of Diseases ICD-10, and (iii) the procedures performed, classified according to the Classification of Interventions and Procedures OPSC-4.6, the UK equivalent of the American Medical Association's CPT coding system. On the staffing side, we have real-time data on the number of midwives in the DU during this period.

In total, 23,300 births occurred in the DU during our observation period, or approximately 13 births per day. In the construction of the main sample we exclude elective C-sections (3,506 births) because care in such cases is already physician-led and not materially affected by midwife decisions. We also exclude 2,672 patients who were transferred to the DU from the adjacent midwifery-led birthing unit. These patients were escalated to the DU at an advanced stage of labor specifically

Figure 1 Number of births by hour of day, day of week and month of year (mean with 95% confidence intervals) and time series of number of births per day in 2011.



because they required additional pain relief, monitoring, or a physician-led delivery. For these patients, the DU midwives do not act as a GP. In addition, to partially homogenize the sample we exclude from the main analysis any patient who is of very high risk and therefore likely to receive one-to-one care and so be shielded from any workload effect. These are identified as any patient with gestation less than 34 weeks (599 patients), any patient whose baby was born weighing less than 2,000g (129 patients), and any delivery that results in a still birth (39 patients). After removing nine observations with missing information, this leaves a final sample of 16,346 births. Importantly, all patients excluded from the analysis sample are still included in the estimation of the workload measures.

Excluding elective C-sections which occur between 8 a.m. and 6 p.m. on weekdays only, there is little within- or between-day variability in the number of deliveries observed (see Figure 1). Indeed, there is statistical evidence to suggest that the homogenous Poisson distribution (with rate of 0.45 arrivals per hour) provides a good fit for the data (and better fit than other continuous or discrete distributions).

5.1. Independent variables

To investigate how workload affects GP behavior we use individual deliveries as the unit of analysis. The main independent variable is GP workload during each delivery, which we define as the standardized time-weighted average number of patients per midwife for the period three hours

prior to birth.[4] Three hours is a somewhat arbitrary choice – it was made to coincide with the average duration of the second (and final) stage of labor. We note that averaging over different time periods (e.g. one, two or four hours) yields highly correlated workload measures and almost identical results.[5] More specifically, if $N_i(t)$ is the number of patients besides focal patient i in the DU at time t (including all patients excluded from the estimation sample, as explained above) and $MW(t)$ is the number of midwives, the (instantaneous) workload at any time t can then be expressed as

$$LOAD_i(t) = \frac{N_i(t)}{MW(t)}. \tag{1}$$

We calculate the time-weighted average load for a patient i who gives birth at time b_i using the averaging formula

$$LOAD_i = \sum_{k \in L(\underline{b}_i, b_i)} \frac{k}{b_i - \underline{b}_i} \int_{\underline{b}_i}^{b_i} \mathbb{1}[LOAD_i(t) = k]\, dt\,, \tag{2}$$

where $\underline{b}_i$ is the time three hours prior to birth, $L(\underline{b}_i, b_i)$ is the set of all observed values of $LOAD_i(t)$ between $t = \underline{b}_i$ and $t = b_i$, and $\mathbb{1}[\cdot]$ is the indicator function, taking the value one if the condition inside the brackets is satisfied and zero otherwise. Since we do not count the focal patient i in the patient counter $N_i(t)$ of (1), $LOAD_i(t)$ and $LOAD_i$ are independent of the length of time that patient i spent in the DU.

To de-trend the workload variable over the five observation years we take its z-score over a 90-day moving window. Specifically, we subtract the mean and divide by the standard deviation of the instantaneous workload, both calculated over a period from 45 days prior to 45 days after the time $t = b_i$ of birth i, giving the standardized time-weighted workload[6]

$$ZLOAD_i = \frac{LOAD_i - \mu(LOAD_i)}{\sigma(LOAD_i)}\,. \tag{3}$$

[4] Other studies measure server workload using occupancy, i.e. the number of patients present in the unit at a certain point in time or averaged over the duration of the patient visit (e.g. KC and Terwiesch 2009, Kuntz et al. 2014). One problem with this measure is that it does not accurately account for variation in staffing. Our detailed staffing data, which includes real-time information on how many midwives were present in the DU, allows us to overcome this limitation.

[5] Instead of calculating the time-weighted average load over a fixed interval (three hours in this case), it may seem more natural to do so by averaging over the total duration of the labor. We do not do this because it has the potential of introducing bias. In particular, as the time spent in labor is correlated with complications, the risk profile of a patient with a short labor will differ from that of a patient with a longer labor. Since averaging workload over a shorter time frame is more likely to generate more extreme levels of average workload, this would cause our dependent variables to be spuriously related to workload.

[6] The mean workload and the (unbiased) standard deviation $\mu(LOAD_i)$ and $\sigma(LOAD_i)$ are given by

$$\sum_{k \in L(\underline{w}_i, \overline{w}_i)} \frac{k}{\overline{w}_i - \underline{w}_i} \int_{\underline{w}_i}^{\overline{w}_i} \mathbb{1}[LOAD_i(t) = k]\, dt \;\text{ and }\; \frac{\sum_{k \in L(\underline{w}_i, \overline{w}_i)} (k - \mu(LOAD_i))^2 \int_{\underline{w}_i}^{\overline{w}_i} \mathbb{1}[LOAD_i(t) = k]\, dt}{V_1 - (V_2/V_1)}\,,$$

respectively, where $\underline{w}_i$ is the time 45 days prior to birth, $\overline{w}_i$ is the time 45 days post birth, $V_1 = \sum_{k \in L(\underline{w}_i, \overline{w}_i)} \int_{\underline{w}_i}^{\overline{w}_i} \mathbb{1}[LOAD_i(t) = k]\, dt$ and $V_2 = \sum_{k \in L(\underline{w}_i, \overline{w}_i)} \left(\int_{\underline{w}_i}^{\overline{w}_i} \mathbb{1}[LOAD_i(t) = k]\, dt \right)^2$. When we do not have activity data for the the entire 90-day time window (i.e. for patients who arrived at the start or end of our observation period) the standardization process occurs over the shorter period for which we have the required information.

Table 1 Descriptive Statistics and Correlation Table

	Descriptive statistics				Correlation table						
Variable	Mean	SD	Min	Max	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Workload	1.10	0.36	0.12	2.84	0.94***	-0.40***	-0.11***	-0.05***	-0.01	-0.00	-0.03***
(2) Standardized workload	-0.11	0.92	-3.12	3.94		-0.27***	-0.11***	-0.05***	-0.00	-0.02*	-0.03***
(3) No of midwives present	8.43	1.35	5.00	13.93			-0.00	0.02*	0.02**	0.01	0.02**
(4) High complexity	0.38	0.49	0.00	1.00				0.23***	0.03***	0.05***	0.23***
(5) Epidural analgesia	0.36	0.48	0.00	1.00					0.30***	0.14***	0.27***
(6) Physician-led delivery	0.38	0.48	0.00	1.00						0.33***	0.46***
(7) Post-birth LOS (hours)	42.95	39.26	2.90	266.98							0.76***
(8) Cost (£)	2,281.51	1,690.24	338.94	9,915.77							

The number of midwives is measured as the time-weighted average over the same time interval as workload.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The variable $ZLOAD_i$ captures the busyness of the DU relative to how busy it is in the three-month time window around the time of the birth.[7]The average workload is 1.10 – see Table 1 – suggesting that, on average, a focal mother experiences a workload of 1.10 other patients per midwife present in the unit. The ideal target is to have one midwife present per patient in active labor (NAO 2013), a target achieved for about 78% of deliveries. The histogram of standardized workload, shown in Figure 2 (left), shows that the distribution is approximately normal with a fair number of patients treated during periods of extreme workload, which aids the empirical identification of workload effects.

The second independent variable, also reported in Table 1, is a binary variable that takes the value one if the patient is of high complexity, operationalized by the need for pharmacological induction, and zero otherwise. In the final sample, 38% of deliveries are high complexity. We note that the mix of complex and non-complex patients does not exhibit any systematic variability (e.g. within-day or between-day variability).

