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 customers to a specialist. We find 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 service 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, non-complex cases are 28.8% less likely to receive an epidural, leading to a cost reduction of 8.7%, while complex cases are 14.2% more likely to be referred for a physician-led delivery, leading to a cost increase of 2.6%. These observations are consistent with overtreatment at both high and low workload levels, albeit for different types of patients, and suggest that smoothing gatekeeper workload would reduce variability in customer service experience.

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

History: February 11, 2016

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) trades off the

desire to protect specialists' valuable time with 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 (which we call the service configuration) 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 or the service configuration 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, which we describe in more detail in §3. 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 (16,355 births) and appropriate econometric models to investigate whether this is indeed the case.

In §4, we build on existing theory to 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: As workload increases, complex cases are more likely to be referred, while less complex cases are more likely to receive less resource-intensive services. Indeed, as we report in §5–7, 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 workload, defined as the patient–midwife ratio,

increases from two standard deviations below to two standard deviations above the mean (0.39 to 1.90 patients per midwife), patients with non-complex needs, defined as those with spontaneous onset of labor, are approximately 28.8% less likely to receive a resource-intensive pain management intervention. In contrast, patients with complex needs, defined as those for who labor was pharmacologically induced in the hospital prior to their arrival at the DU, are approximately 14.2% 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, these empirical observations suggest that the modeling literature on the GP problem, which ignores the presence of both of these endogenous workload-management 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 can help to assess the costs associated with workload-induced changes in GP behavior. In the DU case, the rationing-related reduction in the cost of treating non-complex cases at high rather than low workload (two standard deviations above rather than two standard deviations below the mean) is estimated to be approximately £200 per patient (or -8.7%). By contrast, the workload-induced increase in referrals, which only affects complex cases, increases costs by approximately £61 per patient (or +2.6%). In addition to costs, there are also implications for operationally relevant outcomes and, surprisingly, the behavioral change at higher workload does not necessarily lead to uniformly worse outcomes for these patients. For example, the rationing of non-complex cases leads to a reduction in maternal post-birth length of stay (LOS). 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 lower complexity cases due to the rationing effect. By understanding how workload affects GP behavior, 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. We investigate this further in §8.

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 such unnecessary care, a point we illustrate further in the context of the DU in §8.

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. In a similar vein, Zhang et al. (2011) present a two-tier system for security-check queues.

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 often 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. Mas and Moretti (2009) show also that productivity and service times can be affected by peer effects, with supermarket cashiers speeding up in the presence of highly productive coworkers. 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. Due to data availability, as well as the importance of the setting, 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 (2015) 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 (2015) 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.[1] 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, with a reduction in referrals for complex cases). 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, with an increase in discretionary interventions for non-complex cases). 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 characteristics of the customers' service needs, 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.

[1] 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 by 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 higher risk 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 midwife-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 – 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,[2] (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. This means that staff have no personal financial incentive to advise for or against any particular course of treatment (Lilley 2003). The unit in question is staffed by three types of employees:

[2] 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).

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, with the rota scheduled at least two months in advance, 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 doctors 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 a physician, 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. The DU, for example, aims 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 regularly goes into parallel-processing mode, where a single midwife is in charge of more than one delivery. 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 management of labor and 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. Such 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 previous literature (e.g., 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 (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, providing guidance and support during the final stages of labor, caring 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 §3, 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. We note that this reduction in epidural propensity at higher workload is expected to take place at the margin, i.e., to not affect those patients experiencing the most severe pain for whom the epidural decision is less discretionary.

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. Congestion, however, creates the need for GPs to speed up, leading to decisions based on less complete evidence (Alizamir et al. 2013). 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 a physician-led birth – either an instrument delivery or an emergency C-section – when information becomes available that renders the service too challenging for them to manage safely without physician assistance. (Note that we bundle together all physician-led deliveries since the decision whether to perform an instrumental delivery or an emergency C-section lies with the physician and not the midwife.) Similar to discretionary services, specialists become a buffer that the midwives can use to manage their workload. We would therefore expect referrals for physician-led deliveries to increase when midwife workload increases.

4.3. The role of complexity

Customers in the service system are typically heterogeneous in their service needs (Shen and Su 2007). More specifically, some customers will be relatively straightforward to serve, and the GP will be well placed to do so. Others will exhibit more complex needs that require specialized knowledge and/or skills that go 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 customers 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 customers with complex service needs than non-complex needs for two reasons. First, customers with complex needs 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 case, for which, by definition, the needs are greater. 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 cases, rationing becomes a relatively more important workload management method for such customers. Put differently, 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 customers with non-complex needs, 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 non-complex needs than for customers with complex needs.

Together H3 and H4 suggest that there is a divergence in the service experience of customers 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.