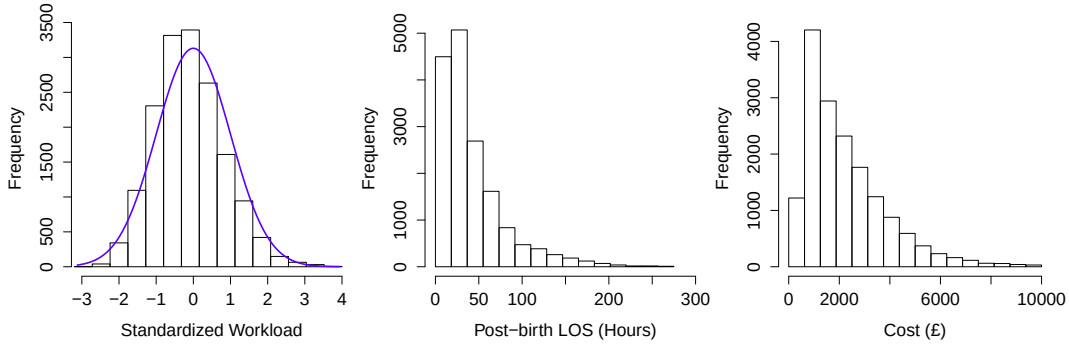
5.2. Dependent variables

The goal of this study is to understand the impact of workload on the propensity to include discretionary service activities in service plans and refer patients to specialists. Therefore, the two main dependent variables, also reported in Table 1, are (i) an indicator variable that takes the value one if the patient received epidural anesthesia and zero otherwise and (ii) an indicator variable that takes the value one if the patient was referred to a physician and zero otherwise.

Finally, to examine the implications of workload for GP behavior we collected data on a number of operational, financial and clinical metrics. We focus on: (i) post-birth LOS, measured in hours, and (ii) the cost associated with the delivery, measured in British pounds (£). Summary statistics and histograms of these measures appear in Table 1 and Figure 2, respectively. Patient LOS is

[7] All of the headline results hold up when using the unstandardized workload, $LOAD_i$, but standardization is preferred as it ensures that workload is stationary over time and so allays concerns about whether our findings are driven by time-dependent changes in both workload and unobserved characteristics of the patient population.

Figure 2 Histogram of standardized workload (left), post-birth LOS (middle), and cost (right).



often used as a proxy for resource utilization (Andritsos and Tang 2014) and quality of care (Kim et al. 2014). Cost is an important financial metric; most NHS hospitals are under pressure to reduce costs. We also collect information on three other measures, which we mention here but do not focus on for the purposes of this study: (i) a baby-related measure, the Apgar score, which is a number between zero and ten used to quickly summarize the health of babies immediately after birth; (ii) a mother-related measure, the incidence of severe (third- or fourth-degree) perineal tearing, which is a complication that occurs in some vaginal deliveries; and (iii) the length of time spent in the DU by the mother (summary statistics not reported here).

5.3. Controls

In addition to the variables of interest, we include a wide range of controls in our study. These can be broadly categorized into features relating to the mother and the pregnancy, time-related factors, medical complications during delivery, contextual factors, and operational factors. Together these account for much of the across-patient heterogeneity. A full list of controls and relevant additional information can be found in Appendix A.

6. Econometric Models and Results I: Service Configuration and Referrals

We begin our empirical investigation by seeking to identify the impact of GP workload on the rationing of discretionary service components and the rate of referrals, as per Hypotheses 1-4.

6.1. Econometric specification and instrumental variables

To examine whether the provision of discretionary services – operationalized by the provision of epidural analgesia – is affected by workload (H1), we estimate a latent variable model (probit) for the epidural decision with the standardized workload defined in (3) as the explanatory variable of interest, controlling for a wide range of factors. This model takes the form

$$EPI_i^* = \alpha_0 + \mathbf{W}_i\boldsymbol{\alpha}_1 + ZLOAD_i\alpha_2 + \delta_i \tag{4}$$

$$EPI_i = \mathbb{1}[EPI_i^* > 0], \tag{5}$$

where $\delta_i \sim \mathcal{N}(0,1)$, EPI_i^* is a latent variable, the vector $\mathbf{W}_i$ contains the set of controls (see Appendix A), EPI_i is the observed dichotomous variable indicating epidural administration, and $\mathbb{1}[\cdot]$ is the indicator function.

To examine whether workload affects the rate at which GPs refer patients to a specialist (H2) – operationalized by whether or not the delivery was physician-led – we proceed similarly. In this case, however, we include epidural analgesia as an additional control. This is done to allow for the possibility that an epidural, which if administered will always be given prior to a physician-led delivery, may increase the risk of a patient being referred to a physician (Liu and Sia 2004). Therefore, the model takes the form

$$PHYS_i^* = \beta_0 + \mathbf{W}_i\boldsymbol{\beta}_1 + ZLOAD_i\beta_2 + EPI_i\beta_3 + \epsilon_i \tag{6}$$

$$PHYS_i = \mathbb{1}[ASST_i^* > 0], \tag{7}$$

where $\epsilon_i \sim \mathcal{N}(0,1)$ and $PHYS_i^*, PHYS_i$ are the latent and observed variables, respectively, for a physician-led delivery.

To investigate whether customer complexity has a differential impact on the effect of workload on referral rates (H3) or the provision of discretionary services (H4) we also estimate the two models above separately for high- and low-complexity patients.

In addition to the fact that the workload-induced rationing of discretionary service components may have a direct effect on referral decisions (which we account for by adding the epidural variable as a control in the referral equation), a second concern is that these two decisions might be taken simultaneously rather than independently, e.g. a patient identified for referral may be given an epidural at the same time in order to reduce discomfort during the more invasive delivery. The two decisions may also be correlated due to unobserved heterogeneity, which would be the case if there are some variables – which are observable to the GP but not to us, the researchers – that make a patient more likely to experience a physician-led delivery and *also* more (or less) likely to receive epidural analgesia. Ignoring the possibility that decisions may be taken jointly would lead to simultaneity (or reverse causality) bias, while ignoring unobservables would lead to correlated omitted variable bias in the estimation of the coefficients of interest.

We correct for these problems by estimating the two models simultaneously using the recursive bivariate probit (BivProbit) model (Maddala 1983, p. 123–129). The main change is in the structure of the errors, which are assumed to be jointly distributed according to the standard bivariate normal distribution with correlation coefficient ρ. The correlation coefficient is a parameter to be estimated. More specifically, the equations to estimate remain the same (i.e. equations (5)

and (7)) but the errors become $(\delta_i, \epsilon_i) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_{\delta\epsilon}\\ \rho_{\delta\epsilon} & 1\end{pmatrix}\right)$. This model is identified and can be estimated efficiently using full information maximum likelihood (e.g. Greene 2002, p. 716). Nevertheless, it is well known that the estimation of such models is greatly improved if there is at least one exogenous variable in the first equation (selection equation) that does not appear in the second equation (outcome equation) (Wilde 2000, Maddala 1983). This condition is often referred to as the exclusion restriction and the variable in question as an instrumental variable (IV). For an IV to be valid it needs to satisfy the condition of being both relevant and exogenous. The relevance condition requires that the IV has a significant effect on the selection equation, i.e. the propensity of receiving an epidural, after we control for all observable variation. More formally, the IV needs to be correlated with the error term in the selection equation, i.e. $cor(IV, \delta) \neq 0$ in (4). The exogeneity condition requires that the IV affects the outcome equation, i.e. whether the delivery is physician-led or not, neither directly nor indirectly through correlated omitted variables. More formally, the IV needs to be uncorrelated with the error term in the outcome equation, i.e. $cor(IV, \epsilon) = 0$ in (6). We identify two such IVs.

The first IV is the time-weighted average operating theater usage by patients other than the focal patient in the period from four to two hours prior to the time of birth. Operating theater use is expected to be relevant in the epidural equation since an epidural can only be given when certain resources are available. Specifically, as discussed in §3, an epidural must be administered by anesthesiologists, who become less available when operating theaters are busy, potentially affecting the likelihood of a patient receiving an epidural. The time lag between measuring operating theater use (two to four hours before birth) and the assistance decision (which occurs near to the time of birth) makes it unlikely that it will have any direct impact on the outcome equation. To make sure that this is the case, we also control for operating theater use at the time of birth in the outcome equation to remove any potential residual effect resulting from serial correlation. Details on how this IV is calculated can be found in Appendix B.