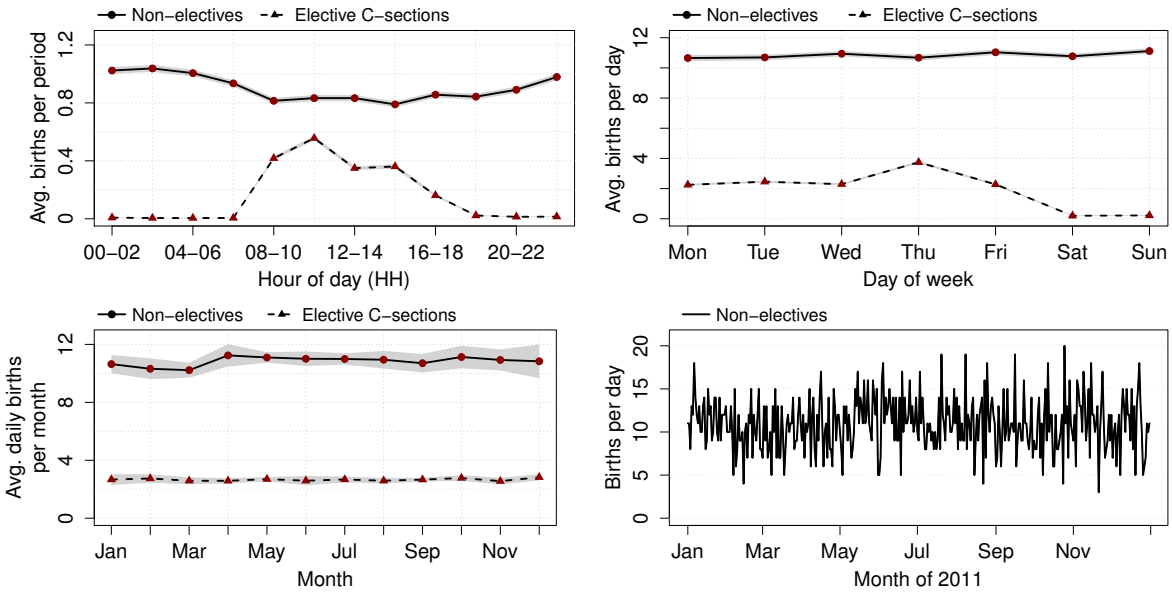
In the specific setting of this study we operationalize the service episode complexity using the type of onset of labor – specifically, whether the labor started spontaneously or it was pharmacologically induced in the hospital prior to the arrival at the DU. Women with spontaneous onset of labor tend to have less complex needs than induced patients as induction changes the birth process in several ways (Lothain 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 cases. Equally importantly, the mode of labor onset is both exogenous to the DU workload and is readily observable by the midwife (as opposed to other measures of complexity that are only observable ex post). Finally, inductions are sufficiently frequent to provide the requisite statistical power. In our context we expect H3 to translate into patients with more complex service needs (i.e., those that arrive with pharmacologically induced labor) being more likely to be referred for a physician-led delivery as workload increases vis-à-vis patients with less complex needs (i.e., those that arrive directly from the community after the spontaneous onset of labor). Similarly, we expect H4 to translate into patients with less complex service needs being less likely to receive epidural pain relief as workload increases vis-à-vis patients with more complex needs.

5. Data and Variable Description

To investigate the hypotheses we collaborated closely with the DU of the 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 at any time during this period.

In total, 23,300 births occurred in the DU during the observation period, or approximately 13 births per day. In the construction of the main sample we exclude elective C-sections (3,506 births)

Figure 1 **Number of births by hour of day, day of week and month of year (mean with 95% CI) for all 5 years, and time series of number of births per day in 2011.**



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 midwife-led birthing unit. These patients were escalated to the DU at an advanced stage of labor specifically because a specialist was needed to manage their service, meaning that DU midwives do not act as GPs. 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). This leaves a final sample of 16,355 births. Importantly, all patients excluded from the analysis sample are still included in the estimation of the workload measures since a DU midwife is still assigned to assist with their care.

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 patient episodes (PEs) as the unit of analysis. A PE begins when the patient arrives to the DU and ends with the delivery. The main independent variable, GP workload, is the standardized time-weighted average number of

patients per midwife for the period three hours prior to birth, which will be described in more detail below. We note that as in other studies in the healthcare context (e.g. KC and Terwiesch 2009, Kuntz et al. 2014), we measure workload at the organization rather than the server level (e.g. as in Tan and Netessine 2014). This is partly because accurate real-time information on the midwife-patient assignment is unavailable, but also because the endogenous allocation of heterogeneous patients – who place different levels of demand on resources – to midwives whose skill and experience levels vary means that server level workload will also be endogenous. We further note that, in contrast to the aforementioned healthcare studies that measure server workload using patient-only measures, our detailed staffing data – which includes real-time information on how many midwives were present in the DU – allows us to accurately account for variation in server availability.

More specifically, to calculate workload, 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 analysis sample, as explained above) and $MW(t)$ is the number of midwives, the (instantaneous) workload at any time t can be expressed as

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

The time-weighted average load for a patient i who gives birth at time b_i is then calculated 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. The three hour averaging period was chosen to coincide with the average duration of the second and final (prior to delivery) stage of labor. Averaging over different time periods (e.g. one, two or four hours) yields highly correlated workload measures and almost identical results. Focal patient i is excluded from the patient counter $N_i(t)$ in (1) to avoid the reverse causality problem. Specifically, this ensures that $LOAD_i(t)$ and $LOAD_i$ are independent of the length of time that patient i spent in the DU, and so independent also of the impact of any GP decision that affects that patient. Nevertheless, including the focal patient in $N_i(t)$ does not invalidate our conclusions (see §2.3 of the online supplement for further details).

To ensure that the workload variable remains stationary over the five observation years we take its z-score over a 12-month moving window. To do this we subtract the mean and divide by the standard deviation of the instantaneous workload, both calculated over a period from 6 months prior to 6 months after the time $t = b_i$ of birth i, giving the standardized time-weighted workload

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

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	1.11	0.37	0.12	3.20	0.97***	-0.44***	-0.11***	-0.05***	-0.01	-0.00	-0.03***
(2) Standardized workload	-0.09	0.93	-3.02	4.37		-0.34***	-0.11***	-0.05***	-0.00	-0.01	-0.02*
(3) No. of midwives present	7.93	1.13	4.00	12.00			0.02**	0.02*	0.01†	-0.00	0.03***
(4) Complex patient episode	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)	42.97	39.29	2.90	266.95							0.76***
(8) Cost (£)	2,281.82	1,691.59	337.99	9,929.23							

The number of midwives is measured as the time-weighted average over the same time interval as workload; $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{\dagger}p<0.10$.

where, $\mu(LOAD_i)$ and $\sigma(LOAD_i)$ are given by

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

respectively, where $\underline{w}_i$ is the time 6 months prior to birth, $\overline{w}_i$ is the time 6 months post birth, and $V_1=\sum_{k\in L(\underline{w}_i,\overline{w}_i)}\int_{\underline{w}_i}^{\overline{w}_i}\mathbb{1}[LOAD_i(t)=k]\,dt$. When we do not have activity data for the the entire 1-year time window the standardization process occurs over a shifted 1-year time window, centered at the date closest to $t=b_i$ for which sufficient activity data is available. (For more information on the standardization of workload see §2.4.3 of the online supplement.)

The average (unstandardized) workload is 1.11 – see Table 1 – suggesting that, on average, a focal mother experiences a workload of 1.11 other patients per midwife present in the unit. The target is to have one midwife present per patient in active labor (NAO 2013), a target achieved for about 74% of deliveries. The histogram of (standardized) workload, shown in Figure 2 (left), shows that the workload distribution is approximately normal with a fair number of patients treated during periods of extreme workload, which aids the empirical identification of workload effects.

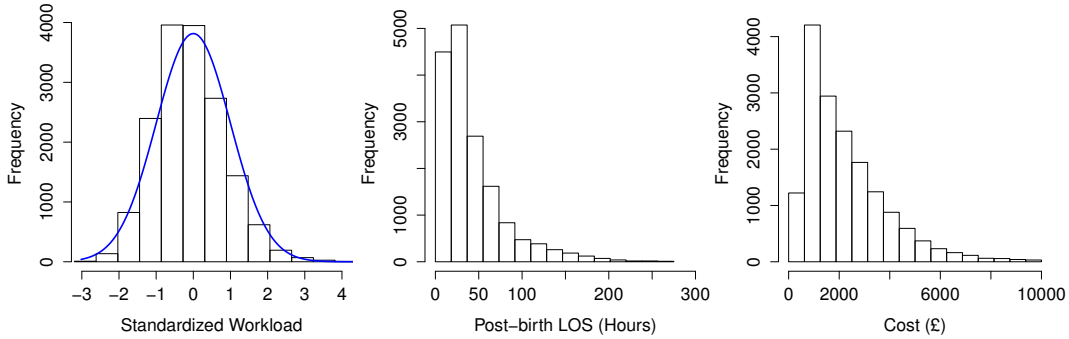
The second independent variable, also reported in Table 1, is a binary variable that takes the value one if the PE is complex, operationalized by the need for pharmacological induction, and zero otherwise. In the final sample 38% of PEs are complex, with the mix of non-complex and complex PEs not exhibiting any systematic variability (e.g. within- or between-day variability).[3]

5.2. Dependent variables

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 analgesia and zero otherwise and (ii) an indicator variable that takes the value one if the patient was referred to a physician and zero otherwise.

[3] We note that the small correlation between a complex PE and a physician-led delivery (0.03) in Table 3 is misleading as it does not account for confounders; doing so we find the complex cases to be 13.4% more likely to be physician-led.

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



Additionally, to examine the implications of workload induced changes in GP behavior, data was collected 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 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 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 may occur during vaginal delivery; 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 may increase the risk of a patient being referred to a physician (Liu and Sia 2004). Therefore, the model takes the form

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

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

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

To investigate whether complexity of the PE 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 non-complex and complex PEs.

Although it is possible to estimate the two models above sequentially using the standard maximum-likelihood probit methodology, one concern is simultaneity/endogeneity bias. For example, a patient identified for referral may also be given an epidural in order to reduce discomfort during the more invasive delivery, or there may be omitted variables, which are observable to the midwife but not to us, the researchers, that affect both the decision to administer epidural analgesia and to refer to a physician. Not accounting for simultaneity/endogeneity could lead to a biased estimate of the epidural coefficient, β_3, in (6). Furthermore, if the epidural decision is also correlated with workload, that is if $\alpha_2 \neq 0$ in (4), then the coefficient β_2 of the workload variable in (6) would also be biased. To account for this we also estimate equations (5) and (7) simultaneously using the recursive bivariate probit (BiProbit) model (Maddala 1983, p. 123–129). The main change is in the structure of the errors, which are jointly distributed according to the standard bivariate normal distribution with unit variances and correlation coefficient ρ. In this model, ρ is

a parameter to be estimated.[4] To improve the reliability of the estimation we include two instrumental variables (IVs) (Wilde 2000, Maddala 1983). These are variables that affect the epidural decision, and therefore appear in the first equation (i.e., are relevant), but do not affect the referral decision, and therefore do not appear in the second equation (i.e., are valid). The two IVs used in this analysis are introduced below. We leave the details of their calculation for Appendix B.

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 referral decision (which occurs near to the time of birth) makes it unlikely that it will have any direct impact on the outcome equation. To ensure that this is the case, the instantaneous operating theater use at the time of birth is controlled for in the outcome equation to remove any potential residual effect resulting from serial correlation.