The second IV that we include is the distance between the hospital and the patient's place of residence. This IV is specific to patients who present themselves after the spontaneous onset of labor, i.e. low-complexity patients, and is therefore only used in estimations based on this sub-sample. For these patients, the distance between the hospital and their place of residence is expected to be relevant in the selection equation (the epidural decision) since the further away the patient lives, the more likely they are to arrive at the hospital in a more advanced stage of labor, when epidural analgesia is contraindicated. Furthermore, there is no evidence or reason to suspect that distance from the hospital directly affects the likelihood of a patient receiving a physician-led delivery if required. Details on how the distance IV is calculated can be found in Appendix B.

Table 2 Descriptive Statistics and Correlation Table for the Instrumental Variables

	Descriptive statistics				Correlation table							
Variable	Mean	SD	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(8) Inst. op. tht. use	0.27	0.49	0.00	3.00	0.11^{***}	0.13^{***}	0.05^{***}	-0.02^{**}	-0.01	-0.07^{***}	-0.02^{**}	-0.03^{***}
(9) 2–4h op. tht. use	0.31	0.39	0.00	2.54	0.18^{***}	0.20^{***}	0.10^{***}	-0.03^{***}	-0.01	0.04^{***}	0.02^{*}	0.04^{***}
(10) Dist. to home	16.74	19.36	0.31	469.26	-0.00	-0.00	-0.01	0.02^{*}	-0.00	0.00	0.03^{**}	0.04^{***}

(1) Workload, (2) Std. workload, (3) Num. midwives, (4) Induced, (5) Epidural, (6) Physician-led delivery, (7) Post-birth LOS, (8) Cost
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

To ensure that this second IV does not have a direct impact on the likelihood of a physician-led delivery, we also control for the level of economic deprivation (e.g. level of income, employment, health, education, etc.) of the patient's home location using two government-produced localized indexes: one measuring general deprivation and the other, health deprivation (DCLG 2011). We do this as economic deprivation may in some way be correlated with the decision for a physician-led delivery.

In summary, we believe that the proposed IVs satisfy both of the necessary relevancy and exogeneity conditions. In addition, for all models estimated in this paper an F-test has been conducted on the null hypothesis that the instruments are weak, with the null being rejected in all cases (with p-values < 0.01). Table 2 presents summary statistics for instantaneous operating theater use (Inst. op. tht. use), operating theater use two to four hours prior to delivery (2–4h op. tht. use), and the distance in kilometers from the hospital to the patient's place of residence (Dist. to home), along with correlations with the variables presented earlier in Table 1.

The models described in this section were estimated using maximum likelihood estimation techniques (Greene 2002, p. 508-512), implemented in the `cmp` function in Stata 12.1 (Roodman 2011).

6.2. Results

Tables 3 and 4 report estimated average partial (marginal) effects (APE) with robust standard errors for the service configuration and referral decisions, respectively. Recall that the APE of a continuous variable estimates the average effect across the population of exposing all individuals to a one-unit increase in the independent variable of interest, with all other covariates held fixed. Examining Probit (1) in Table 3, we find evidence of rationing behavior by GPs at higher workload: As workload increases by one standard deviation, the rate of provision of discretionary services, i.e. epidural analgesia, decreases by 2.4% (APE $= -0.024$, p-value < 0.001). Re-estimating this model separately for low- and high-complexity patients in Probits (2) and (3), respectively, we find that workload has a strong effect on the service configuration for the low-complexity segment (APE $= -0.024$, p-value < 0.001) but does not appear to affect the high-complexity segment (APE $= -0.005$, p-value $= 0.486$). We note that one of the variables included in the models in Table 3 is operating theatre usage two to four hours before delivery. This variable partially controls for the busyness of

Table 3 Average Marginal Effects for Discretionary Service Component (Epidural)

	Probit		
Complexity	(1) Epidural *All*	(2) Epidural *Low*	(3) Epidural *High*
Std. workload	-0.024***	-0.024***	-0.005
	(0.004)	(0.005)	(0.007)
Dist. to home	-0.008†	-0.018***	0.002
	(0.004)	(0.005)	(0.007)
2–4h op. tht. use	-0.023	-0.050**	0.024
	(0.014)	(0.017)	(0.024)
Inst. op. tht. Use	-0.001	0.001	-0.010
	(0.008)	(0.009)	(0.013)
N	16346	10084	6262
Log-lik	-9616.49	-5311.48	-3833.84
Pseudo-R^2	0.098	0.098	0.117

Robust standard error in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$ $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

DU physicians and/or anesthesiologists at the time that epidural injections are most likely to be administered (specifications using different time windows give similar results).

In Table 4 we report the effect of workload on GP referral rates. Probits (1)–(3) do not include epidural analgesia as a regressor and estimate the total effect of workload on referrals, whether mediated by the effect on epidural rates or not, while Probit (4)–(6) include the epidural control and therefore estimate the residual effect of workload after accounting for the workload effect on epidural rates. Probits (1) and (4) show that in the full sample there is no apparent effect of workload on referral rates, regardless of whether we control for epidural analgesia (APE $=$ 0.001, p-value $= 0.791$) or not (APE $= -0.004$, p-value $= 0.318$). Interestingly, the referral rates in the sub-samples of low- and high-complexity patients *do* appear to be affected by workload. However, because the directions of the effects on the sub-samples are opposing they cancel out when estimating the aggregated model.

For the low-complexity segment, in Probit (2) of Table 4 we find evidence that at higher levels of workload these patients are referred at a lower rate (APE $= -0.011$, p-value $= 0.012$). However, when introducing epidural analgesia as a control in Probit (5) the workload effect becomes insignificant (APE $= -0.007$, p-value $= 0.129$). This indicates that for low-complexity patients the decrease in referrals at higher levels of workload is primarily a consequence of the rationing of epidural analgesia. Since administering an epidural increases a patient's likelihood of requiring specialist assistance (APE $= 0.202$, p-value < 0.001) the decrease in epidural rates at higher workload has the effect of reducing the referral rate for physician-led deliveries.

Since we have shown in Probit (2) of Table 3 that workload is a significant predictor in the epidural (selection) equation for low-complexity patients, we also estimate the BivProbit model to correct for any simultaneity/correlated omitted variable bias. The BivProbit model in Table

Table 4 Average Marginal Effects for Referral (Physician-led Delivery) Decision

	Probit						BivProbit	
Complexity	(1) Phys. *All*	(2) Phys. *Low*	(3) Phys. *High*	(4) Phys. *All*	(5) Phys. *Low*	(6) Phys. *High*	(1) Epidural *Low*	(2) Phys. *Low*
Std. workload	-0.004	-0.011*	0.012*	0.001	-0.007	0.012*	-0.022***	-0.001
	(0.004)	(0.004)	(0.006)	(0.003)	(0.004)	(0.006)	(0.004)	(0.005)
Epidural	–	–	–	0.192***	0.202***	0.199***	–	0.194***
	–	–	–	(0.007)	(0.010)	(0.011)	–	(0.010)
Dist. to home	-0.006	-0.009†	-0.001	-0.004	-0.005	-0.001	-0.017***	–
	(0.004)	(0.005)	(0.006)	(0.004)	(0.004)	(0.006)	(0.004)	–
2–4h op. tht. use	0.001	-0.002	0.001	0.006	0.008	-0.004	-0.042**	–
	(0.012)	(0.015)	(0.021)	(0.012)	(0.015)	(0.020)	(0.015)	–
Inst. op. tht. use	-0.063***	-0.065***	-0.056***	-0.063***	-0.064***	-0.054***	0.000	-0.063***
	(0.007)	(0.009)	(0.011)	(0.007)	(0.008)	(0.011)	(0.008)	(0.009)
N	16,346	10,084	6,262	16,346	10,084	6,262	10084	
Log-lik	-8,047.29	-4,831.77	-3,110.73	-7,665.39	-4,595.33	-2,949.52	-9,898.4	
Pseudo-R^2	0.258	0.272	0.260	0.293	0.307	0.298	–	
ρ	–	–	–	–	–	–	-0.461*** (0.090)	

Robust standard error in parentheses for Probit; *Bootstrapped standard error* in parentheses for BivProbit, 10,000 simulations; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$ $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

4 estimates a negative and highly significant correlation between the error terms of the bivariate equations ($\rho = -0.461$, p-value < 0.001), indicating that there is evidence of simultaneity and/or correlated omitted variable bias. More specifically, there exist unobservable variables that result in a patient who is less likely to receive an epidural being more likely to be referred for a physician-led delivery, and vice versa. This is to be expected: If the unobservable factors are clinical indicators of risk, then it seems reasonable that higher-risk patients would be less likely to receive an epidural as the epidural may exacerbate any existing complications (Anim-Somuah et al. 2011). Because the correlation ρ is significant, the BivProbit model is more appropriate than Probit (5) in Table 4, suggesting that after controlling for the effect of workload on epidurals and any factors that affect rationing and referral simultaneously, the residual effect of workload on referral rates for low-complexity patients effectively becomes zero (APE $= -0.001$, p-value $= 0.874$).

Turning to the high-complexity segment, Probit (3) of Table 4 indicates that at higher workload levels high-complexity patients are more likely to be referred (APE $= 0.012$, p-value $= 0.048$), and Probit (6) shows that, unlike for low-complexity patients, the effect of workload on referral rates is not driven by changes in the earlier decision whether to provide epidural analgesia (APE $= 0.012$, p-value $= 0.037$). This is not surprising, because unlike for low-complexity patients, we did not find a significant effect of workload on epidural rates for high-complexity patients (see Probit (3) of Table 3). As there is no such effect, it is not necessary to estimate a BivProbit model for high-complexity patients: The estimate of interest – the workload coefficient in Probit (6) of Table 4 – will not be biased even in the presence of simultaneity.

In summary, we find strong support for all hypotheses in §4. The effect of workload is both statistically and clinically significant. To illustrate the magnitude of the effect, we compare estimated

epidural and referral rates for a low-workload scenario, with workload two standard deviations below the sample mean (-1.95, or approx. 0.40 patients per midwife), and a high-workload scenario, with workload two standard deviations above the sample mean (1.73, or approx. 1.88 patients per midwife). For low-complexity patients the estimated epidural rate falls from 31.9% at low workload to 22.9% at high workload, a relative decrease of 28.1%. This reduction in epidural rates leads to a decrease in the physician-led delivery rate from 39.7% to 35.7%, a relative decrease of 10.1%.[8] After controlling for the changing referral rate as a consequence of the lower epidural rate, we do not find a significant additional direct effect of workload on referrals for low-complexity patients. For high-complexity patients, we find no significant effect on the epidural rate, while the estimated physician-led delivery rate increased from 37.4% to 41.9% as workload increased from the low- to the high-workload scenario, a relative increase of 11.8%.

To provide some context, comparing the effect of workload against a number of clinical factors known to influence epidural and referral decisions, we find that the size of the effect is large. The impact of workload on the epidural rate for low-complexity patients is commensurate with factors such as having given birth once before (APE$=-12.3\%$), an increase in gestation by two weeks (APE$=6.7\%$), having previously had a C-section (APE$=11.7\%$), and maternal diabetes (APE$=-9.6\%$), while the effect is about half the size of the strongest clinical factors, including breech birth (APE$=-19.2\%$). For physician-led delivery rates among high-complexity patients, the effect of workload is similar to that of maternal diabetes (APE$=6.7\%$), an increase in gestation by two weeks (APE$=3.1\%$), and maternal obesity (APE$=2.6\%$), but is smaller than for other medical conditions such as having previously had a C-section (APE$=39.3\%$) or a breech birth (APE$=36.8\%$).

7. Econometric Models and Results II: Outcomes

In this section we turn our attention to the operational and cost implications of workload-induced changes in GP behavior. More specifically, we focus on whether post-birth LOS and the overall cost of delivery are affected by workload-related changes in GP behavior.

7.1. Econometric specification and instrumental variables

We start our investigation by constructing two linear regression models, one for post-birth LOS ($PbLOS_i$) and another for costs ($COST_i$). The histograms in Figure 5.3 suggest models based on logarithmic transformations of the dependent variables.

$$\ln(PbLOS_i) = \gamma_0 + \mathbf{X}_i\boldsymbol{\gamma}_1 + ZLOAD_i\gamma_2 + EPI_i\gamma_3 + ASST_i\gamma_4 + \mu_i \tag{8}$$

[8] We note that these two effects work in opposite directions: To reduce her workload the GP rations the provision of epidural analgesia but, as a result, she becomes less able to refer these patients to the physician. The magnitude of the first effect, though, is much larger than the second, which would suggest that by rationing epidurals the GP reduces her overall workload.

$$\begin{aligned} \mu_i &\sim \mathcal{N}(0, \sigma_\mu^2)\,. \\ \ln(COST_i) &= \lambda_0 + \mathbf{X}_i\boldsymbol{\lambda}_1 + ZLOAD_i\lambda_2 + EPI_i\lambda_3 + ASST_i\lambda_4 + \nu_i \\ \nu_i &\sim \mathcal{N}(0, \sigma_\nu^2)\,. \end{aligned} \tag{9}$$

The control vector $\mathbf{X}_i$ of (8) and (9) is similar to $\mathbf{W}_i$ that is used in modeling rationing and referral behavior in equations (4) and (6) (see Appendix A) but with one addition: Here, we also control for time-weighted occupancy in the post-natal unit in the six-hour period prior to the discharge of the focal patient. This additional control is included so that we can isolate the effect of GP workload in the DU on a patient's post-birth LOS and delivery cost rather than erroneously capturing post-delivery discharge pressure due to the effect of DU workload on occupancy in the post-natal unit. Since recent studies have found non-linear workload effects on discharge (e.g. Kuntz et al. 2014), we also include the square of time-weighted post-natal unit occupancy in $\mathbf{X}_i$.

As in the referral estimation in §6, endogeneity is potentially a problem when estimating the workload effect in the outcomes models. For example, there might be an unobservable variable that makes a patient more likely to both receive discretionary services (in this case, an epidural) and have longer post-birth LOS and/or higher costs. To account for this we supplement the models in (8) and (9) with the Heckman treatment effects (HeckTreat) model (Maddala 1983, p. 123–129). The HeckTreat model is similar to the BivProbit model but with a few differences, which we explain below.[9] Before we do so, we find it necessary to distinguish between the low- and high-complexity patient segments.

We have shown that for low-complexity patients workload has a direct effect on the service configuration decision (epidural analgesia) (4) but no direct effect on referral propensity (see §6.2). Therefore, we need only be concerned with the potential for endogeneity between the epidural decision and the outcomes – without correcting for it the coefficient of the epidural variable in the outcome equations may be biased.[10] This suggests that the appropriate endogeneity-corrected model for the low-complexity patient segment is a HeckTreat model where the first-stage (selection) equation is the decision to administer an epidural (i.e. equation (4)) and the second-stage equation for post-birth LOS is as in (8) and with error structure $(\delta_i, \mu_i) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_{\delta\mu}\sigma_\mu \\ \rho_{\delta\mu}\sigma_\mu & \sigma_\mu^2\end{pmatrix}\right)$. A similar formulation would hold for costs. Estimation is made using full information maximum likelihood (Fraser and Guo 2009). Furthermore, to improve the robustness of the estimation we need to identify at least one appropriate IV. For the IV to be valid it must affect the rate of

[9] Note that HeckTreat differs from the standard Heckman sample selection model in two ways: First, the dummy variable indicating treatment appears in the outcome equation; second, the outcome variable is observed regardless of whether the individual receives the treatment (Fraser and Guo 2009).

[10] Refer to §6.1 for a discussion of why this is the case.

epidural analgesia but have no effect on outcomes (post-birth LOS or costs). The distance between the hospital and the patient's home does not satisfy the exogeneity condition: It is possible that a patient who lives further from the hospital is more likely to be delayed in being discharged, since if a problem subsequently arises it will take longer for the patient to return to the hospital. This means that post-birth LOS may be higher for these patients and, as a result, costs may also increase. For the low-complexity segment we use operating theater utilization (by mothers other than the focal mother) two to four hours prior to birth, since there is no reason to believe that this would have an impact on post-birth LOS or costs other than through the epidural decision.