The second IV is the distance between the hospital and the patient's place of residence. The distance to facilities is used commonly in the medical literature as an IV for exposure to available treatments at those facilities (see e.g., Brookhart et al. 2010). This IV is specific to those patients who present after spontaneous onset of labor, and so is only usable in estimations for the non-complex PEs. For these patients, this IV will affect whether or not they receive an epidural (i.e., be relevant) since the further the distance they must travel, the more likely they are to arrive at the hospital in a more advanced stage of labor, when epidural analgesia is contraindicated. There is also no reason to suspect that distance from the hospital directly affects the likelihood of a patient receiving a physician-led delivery if necessary, and no evidence that there exist effective preventative actions that might be taken earlier in labor (see NICE 2012). To be sure also that patients who live further from the hospital do not differ in their risk, we control for the level of 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).

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

[4] Note that if the epidural decision is not correlated with workload, i.e. if $\alpha_2 = 0$ in (4), then the coefficient of workload, β_1, in (7) would not be biased even if simultaneity/endogeneity were an issue. In this case, if the goal is to find an unbiased estimate of the workload coefficient β_1, estimating the simpler univariate model will suffice.

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	0.25	0.47	0.00	2.00	0.13***	0.14***	0.01	-0.03***	-0.02**	-0.11***	-0.05***	-0.07***
(9) 2–4h op. tht. use	0.30	0.37	0.00	2.54	0.20***	0.21***	0.04***	-0.03***	-0.01	0.04***	0.02**	0.03***
(10) Dist. to home (km)	15.97	15.55	0.05	469.34	0.00	0.00	-0.01†	0.02**	-0.01	-0.00	0.02**	0.04***

(1) Workload, (2) Std. workload, (3) No. midwives, (4) Complex PE, (5) Epidural, (6) Phys.-led delivery, (7) Post-birth LOS, (8) Cost
$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{\dagger}p<0.10$.

correlations with the variables presented earlier in Table 1. Formal hypothesis testing (see §6 of the online supplement) of the IVs for under-, over- and weak identification using standard testing procedures provides strong evidence that the IVs are relevant (p-value < 0.01), have low maximal relative bias (of between 10% and 15%), and are not invalid (p-value > 0.05, as required).

6.2. Results

Tables 3 and 4 report estimated average partial (marginal) effects (APEs) with robust standard errors for the service configuration and referral decisions, respectively. 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 2.5% (APE $= -0.025$, p-value < 0.001). Re-estimating this model separately for the non-complex and complex PEs in Probits (2) and (3), respectively, we find that workload has a strong effect on the service configuration for the non-complex segment (APE $= -0.025$, p-value $<$ 0.001) but does not appear to affect the complex segment (APE $= -0.006$, p-value $= 0.415$).

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 $= 0.002$, p-value $= 0.542$) or not (APE $= -0.002$, p-value $= 0.497$). Separating the non-complex and complex PEs and re-estimating, in Probit (2)–(3) of Table 4 we find evidence that workload does in fact affect referrals: At higher levels of workload the referral rate for the non-complex PEs is lower (APE $= -0.010$, p-value $= 0.021$), while the referral rate is higher for the complex PEs (APE $=$ 0.014, p-value $= 0.026$). Interestingly, the directions of the effects on the sub-samples are opposing and so cancel out in the aggregated model, explaining the null results in Probit (1) and (4).

When we include epidural analgesia as a possible mediator variable, the estimated workload effect for complex PEs is not affected. More specifically, the estimated workload effect in Probit (6) of Table 4, which includes epidural analgesia as an explanatory variable, remains positive

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

	Probit		
Complexity	(1) Epidural *All*	(2) Epidural *nC*	(3) Epidural *C*
Std. workload	-0.025***	-0.025***	-0.006
	(0.004)	(0.005)	(0.007)
Dist. to home	-0.006	-0.017***	0.004
	(0.004)	(0.005)	(0.007)
2–4h op. tht. use	-0.013	-0.035*	0.025
	(0.015)	(0.018)	(0.025)
Inst. op. tht. use	-0.008	-0.005	-0.017
	(0.008)	(0.009)	(0.013)
N	16,355	10,091	6,264
Log-lik	-9,624.01	-5,318.66	-3,835.32
Pseudo-R^2	0.098	0.097	0.117

nC and *C* refer to non-complex and complex patient episodes, respectively; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

(APE = 0.015, p-value = 0.015) and nearly identical in value to that of Probit (3) of Table 4, which did not include epidural analgesia as an explanatory variable. Note that for complex PEs there is no need to estimate a bivariate probit model as workload has no impact on the epidural decision for these patients (see Probit (3) of Table 3) and so, as explained in footnote 4, the presence of simultaneity/endogeneity would not bias the estimated workload coefficient.

In contrast, when we include epidural analgesia as a possible mediator variable the estimated workload effect for non-complex PEs effectively disappears. This is evident in both Probit (5) of Table 4 (APE = -0.006, p-value = 0.177), which does not account for endogeneity or simultaneity, as well as the BiProbit models of Table 4 (APE = -0.000, p-value = 0.993), which does correct for such issues. This indicates that for non-complex PEs 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 a physician's assistance (APE = 0.193, p-value < 0.001) the decrease in epidural rates at higher workload has the effect of reducing the rate of referrals to physicians.