We have shown that for the high-complexity patient segment workload has no effect on the service configuration decision but does increase the rate of referrals. As a result, in the endogeneity-corrected models the referral decision in (6) is the first-stage (selection) equation, while the second stage would be post-birth LOS, as in (8) (or cost as in (9)). The error structure is equivalent to that of the low-complexity patient segment. A valid IV in this case should affect the referral rate but not have a direct impact on either of the two outcomes. In §6.1 we have already shown that operating theater utilization two to four hours prior to birth and the patient's distance from the hospital do not satisfy the relevancy condition (i.e. they do not affect the referral rate directly). Instead, for the referral decision we propose using operating theater utilization at the time of birth as the IV. It is clear from Probit (6) of Table 4 that this variable will satisfy the relevancy condition: If the operating theater is busy with other patients when the focal patient gives birth, then the focal patient will be significantly less likely to receive a referral for a physician-led delivery (APE $= -0.054$, p-value < 0.001). In addition, the busyness of the operating theater at the time of birth by mothers other than the focal mother should have no impact on post-birth LOS or costs other than through the referral rate. Therefore, this IV is expected to be both relevant and exogenous.

7.2. Results

Tables 5 and 6 report the estimated coefficients, standard errors, and model summary statistics of the outcome-related regressions. As in §6.2, in discussing effect sizes we use workload two standard deviations below (above) the mean to denote the low- (high-)workload scenario.

7.2.1. Post-birth LOS The first outcome we examine is post-birth LOS. The two types of model we examine are OLS and HeckTreat. The former ignores endogeneity while the latter accounts for it. Model coefficients are presented in Table 5.

For low-complexity patients there is evidence in OLS (1) that an increase in workload leads to a decrease in post-birth LOS (coef. $= -0.019$, p-value $= 0.031$). In order to determine the extent to which this is a direct workload effect rather than a consequence of the earlier service configuration

Table 5 Coefficient Estimates in Statistical Models for Post-birth LOS

	OLS				HeckTreat			
Complexity	(1) PbLOS *Low*	(2) PbLOS *High*	(3) PbLOS *Low*	(4) PbLOS *High*	(1) Epidural *Low*	(2) PbLOS *Low*	(3) Phys. *High*	(4) PbLOS *High*
Std. workload	-0.019*	0.004	-0.010	-0.003	-0.079***	0.008	0.067**	0.009
	(0.009)	(0.011)	(0.009)	(0.010)	(0.017)	(0.009)	(0.022)	(0.011)
Epidural	–	0.315***	0.343***	0.225***	–	1.092***	0.702***	0.343***
	–	(0.019)	(0.016)	(0.018)	–	(0.049)	(0.039)	(0.025)
Phys. delivery	–	–	–	0.432***	–	–	–	-0.142†
	–	–	–	(0.020)	–	–	–	(0.079)
2–4h op. tht. use	-0.045	-0.045	-0.028	-0.042	-0.156**	–	–	–
	(0.029)	(0.036)	(0.029)	(0.034)	(0.052)	–	–	–
Inst. op. tht. use	-0.042**	-0.003	-0.041**	0.019	0.003	-0.040*	-0.176***	–
	(0.016)	(0.020)	(0.016)	(0.019)	(0.029)	(0.017)	(0.039)	–
N	10,084	6,262	10,084	6,262	10,084		6,262	
Log-lik	-11,273.05	-6,570.48	-11,078.63	-6,370.87	-16,362.43		-9,317.79	
Adj-R^2	0.312	0.279	0.338	0.323	–		–	
ρ	–	–	–	–	-0.561*** (0.029)		0.485*** (0.057)	

Robust standard error in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

decision we must examine the models that control for epidural analgesia, given in OLS (3) and HeckTreat (1–2). The negative and significant value of ρ (coef. $= -0.561$, p-value < 0.001) in the HeckTreat model is evidence of non-random selection, with the negative sign indicating that post-birth LOS is higher for patients who are not selected to receive an epidural. Based on our earlier discussion, this is reasonable: Since an epidural increases the chance of adverse outcomes, it is less likely to be given to high-risk patients, who also have higher expected post-birth LOS. Therefore, the appropriate model for reporting our results for low-complexity patients is an endogeneity-corrected HeckTreat model.

The workload coefficient in HeckTreat (2) captures the residual effect of increasing DU workload on post-birth LOS that is not explained by the effect of workload on epidural rates; this residual effect is not significant at conventional levels (coef. $= 0.008$, p-value $= 0.394$). Therefore, the model suggests that the aggregate decrease in post-birth LOS reported in OLS (1) is due entirely to the indirect effect of rationing epidural analgesia at higher levels of workload. The estimated average treatment effect (ATE) of epidural analgesia on post-birth LOS is an increase of 31.0%.[11] As a result, the indirect effect (through the reduction in epidural rates) of moving from low- to high-workload conditions is estimated to be an 8.5% decrease in post-birth LOS. This indirect effect is calculated using the full marginal effect of the HeckTreat model (details available from the authors).

For high-complexity patients there is little evidence that increased workload leads to an increase in post-birth LOS. The HeckTreat (3–4) model indicates the presence of positive selection, as

[11] Observe that the ATE is not equal to the coefficient estimate for epidural analgesia in the outcome equation. This is because the conditional expectations of the outcome include terms that account for selection bias. See e.g. Brown and Mergoupis (2011) for the derivation of the ATE in a HeckTreat model.

Table 6 Coefficient Estimates in Statistical Models for Cost

	OLS				HeckTreat			
Complexity	(1) Cost *Low*	(2) Cost *High*	(3) Cost *Low*	(4) Cost *High*	(1) Epidural *Low*	(2) Cost *Low*	(3) Phys. *High*	(4) Cost *High*
Std. workload	-0.015*	0.014†	-0.007	0.007	-0.075***	0.014†	0.068**	0.013†
	(0.007)	(0.008)	(0.006)	(0.007)	(0.016)	(0.008)	(0.022)	(0.007)
Epidural	–	0.326***	0.353***	0.240***	–	1.172***	0.693***	0.269***
	–	(0.014)	(0.012)	(0.013)	–	(0.035)	(0.039)	(0.019)
Phys. delivery	–	–	–	0.415***	–	–	–	0.267***
	–	–	–	(0.015)	–	–	–	(0.068)
2–4h op. tht. use	-0.010	-0.031	0.008	-0.028	-0.102*	–	–	–
	(0.022)	(0.026)	(0.022)	(0.025)	(0.045)	–	–	–
Inst. op. tht. use	-0.047***	-0.016	-0.047***	0.005	-0.005	-0.045**	-0.191***	–
	(0.012)	(0.014)	(0.011)	(0.014)	(0.028)	(0.014)	(0.042)	–
N	10,084	6,262	10,084	6,262	10,084		6,262	
Log-lik	-8,513.24	-4,568.92	-8,150.39	-4,211.41	-13,377.13		-7,170.72	
Adj-R^2	0.393	0.365	0.435	0.433	–		–	
ρ	–	–	–	–	-0.755*** (0.021)		0.186* (0.082)	

Robust standard error in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

indicated by the positive and significant value of ρ (coef. $= 0.485$, p-value < 0.001), and is therefore more appropriate that OLS (2) or OLS (4). After controlling for endogeneity, the direct effect of workload on post-birth LOS is not significant (coef. $= 0.009$, p-value $= 0.387$), while the effect of a physician-led delivery is negative but only weakly significant (coef. $= -0.142$, p-value $= 0.072$). These estimates tentatively suggest a potential reduction in post-birth LOS at high workloads for high-complexity patients as a consequence of an increased referral rate for these patients.

7.2.2. Cost The second outcome we examine is cost. As with post-birth LOS, the two types of model we examine are OLS and HeckTreat. Model coefficients are presented in Table 6.

The unmediated model for low-complexity patients, given in OLS (1), suggests an overall decrease in costs with increasing workload (coef. $= -0.015$, p-value $= 0.022$). Model OLS (3) suggests that this decrease is explained by the reduction in epidural rates as the direct effect of workload becomes insignificant after controlling for epidural analgesia. The model HeckTreat (1–2) accounts for endogeneity in the epidural decision and estimates a significantly negative correlation ρ (coef. $= -0.755$, p-value < 0.001). This suggests that patients who are selected to receive an epidural are likely to have unobserved characteristics that make their deliveries less costly. Given the highly significant correlation ρ, HeckTreat (1–2) is the most appropriate model for the interpretation of cost effects. The model provides weak evidence of a positive direct effect of DU workload on delivery costs (coef. $= 0.014$, p-value $= 0.062$), after controlling for the effect of workload on epidurals and the endogeneity of the epidural decision. There is, however, strong evidence of an indirect cost effect of workload via the rationing response on epidurals, as epidurals have a strong positive effect on costs after controlling for endogeneity (coef. $= 1.172$, p-value < 0.001), and workload has a strongly significant effect on a mother's likelihood of receiving an epidural (coef. $= -0.075$, p-value < 0.001).