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 workload two standard deviations below the sample mean (-1.95, or approx. 0.39 patients per midwife) and two standard deviations above the sample mean (1.78, or approx. 1.90 patients per midwife), respectively. For non-complex PEs the estimated epidural rate falls from 32.0% at low workload to 22.8% at high workload, a relative decrease of 28.8%. This reduction in epidural rates leads to a decrease in the physician-led delivery rate from 39.6% to 35.8%, a relative decrease of 9.6%, with no additional direct effect of workload.[5] For complex PEs,

[5] We note that these two effects work in opposite directions: To reduce her workload the GP rations the provision of

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

	Probit						BiProbit	
Complexity	(1) Phys. *All*	(2) Phys. *nC*	(3) Phys. *C*	(4) Phys. *All*	(5) Phys. *nC*	(6) Phys. *C*	(1) Epidural *nC*	(2) Phys. *nC*
Std. workload	-0.002	-0.010*	0.014*	0.002	-0.006	0.015*	-0.022***	-0.000
	(0.004)	(0.004)	(0.006)	(0.003)	(0.004)	(0.006)	(0.004)	(0.005)
Epidural	–	–	–	0.191***	0.201***	0.198***	–	0.193***
	–	–	–	(0.007)	(0.010)	(0.011)	–	(0.010)
Dist. to home	-0.007†	-0.011*	-0.001	-0.006	-0.007	-0.002	-0.017***	–
	(0.004)	(0.005)	(0.006)	(0.004)	(0.004)	(0.006)	(0.004)	–
2–4h op. tht. use	0.001	0.006	-0.012	0.003	0.013	-0.018	-0.028†	–
	(0.013)	(0.016)	(0.022)	(0.012)	(0.016)	(0.021)	(0.016)	–
Inst. op. tht. use	-0.097***	-0.091***	-0.100***	-0.095***	-0.088***	-0.097***	-0.004	-0.086***
	(0.007)	(0.009)	(0.012)	(0.007)	(0.009)	(0.012)	(0.008)	(0.009)
N	16,355	10,091	6,264	16,355	10,091	6,264	10,091	
Log-lik	-8,002.64	-4,810.23	-3,089.24	-7,623.05	-4,575.89	-2,927.96	-9,887.06	
Pseudo-R^2	0.262	0.275	0.265	0.297	0.311	0.303	–	
ρ	–	–	–	–	–	–	-0.453*** (0.091)	

nC and *C* refer to non-complex and complex patient episodes, respectively; *Robust standard error* in parentheses for Probit; *Bootstrapped standard error* in parentheses for BiProbit, 10,000 simulations; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

we find no significant effect on the epidural rate, while the estimated physician-led delivery rate increases from 37.1% to 42.3% as workload increases from the low- to the high-workload scenario, a relative increase of 14.2%.

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 non-complex PEs 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 complex PEs, 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.

epidural analgesia but, as a result, fewer of these patients need to be referred 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.

7.1. Econometric specification and instrumental variables

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

$$\ln(PbLOS_i) = \gamma_0 + \mathbf{X}_i\boldsymbol{\gamma}_1 + ZLOAD_i\gamma_2 + EPI_i\gamma_3 + PHYS_i\gamma_4 + \mu_i \tag{8}$$
$$\mu_i \sim \mathcal{N}(0, \sigma_\mu^2)\,.$$
$$\ln(COST_i) = \lambda_0 + \mathbf{X}_i\boldsymbol{\lambda}_1 + ZLOAD_i\lambda_2 + EPI_i\lambda_3 + PHYS_i\lambda_4 + \nu_i \tag{9}$$
$$\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)\,.$$

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 $\mathbf{X}_i$. These models are estimated using the classic ordinary least squares (OLS) method.

To mitigate potential endogeneity concerns, we supplement the OLS models above with Heckman treatment effects (HeckTreat) models, using appropriate IVs (Maddala 1983, p. 123–129). The HeckTreat models ensure that the estimated impact of workload on outcomes is not biased by the presence unobservable variables that make a patient more likely to both receive discretionary services (or be referred to a physician) and have a longer post-birth LOS (or higher costs).

The endogeneity concerns differ depending on whether the PE is complex or not. For non-complex PEs we have found that workload has a direct effect on the service configuration decision (epidural analgesia) 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 for the non-complex PEs (see also footnote 4). In terms of instrumental variables that can help resolve this issue, the distance between the hospital and the patient's home does not satisfy the exogeneity condition: A patient who lives further from the hospital may be more likely to be delayed in being discharged – and so increasing their post-birth LOS and cost – since if a problem subsequently arises it will take longer for the patient to return to the hospital. For the non-complex segment we therefore drop the distance IV and employ only the operating theater use two to four hours

prior to delivery as an IV, since for the latter there is no reason to believe that this would have an impact on post-birth LOS or costs other than through the epidural decision.

In contrast, for complex PEs we have found that workload has no effect on the service configuration decision but does increase the rate of referrals. Therefore, for complex PEs we need to only account for the potential endogeneity of the referral decision in the outcome equations – the logic behind this is similar to that described in footnote 4. Considering next the IVs, in this case it is clear that neither the operating theater usage *prior* to birth nor the distance to the hospital would be suitable as they do not satisfy the relevancy condition (i.e., they do not affect the referral rate directly) – see Probit (6) of Table 4. Instead, for the referral decision we use the instantaneous operating theater use 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 physician-led delivery (APE $= -0.097$, 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, in addition to being relevant, this IV is also expected to be valid.

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 For non-complex PEs the model OLS (1), which does not control for the earlier service configuration (epidural analgesia) decision, suggests that an increase in workload leads to a decrease in post-birth LOS (coef. $= -0.018$, p-value $= 0.043$). However, when we account for epidural analgesia – either with the OLS (3) model, which does not account for non-random selection, or with HeckTreat (1–2) model, which does – the coefficient of workload becomes statistically indistinguishable from zero, suggesting that the aggregate decrease in post-birth LOS reported in OLS (1) is entirely due to the rationing of epidural analgesia that occurs at higher levels of workload. Using the HeckTreat model, we estimate that moving from low- to high-workload conditions indirectly (through the reduction in epidural rates) causes an 8.3% decrease in post-birth LOS. This indirect effect is calculated using the full marginal effect of the HeckTreat model derived in §7 of the online supplement.

For complex PEs there is little evidence that workload affects post-birth LOS. Workload does not affect LOS directly (see OLS (2) or OLS (4) models). Although OLS (4) suggests that women who have had a physician-led delivery have a longer post-birth LOS (coef. $= 0.432$, $p < 0.001$),

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

	OLS				HeckTreat			
	(1) PbLOS	(2) PbLOS	(3) PbLOS	(4) PbLOS	(1) Epidural	(2) PbLOS	(3) Phys.	(4) PbLOS
Complexity	*nC*	*C*	*nC*	*C*	*nC*	*nC*	*C*	*C*
Std. workload	-0.018*	0.007	-0.009	-0.001	-0.078***	0.008	0.071**	0.010
	(0.009)	(0.011)	(0.009)	(0.010)	(0.016)	(0.009)	(0.023)	(0.011)
Epidural	–	0.309***	0.343***	0.220***	–	1.073***	0.705***	0.328***
	–	(0.019)	(0.015)	(0.018)	–	(0.049)	(0.039)	(0.027)
Phys. delivery	–	–	–	0.430***	–	–	–	-0.093
	–	–	–	(0.020)	–	–	–	(0.094)
2–4h op. tht. use	-0.043	-0.046	-0.031	-0.038	-0.114*	–	–	–
	(0.030)	(0.037)	(0.030)	(0.036)	(0.054)	–	–	–
Inst. op. tht. use	-0.061***	-0.035†	-0.059***	0.004	-0.017	-0.054**	-0.319***	–
	(0.016)	(0.020)	(0.016)	(0.020)	(0.031)	(0.017)	(0.046)	–
N	10,091	6,264	10,091	6,264	10,091		6,264	
Log-lik	-11,145.93	-6,516.35	-10,949.43	-6,317.40	-16,220.03		-9,242.71	
Adj-R^2	0.302	0.279	0.329	0.324	–		–	
ρ	–	–	–	–	-0.553*** (0.030)		0.453*** (0.071)	

nC and *C* refer to non-complex and complex patient episodes, respectively; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

once we correct for the presence of non-random selection with HeckTreat (3–4) – which appears to be important as indicated by the positive and significant value of $\rho = 0.453$ (p-value < 0.001) – the effect of a physician-led delivery becomes insignificant (coef. $= -0.093$, p-value $= 0.321$). This suggests there is no indirect effect of workload either.

7.2.2. Cost For non-complex PEs the model OLS (1), which does not control for the earlier service configuration (epidural analgesia) decision, suggests an overall decrease in costs with increasing workload (coef. $= -0.016$, p-value $= 0.021$). However, OLS (3) suggests that this decrease is explained by the reduction in epidural rates at higher workload, since the direct effect of workload becomes insignificant after controlling for epidural analgesia. More reliably, the HeckTreat (1–2) model that accounts for endogeneity of the epidural decision – which appears to be important since the model estimates a significantly negative correlation $\rho = -0.754$ (p-value < 0.001) – finds (weak) evidence of a positive direct effect of DU workload on cost (coef. $= 0.013$, p-value $= 0.083$). Translated into cost terms, the direct effect of moving from low- to high-workload conditions is an increase in costs by £114.67 (5.0%) per PE, or £1,157,200 in total across all non-complex PEs. There is, however, also strong evidence of an opposing negative indirect effect of workload on cost via the rationing of epidurals: Epidurals have a strong positive effect on costs after accounting for endogeneity (coef. $= 1.162$, p-value < 0.001), and workload has a strong negative effect on the likelihood of a patient receiving an epidural (coef. $= -0.074$, p-value < 0.001). This is equivalent to an £199.62 (8.7%) per PE, or £2,014,400 in total, decrease in costs caused by the rationing of epidurals when workload increases from low to high. Put together, moving from low- to high-workload conditions results in a decrease in costs of 3.7%, which is equal to a saving of £84.95 per PE or a total saving of £857,200 across all non-complex PEs.

Table 6 Coefficient Estimates in Statistical Models for Cost

	OLS				HeckTreat			
Complexity	(1) Cost *nC*	(2) Cost *C*	(3) Cost *nC*	(4) Cost *C*	(1) Epidural *nC*	(2) Cost *nC*	(3) Phys. *C*	(4) Cost *C*
Std. workload	-0.016*	0.016*	-0.007	0.009	-0.074***	0.013†	0.072**	0.014†
	(0.007)	(0.008)	(0.007)	(0.007)	(0.016)	(0.008)	(0.023)	(0.008)
Epidural	–	0.321***	0.353***	0.236***	–	1.162***	0.695***	0.253***
	–	(0.014)	(0.012)	(0.013)	–	(0.036)	(0.039)	(0.023)
Phys. delivery	–	–	–	0.412***	–	–	–	0.328***
	–	–	–	(0.015)	–	–	–	(0.094)
2–4h op. tht. use	-0.004	-0.016	0.007	-0.008	-0.054	–	–	–
	(0.023)	(0.027)	(0.022)	(0.026)	(0.047)	–	–	–
Inst. op. tht. use	-0.076***	-0.052***	-0.073***	-0.014	-0.020	-0.067***	-0.355***	–
	(0.012)	(0.015)	(0.012)	(0.015)	(0.029)	(0.014)	(0.047)	–
N	10,091	6,264	10,091	6,264	10,091		6,264	
Log-lik	-8,374.30	-4,518.81	-8,005.69	-4,164.51	-13,221.79		-7,101.73	
Adj-R^2	0.390	0.368	0.433	0.436	–		–	
ρ	–	–	–	–	-0.754*** (0.022)		0.111 (.118)	

nC and *C* refer to non-complex and complex patient episodes, respectively; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