For low-complexity patients the estimated direct effect of moving from a low- to high-workload environment is an increase in costs of 5.3% per patient. At the mean value (£2,281.51) this is equivalent to a £130.00 increase in costs for each low-complexity patient, or an increase in total costs of £1,220,100 across the full low-complexity sub-sample. This is, however, counteracted by an 8.7% (£198.87 at the mean) indirect decrease in costs caused by the workload-induced rationing of epidurals. In total, moving from low- to high-workload conditions results in a decrease in costs of 3.4%, which is equal to a saving of £77.89 per patient or a total saving of £785,300 across the low-complexity sample.

For high-complexity patients there is weak evidence, given in OLS (2), that an increase in workload leads to an increase in costs (coef. = 0.014, p-value = 0.082). Following the same intuition as for post-birth LOS, the positive value of ρ (coef. = 0.186, p-value 0.026) in HeckTreat (3–4) suggests that there is a positive selection effect. Weak evidence of a positive direct effect of workload on costs remains even after controlling for endogeneity (coef. = 0.013, p-value = 0.090). Translated into cost terms, the direct effect at the mean (£2,281.51) of moving from low- to high-workload conditions is an increase in costs of £105.40 (4.6%) per patient, or £660,000 in total across the high-complexity sub-sample. Added to this is an increase of £55.65 (2.4%) per patient resulting from the increased rate of physician-led deliveries, leading to a total cost increase per patient of £161.05, or £1,008,500 when aggregated across all high-complexity patients in the sample.

7.3. Other measures

In addition to post-birth LOS and costs, we investigate whether workload-induced changes in GP behavior affect a baby- and a mother-related health measure. Our main findings (not reported here) are that workload either has no effect (e.g. we observe no reduction in baby Apgar scores), or has an effect that is in the direction predicted by extant literature but unrelated to GP behavior (e.g. we find that perineal tears are more likely at higher workload for mothers who give birth vaginally but that this is independent of GP decisions). We also examine another operational measure, the DU LOS. This is a potentially important measure because it is related to patient throughput. As was the case for post-birth LOS and costs, the impact of GP behavior is confounded by omitted variable bias, i.e. there are unobserved factors that affect both GP decisions and DU LOS. However, unlike for post-birth LOS and costs, this is more difficult to resolve with instrumental variables as all of the variables that have been used as instruments in the previous sections are likely to affect DU LOS directly. Nevertheless, we believe that the workload-induced change in GP behavior is consistent with increasing throughput: Epidural analgesia is known to increase the length of labor (Anim-Somuah et al. 2011) and a physician-led delivery by definition brings forward the delivery by terminating labor before it finishes naturally. Therefore, although the reduction in the provision

of epidural analgesia and increase in the rate of referrals for physician-led deliveries observed at higher workload is done primarily to balance time-sharing across multiple mothers, these responses are consistent with behavior that reduces the average time patients spend in the DU.

7.4. Robustness checks

To confirm the robustness of the results presented in the previous sections we estimate a number of alternative model specifications that: (i) control for midwife fixed effects, Elixhauser et al. (1998) comorbidities, and antenatal unit occupancy; (ii) measure workload over different time windows and post-birth LOS in nights rather than a continuous scale; and (iii) allow for non-linear workload effects. The results are qualitatively similar to those reported in the paper.

8. Discussion

The GP literature is based on the assumption that workload varies exogenously and that GPs' service configuration and referral decisions are unaffected by workload. Our data and analysis suggest that this assumption is invalid: Workload affects both decisions. In this section we discuss whether this has a material impact on one of the most important decisions in such a service setting: decisions on GP staffing levels. Management orthodoxy suggests that increasing staffing will reduce customer waiting times and/or the need for parallel processing, therefore improving service quality, albeit with additional staffing costs. However, if workload affects decisions, then there may be other, more surprising implications; see for example Hopp et al. (2007), Green et al. (2013), and Tan and Netessine (2014), who show that the endogenous response to workload means that an increase in staffing may reduce (i) throughput in a service setting with discretionary service components, (ii) absenteeism of ED nurses, and (iii) restaurant sales, respectively. Does the influence of workload on GP decisions create such counterintuitive comparative statics with regard to staffing?

To investigate this question in the context of our study unit, we evaluate the implications of an increase in midwife staffing to a level that raises that proportion of mothers receiving the desirable one-to-one (or better) level of care during active labor from the current level of 78% to 88% and 95%. The number of additional staff required to achieve this was recently investigated by the National Audit Office (NAO 2013) and Green and Liu (2013), without accounting for the influence of workload on GP decisions. Following NAO (2013), we make the simplifying assumption that staffing levels can be fixed and, therefore, that any variation in workload is caused by fluctuating demand only. Under this assumption, and using the demand variation present in our data, the DU would require eight midwives to achieve the current service level of 78% one-to-one care and would have to increase this to nine or 11 midwives to achieve the 88% or 95% levels, respectively. Using current estimates for the full economic cost of a midwife (Curtis 2012), this would add approximately £184,900 and £831,000, respectively, to the staffing bill per annum. As shown in

Table 7 Relative Change in Gatekeeping Behavior and Expected Outcomes under Alternative Staffing Scenarios.

	88% One-to-one service (Fixed staffing)			95% One-to-one service (Fixed staffing)			95% One-to-one service (Variable staffing)		
Complexity	*All*	*Low*	*High*	*All*	*Low*	*High*	*All*	*Low*	*High*
Avg. workload	0.960	0.985	0.919	0.785	0.806	0.752	0.883	0.899	0.856
Avg. MW per obs.	9	9	9	11	11	11	9.66	9.73	9.53
Epidural analgesia									
– 10^{th}%ile	5.4%	6.5%	0.8%	12.3%	14.6%	1.7%	9.1%	10.2%	1.5%
– Mean	1.7%	3.3%	0.3%	4.0%	7.8%	0.8%	2.8%	5.5%	0.5%
– 90^{th}%ile	0.5%	2.4%	0.2%	1.2%	5.7%	0.5%	0.8%	4.1%	0.3%
Phys. deliveries									
– 10^{th}%ile	0.1%	2.9%	-2.6%	1.9%	7.0%	-6.6%	1.2%	4.7%	-4.6%
– Mean	0.3%	1.1%	-1.0%	0.6%	2.6%	-2.4%	0.5%	1.9%	-1.5%
– 90^{th}%ile	0.0%	0.5%	-0.7%	-0.0%	1.2%	-1.4%	-0.0%	0.9%	-0.9%
PbLOS (hours)									
– 10^{th}%ile	0.6%	0.7%	-0.2%	1.1%	1.5%	-0.3%	0.8%	1.2%	-0.3%
– Mean	0.4%	0.7%	-0.1%	0.8%	1.6%	-0.3%	0.6%	1.2%	-0.2%
– 90^{th}%ile	0.4%	0.8%	-0.1%	1.1%	1.6%	-0.2%	0.7%	1.2%	-0.1%
Cost (£)									
– 10^{th}%ile	0.5%	0.5%	-0.5%	0.8%	1.3%	-1.1%	0.6%	0.8%	-0.6%
– Mean	0.1%	0.6%	-0.4%	0.2%	1.3%	-1.1%	0.2%	0.9%	-0.7%
– 90^{th}%ile	-0.1%	0.4%	-0.7%	-0.0%	1.2%	-1.0%	-0.1%	0.7%	-0.6%

Scenario analysis uses standardized workload excluding the focal patient and with standardization performed using the same mean and s.d. as in (3).