For complex PEs there is evidence, given in OLS (2), that an increase in workload leads to an increase in costs (coef. = 0.016, p-value = 0.040). The insignificant value of ρ (coef. = 0.111, p-value = 0.348) in HeckTreat (3–4) suggests that there is little evidence of a selection effect. Therefore the direct and indirect effects of workload can then be inferred from the change in the workload coefficient when comparing the models in OLS (2) and OLS (4), i.e., in this case the problem reduces to a basic mediation analysis. Translated into cost terms, the direct effect of moving from low- to high-workload conditions is an increase in costs of £78.51 (3.4%) per PE, or £491,800 in total across the complex PEs. Added to this is an increase of £60.66 (2.6%) per patient resulting from the increased rate of physician-led deliveries, leading to a total cost increase per PE of £139.17, or £871,800 when aggregated across all complex PEs 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 (results available in §4 of the online supplement) 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 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 IVs 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. Alternative Explanations and Robustness Checks

We begin this section by discussing two alternative explanations for the results presented in the previous sections. First, in the service setting we study the decision as to whether a patient should be referred to a physician is made by the midwife, but whether or not the patient actually receives physician-led care also depends on the availability of the physicians. Therefore, if midwife and physician workload are not independent, then it is possible that the observed increase in referrals at higher workload may be attributable to changes in physician behavior rather than changes in referral behavior by the midwives. We note that we partially control for physician workload with the occupancy of the operating theatres (a measure that is correlated with demands on physician time) and with a set of time-fixed effects (which are correlated with physician supply). Furthermore, if anything, we would expect this effect to bias our results towards zero: physician workload is more likely to be positively correlated with midwife workload and physicians are less likely to be able to accept a referral when their workload is high (and not more likely, as we find). Second, it is also possible that the increase in referrals is not due to gatekeeping behaviour but is, in fact, necessitated by unobservable safety concerns that arise because of a deterioration in the quality of care provided at higher levels of workload. If this were the case, however, then we should also expect there to be a direct effect of workload on observable quality measures associated with the health status of the mother or the baby. Since we find no such effects (e.g., for post-birth LOS, baby Apgar scores, and other measures not reported here) we believe that this is not the case. In §5 of the online supplement these two alternatives are discussed in more detail.

To confirm the robustness of the results presented in the previous sections we also estimate a number of alternative model specifications that: (i) expand the definition of a complex PE to include other ex ante observable factors such as breech and multiple births; (ii) control for midwife fixed effects and antenatal unit occupancy; (iii) employ interaction effect models to compare complex and non-complex PEs instead of subgroup analysis; (iv) measure workload over different time windows and post-birth LOS in nights rather than on a continuous scale; and (v) allow for non-linear workload effects and the inclusion of the focal patient in the workload measure. The results

are qualitatively similar to those reported in the paper, and can be found in §§2–3 of the online supplement.

8. Discussion and Implications for Staffing

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 the DU, we evaluate the implications of an increase in midwife staffing to a level that raises the proportion of mothers receiving the desirable one-to-one (or better) level of care during active labor from the current level of 74% to 90% 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 (2015), 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 74% one-to-one care and would have to increase this to nine or 10 midwives to achieve the 90% or 95% levels, respectively. Using current estimates for the full economic cost of a midwife (Curtis 2012), this would add approximately £355,500 and £678,200, respectively, to the staffing bill per annum.

As shown in Table 7, increasing staffing to nine midwives (Columns 2–4, 90% one-to-one service) is associated with an average relative increase (at the mean workload level and compared with the observed workload in the data) of 1.9% in the epidural rate and of 0.2% in the rate of physician-led deliveries across all PEs (Column 2). The use of discretionary services and specialists increases further when staffing rises to 10 midwives (Columns 5–7, 95% one-to-one service), to increases of 3.2% in epidural rates and 0.3% in referral rates. Contrary to conventional assumptions, the increase in staffing leads to a *deterioration* in average service outcomes, as indicated by an increase

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

	90% One-to-one service (Fixed staffing)			95% One-to-one service (Fixed staffing)			95% One-to-one service (Variable staffing)		
Complexity	*All*	*nC*	*C*	*All*	*nC*	*C*	*All*	*nC*	*C*
Avg. workload	0.960	0.985	0.919	0.864	0.886	0.827	0.929	0.945	0.898
Avg. MW per obs.	9	9	9	10	10	10	9.17	9.22	9.08
Epidural analgesia									
– 10^{th}%ile	5.8%	6.6%	0.7%	9.1%	10.9%	1.2%	6.9%	7.9%	0.7%
– Mean	1.9%	3.6%	0.4%	3.2%	6.1%	0.7%	2.4%	4.6%	0.5%
– 90^{th}%ile	0.6%	2.5%	0.2%	1.0%	4.4%	0.4%	0.7%	3.4%	0.4%
Phys. deliveries									
– 10^{th}%ile	0.8%	3.5%	-3.0%	1.7%	5.3%	-5.0%	1.3%	3.8%	-3.2%
– Mean	0.2%	1.1%	-1.2%	0.3%	1.9%	-2.1%	0.3%	1.4%	-1.5%
– 90^{th}%ile	0.1%	0.5%	-0.8%	0.1%	0.9%	-1.2%	0.1%	0.6%	-0.9%
PbLOS (hours)									
– 10^{th}%ile	0.5%	0.7%	-0.3%	0.6%	1.0%	-0.5%	0.4%	0.8%	-0.3%
– Mean	0.3%	0.7%	-0.2%	0.5%	1.1%	-0.4%	0.4%	0.8%	-0.3%
– 90^{th}%ile	0.2%	0.7%	-0.1%	0.6%	1.1%	-0.1%	0.2%	0.8%	-0.0%
Cost (£)									
– 10^{th}%ile	0.4%	0.8%	-0.6%	0.5%	1.1%	-1.1%	0.3%	0.9%	-0.5%
– Mean	0.0%	0.6%	-0.6%	0.1%	1.0%	-1.0%	0.1%	0.7%	-0.7%
– 90^{th}%ile	-0.3%	0.4%	-0.7%	-0.3%	0.8%	-1.0%	-0.2%	0.6%	-0.7%