Table 7, increasing staffing levels to nine midwives (Columns 2–4, 88% One-to-one service) is associated with a relative increase in the epidural rate across all patients (Column 2) of 1.7% and physician-led deliveries, of 0.3% (at the mean workload and compared with the observed workload in the data). The use of discretionary services and specialists increases further when staffing levels rise to 11 midwives (Columns 5–7, 95% One-to-one service), to increases of 4.0% in epidural analgesia rates and 0.6% in referral rates. Contrary to conventional assumptions, the increase in staffing leads to a *deterioration* in average service outcomes, as indicated by an increase in post-birth LOS by 0.4% in the nine-midwives scenario and by 0.8% in the 11-midwives scenario. Furthermore, we find that the increase in staffing raises costs by 0.1% and 0.2%, respectively, in the two staffing scenarios (over and above the increase in staffing costs) due to GP response to workload. While the aggregate effects across all patients are relatively small, they become more pronounced for the sub-samples of low- and high-complexity patients (Columns 3 and 6 for low-complexity patients and Columns 4 and 7 for high-complexity patients). In particular, in the 11-midwives scenario (Columns 6 and 7), physician-led delivery rates and costs increase markedly for low-complexity patients, by 2.6% and 1.3%, respectively, while referral rates and costs *decrease* for high-complexity patients by 2.4% and 1.1%, respectively. This shows that the overall effect of changes in GP staffing, in magnitude as well as in sign, depends on the case mix.

GPs play an important role in assigning patients to the most appropriate treatment route and, thereby, keeping costs under control. The behavioral effects of workload suggest that too much work for GPs results in a tendency to increase referrals to specialists, while too little work may

result in a tendency to provide more discretionary service features. Our findings are therefore consistent with overtreatment at both high and low workload. However, since we cannot know the appropriate level of treatment for any specific mother, it is not possible to quantify overtreatment or provide more specific evidence for the relationship between GP workload and overtreatment. This could be a fruitful avenue for further research. The observation, however, that compared to average workload GPs tend to overtreat at both high and low levels of workload does suggest that operational interventions that aim to better match GP supply with demand have the potential to reduce overtreatment. Motivated by this observation, we also examine a perfectly flexible staffing regime in which the unit adds staff above their existing levels only when workload exceeds the desired one-to-one level of care up to a maximum of 11 staff members. This purely hypothetical way of staffing has the potential to increase one-to-one care levels from 78% to 95% without substantially increasing the average number of staff required (a 14.6% – or £328,000p.a. – increase as compared to 30.5% – or £831,000p.a. – increase under the equivalent fixed staffing policy), while partially negating the workload-related changes in discretionary service and referral rates by GPs (20–30% lower than the equivalent fixed staffing policy).

Our findings may also have implications for service specialization in the context of services that include GPs. We show that customer complexity moderates the effect of workload: Non-complex customers experience cuts in discretionary services, while complex customers are referred to specialists more often. If operational changes such as unit specialization change the case mix, for example, by diverting non-complex patients to midwifery-led birthing units and thereby increasing the complexity of the residual patients in the standard DU, then the referral effect for complex patients may become even more pronounced because GPs have fewer patients with which they can apply the second lever – the rationing of discretionary services – to regulate their workload. It may be possible to analyze this by interacting the workload variable with a variable that captures the percentage of non-complex patients in the unit during the period prior to birth. Our sample, however, does not have enough statistical power to enable such an analysis, and we must leave this question for further research.

Appendix

A. Controls

In Table 8 we list all of the exogenous regressors (controls) for the models presented in Tables 3–6. These can broadly be broken down into six categories: factors related to the mother, those specific to the pregnancy, time controls, a subset of the clinical conditions that may affect outcomes (chosen from the relevant medical literature), contextual controls, and organizational factors that were not the focus of this paper. The number following the variables specified as categorical indicates the number of categories. We indicate the models in which the controls were included by either the direction of their effect, as indicated by the sign and

Table 8 Table of Controls.

	Type	Epidural	Physician-led delivery	Post-birth LOS	Cost
Maternal Characteristics					
- Age	Categorical (4)	Y	Y	Y	Y
- Body mass index	Categorical (3)	Y	Y	Y	Y
- Num. prev. births	Categorical (4)	Y	Y	Y	Y
- Previous C-section	Binary	+	+	+	+
Pregnancy Characteristics					
- Gestation	Categorical (7)	Y	Y	Y	Y
- Baby weight	Continuous	+	+	−	−
- Baby weight sq.	Continuous	0	+	+	+
Temporal					
- Year-qtr	Categorical (20)	Y	Y	Y	Y
- Hour of birth (2-hourly)	Categorical (12)	Y	Y	N	Y
- Weekend	Binary	0	0	0	0
Clinical Complications					
- Breech	Binary	−	+	+	0
- Malpresentation	Binary	+	+	+	+
- Shoulder dystocia	Binary	0	0	+	0
- Obstructed labor	Binary	0	+	+	+
- Diabetes	Binary	+	+	+	+
- Hypertension	Binary	+	+	+	+
- PROM	Binary	+	+	+	+
- COPD	Binary	+	+	+	+
- Other complications	Binary	+	+	+	+
Contextual Factors					
- Deprivation index	Continuous	0	0	0	0
- Health deprivation index	Continuous	−	0	+	+
- Unkn. dist. to hospital	Binary	0	0	0	0
- Antenatal stay	Binary	+	+	+	+
- Num. antenatal visits	Categorical (4)	Y	N	Y	Y
Other Operational Factors					
- Proportion epidural	Continuous	0	0	0	0
- Proportion physician led	Continuous	0	0	0	0
- Proportion escalated	Continuous	0	0	0	0
- Post-birth workload	Continuous	N/A	N/A	0	+
- Post-birth workload sq.	Continuous	N/A	N/A	0	0

PROM: indicates that a patient experienced premature rupture of membranes. *COPD*: indicates that a patient had chronic obstructive pulmonary disease. *Deprivation index*: an index of multiple deprivation. *Health deprivation index*: an index of health deprivation. *Unkn. dist. to hospital*: indicates that it was not possible to identify the distance between the patient's home and the hospital. *Proportion epidural*: the time-weighted proportion of other patients in the DU that received an epidural in the three-hour period prior to the time of birth for the focal patient. *Proportion physician led*: as above, but the proportion who experienced a physician-led delivery. *Proportion escalated*: as above, but the proportion of patients escalated from the midwifery-led unit.

significance of the estimated coefficient (+ for positive and significant, − for negative and significant, 0 for insignificant, all at the 5% level), and for categorical variables by Y if one or more of the levels was significant at the 5% level and N otherwise.

A useful exercise is to check that the direction of the reported effects in the models corresponds with intuition and with medical literature (e.g. Bragg et al. 2010, Renfrew et al. 1998, Eason et al. 2000). For example, larger babies are associated with increased likelihood that the mother will require pain relief and need physician assistance during delivery; therefore, we should expect a positive coefficient for the "Baby weight" variable in the "Epidural" and "Physician-led delivery" equations, which we do find. Clinical complications in general have been shown to lead to poorer outcomes (in terms of increased need for physician-led deliveries, increased LOS, and higher costs), which, again, we see.

B. Calculation of instrumental variables

We use two instrumental variables: (i) operating theater usage by patients other than the focal patient in the period two to fours hours prior to the time of birth and (ii) the distance between the hospital and the patient's place of residence. The exact calculation of the first variable is as follows. For deliveries after

April 1, 2009 (78.6% of patients), actual obstetric operating theater data is used in this calculation. For the remaining patients, we do not have operating theater data as it was not collected. Instead we infer operating theater time stamps by assuming that an operating theater is in use in the 45 minutes prior to and 45 minutes after the birth of a baby delivered in theater. These are the mean (and also median) times in the observed data. To formalize the calculation of this IV, define b_i to be the time that patient i gives birth and P_{OT} to be the set of patients who delivered in an obstetric operating theater. If operating theater data is available, for each patient $j \in P_{OT}$ let $\underline{b}_j$ be the time the operation begins and $\overline{b}_j$ be the time the operation ends. If operating theater time stamps are not available, for patient $j \in P_{OT}$ set $\underline{b}_j$ and $\overline{b}_j$ to be the times 45 minutes prior to and post b_j, respectively. At any time t the operating theater use by patients other than the focal patient i will be equal to $OT_i(t) = \sum_{j \in P_{OT} \setminus \{i\}} \mathbb{1}\left[t \in [\underline{b}_j, \overline{b}_j]\right]$. Then, the (instantaneous) operating theater use at time of birth for patient i is given by $OT_i^{INS} = OT_i(b_i)$. Therefore, the IV is given by $OT_i^{PRI} = \sum_{k \in L_i(r_i, s_i)} \frac{k}{s_i - r_i} \int_{r_i}^{s_i} \mathbb{1}[OT_i(t) = k]\, dt$, where r_i and s_i are the times four and two hours prior to birth, respectively, and $L_i(r_i, s_i)$ is the set of all observed values of $OT_i(t)$ between $t = r_i$ and $t = s_i$. As a robustness check, for the 78.6% of the patients for whom we can generate OT_i^{PRI} using both actual and inferred operating theater time stamps we find that the two are strongly positively correlated ($\rho = 0.81$, p-value < 0.001), as desired.

The exact calculation of the second IV, the distance between the hospital and the mother's place of residence, proceeds as follows. For 68.7% of patients we know the residential postcode (which is a very localized measure in the UK), and using this information we can calculate the distance from the residence to the hospital. For the remaining patients, the residential postcode is not known. However, for the majority of these patients we know the address of the primary care practice (PCP) and can therefore use the distance between the hospital and the patient's PCP as a proxy for the distance from home. For patients where we can observe both the place of residence and the PCP, 34%, 51%, 71%, and 83% live within 1km, 2km, 5km, and 10km of the PCP, respectively, indicating that the location of the PCP is generally a good proxy for the place of residence. After this, there remains 1.0% of patients for whom we have no location information. For these, we set the distance equal to the average of all other patients, introduce a dummy to capture any unobserved differences, and include this dummy in both the selection and outcome equations. Finally, to reduce the skewness of the distribution of distance observed in the data, we take its natural logarithm.

References

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

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

Andritsos, D.A., C.S. Tang. 2014. Linking process quality and resource usage: An empirical analysis. *Production and Operations Management* **23**(12) 2163–2177.

Anim-Somuah, M., R. Smyth, C. Howell. 2011. Epidural versus non-epidural or no analgesia in labour. *Cochrane Database of Systematic Reviews* **4**.

Batt, R., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. The Wharton School, Working paper.

Bendoly, E., K. Donohue, K. L. Schultz. 2006. Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management* **24**(6) 737–752.

Berry-Jaeker, J., A. Tucker. 2013. An empirical study of the spillover effects of workload on patient length of stay. HBS Working Paper.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

Boudreau, J., W. Hopp, J.O. McClain, L. J. Thomas. 2003. On the interface between operations and human resources management. *Manufacturing & Service Operations Management* **5**(3) 179–202.

Bragg, F., D. Cromwell, L. Edozien, I. Gurol-Urganci, T. Mahmood, A. Templeton, J. van der Meulen. 2010. Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: Cross sectional study. *BMJ* **341**.

Brekke, K. R., R. Nuscheler, O. R. Straume. 2007. Gatekeeping in health care. *Journal of Health Economics* **26**(1) 149–170.

Brown, G. K., T. Mergoupis. 2011. Treatment interactions with nonexperimental data in Stata. *Stata Journal* **11**(4) 545–555.

Chan, C. W., V. F. Farias, G. Escobar. 2015. The impact of delays on service times in the intensive care unit. *Columbia GSB Working Paper* .

Clover, B. 2010. Midwife workloads too high to be safe. *Nursing Times* Published: 2010-09-23. Accessed: 2014-03-11.

Curtis, L. 2012. Unit costs of health & social care 2012. Tech. rep., Personal Social Services Research Unit.

DCLG. 2011. The english indices of deprivation. Tech. rep., Department for Communities and Local Government.

Debo, L.G., L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.

DH. 2012. Patient level information and costing systems (PLICS) and reference costs best practice guidance for 2011-12. Tech. rep., Department of Health.

Eason, E., M. Labrecque, G. Wells, P. Feldman. 2000. Preventing perineal trauma during childbirth: a systematic review. *Obstetrics & Gynecology* **95**(3) 464–471.

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

Fraser, M., S. Guo. 2009. Sample selection and related models. *Propensity Score Analysis*. Sage Publications, Inc., 85–125.

González, P. 2010. Gatekeeping versus direct-access when patient information matters. *Health economics* **19**(6) 730–754.

Green, L., N. Liu. 2013. Capacity planning for hospital obstetrics units: Insights from a New York City study. Columbia University Working Paper.

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

Greene, W. 2002. *Econometric Analysis*. 5th ed. Prentice Hall.

Hamilton, B. E., P. D. Sutton. 2013. Recent trends in births and fertility rates through December 2012. Tech. rep., National Center for Health Statistics.

Hasija, S., E. J. Pinker, R. A. Shumsky. 2005. Staffing and routing in a two-tier call centre. *International Journal of Operational Research* **1**(1/2) 8–29.

Henderson, J., R. McCandlish, L. Kumiega, S. Petrou. 2001. Systematic review of economic aspects of alternative modes of delivery. *British Journal of Obstetrics and Gynaecology* **108**(2) 149–157.

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

Huckman, R.S., B.R. Staats, D.M. Upton. 2009. Team familiarity, role experience, and performance: Evidence from Indian software services. *Management science* **55**(1) 85–100.

Kane, R., T. Shamliyan, C. Mueller, S. Duval, T. Wilt. 2007. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care* **45**(12) 1195–1204.

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

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

Kesavan, S., B. R. Staats, W. Gilland. 2014. Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science* **60**(8) 1884–1906.

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

Kostami, V., S. Rajagopalan. 2013. Speed-quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management* **16**(1) 104–118.

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

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

Liu, E., A. Sia. 2004. Rates of caesarean section and instrumental vaginal delivery in nulliparous women after low concentration epidural infusions or opioid analgesia: Systematic review. *BMJ* **328** 1410–1415.

Luo, J., J. Zhang. 2013. Staffing and control of instant messaging contact centers. *Operations Research* **61**(2) 328–343.

Maddala, G. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

Malcomson, J. M. 2004. Health service gatekeepers. *The RAND Journal of Economics* **35**(2) 401–421.

Mariñoso, B. G., I. Jelovac. 2003. GPs' payment contracts and their referral practice. *Journal of Health Economics* **22**(4) 617–635.

Martin, A. B., M. Hartman, L. Whittle, A. Catlin. 2014. National health spending in 2012: Rate of health spending growth remained low for the fourth consecutive year. *Health Affairs* **33**(1) 67–77.

NAO. 2013. Maternity services in England. Tech. rep., National Audit Office.

Needleman, J., P. Buerhaus, V. Pankratz. 2011. Nurse staffing and inpatient hospital mortality. *N. Engl. J. Med.* **364**(11) 1037–1045.

OAA. 2013. Guidelines for obstetric anaesthetic services. Tech. rep., The Obstetric Anaesthetists' Association and the Association of Anaesthetists of Great Britain & Ireland. URL `http://www.aagbi.org/sites/default/files/obstetric_anaesthetic_services_2013.pdf`.

Oliva, R., J. Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7) 894–914.

Paç, M. F., S. Veeraraghavan. 2015. False diagnosis and overtreatment in services. *Working Paper, The Wharton School* .

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

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

Reed, R. 2011. Induction: a step by step guide. *MidwifeThinking* URL `http://midwifethinking.com/2011/07/17/induction-a-step-by-step-guide/`. Published: 2011-07-17. Accessed: 2015-05-30.

Renfrew, M., W. Hannah, L. Albers, E. Floyd. 1998. Practices that minimize trauma to the genital tract in childbirth: A systematic review of the literature. *Birth* **25**(3) 143–160.

Robinson, L. W., R. R. Chen. 2011. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management* **13**(1) 53–57.

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

Rosenthal, E. 2013. American way of birth, costliest in the world. *The New York Times* Published 2013-06-30. Accessed: 2015-05-15.

Schultz, K.L., D.C. Juran, J.W. Boudreau, J.O. McClain, L.J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Science* **44**(12-part-1) 1595–1607.

Shen, Z.-J. M., X. Su. 2007. Customer behavior modeling in revenue management and auctions: A review and new research opportunities. *Production and operations management* **16**(6) 713–728.

Shumsky, R. A., E. J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Smith, M., R. Saunders, L. Stuckhardt, J. M. McGinnis, eds. 2012. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. National Academies Press, Washington, D.C.

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

Tan, T. F., S. Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

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