nC and *C* refer to non-complex and complex patient episodes, respectively; Scenario analysis uses standardized workload excluding the focal patient and with standardization performed using the same mean and s.d. as in (3).

in post-birth LOS by 0.3% in the nine-midwives scenario and by 0.5% in the 10-midwives scenario. Furthermore, with 10 midwives costs rise by 0.1% (over and above the increase in staffing costs).[6]

While the aggregate effects across all PEs are relatively small, they become more pronounced for the sub-samples of non-complex and complex PEs (Columns 3 and 6 for non-complex PEs and Columns 4 and 7 for complex PEs). In particular, in the 10-midwives scenario (Columns 6 and 7), epidural rates, physician-led delivery rates and costs increase markedly for non-complex PEs, by 6.1%, 1.9% and 1.0%, respectively, while referral rates and costs *decrease* for complex PEs by 2.1% and 1.0%, 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. While this behavior may be rational from a load-balancing perspective, from a patient perspective these findings are consistent with overtreatment at both high and low workload. Since we cannot know the appropriate level of treatment for any specific mother, it is not possible to quantify overtreatment or provide more

[6] Despite finding no evidence of non-linear workload effects (see §7.4) in our sample, increasing staffing beyond 10 midwives, at which point almost all patients will receive one-to-one care, suggests that there is likely a limit to the extent to which staffing will affect rationing and referral behavior, and subsequently outcomes and costs. Such extrapolation is beyond the scope of this analysis.

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 10 staff members. This purely hypothetical way of staffing generates smoother GP workloads and leads to an increase in the proportion of patients receiving one-to-one care from 74% to 95% without substantially increasing the average number of staff required (a 13.7% – or £350,900p.a. – increase as compared to 26.6% – or £678,200p.a. – increase under the equivalent fixed staffing policy). Furthermore, flexible staffing reduces the variability in customer experience as workload-related changes in discretionary service and referral rates by GPs are 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 complexity moderates the effect of workload: Customers with non-complex needs experience cuts in discretionary services, while those with complex needs are referred to specialists more often. The interaction between workload and the case mix (e.g. the percentage of complex PEs) – which we do not consider in our analysis because our sample does not have sufficient statistical power – may provide additional insight as to how operational changes such as unit specialization that change the case mix impacts on service performance. For example, by diverting non-complex PEs to midwife-led birthing units the complexity of the residual PEs in the standard DU is thereby increased. The referral effect for the complex PEs may then become even more pronounced because GPs have fewer patients with whom they can apply the second lever – the rationing of discretionary services – to regulate their workload. Such effects may be important to consider when organizing or reconfiguring services that include GPs.

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 significance of the estimated coefficient (+ for positive and significant, − for negative and significant, 0 for

Table 8 Table of Controls

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

nC and *C* refer to non-complex and complex patient episodes, respectively. All estimations made using standard OLS/Probit, without controlling for epidural analgesia or physician-led deliveries. *PROM*/COPD: indicates that a patient had premature rupture of membranes/chronic obstructive pulmonary disease. *Deprivation index*: an index of multiple deprivation. *Health 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 who received an epidural in the four-hour period prior to the focal patient's time of delivery. *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 midwife-led unit.

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.

It is useful 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 known to be associated with an increased likelihood of a patient requiring pain relief and physician assistance during delivery; therefore, a positive coefficient is expected for the "Baby weight" variable in Columns (3–6) in Table 8, as is the case. Furthermore, 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), consistent with the positive coefficient estimates reported in Table 8.

B. Calculation of instrumental variables

The two primary instrumental variables are: (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. 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. 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. 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 $t = s_i$.

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. 2015. How a service process adapts to load: An econometric analysis of patient treatment in the emergency department. 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. 2015. Past the point of speeding up: The negative effects of workload saturation on efficiency and quality. *Management Science (forthcoming)* .

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.

Brookhart, M.A., J.A. Rassen, S. Schneeweiss. 2010. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety* **19**(6) 537–554.

Chan, C. W., V. F. Farias, G. Escobar. 2015. The impact of delays on service times in the intensive care unit. *Management Science (forthcoming)* .

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.

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

Green, L., N. Liu. 2015. A study of New York City obstetrics units demonstrates the potential for reducing hospital inpatient capacity. *Medical Care Research and Review* **72**(2) 168–186.

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

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.

Lilley, R. 2003. *The Insider's Guide to the NHS: How It Works and Why It Sometimes Doesn't*. Radcliffe Publishing.

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.

Lothain, J. 2006. Birth plans: The good, the bad, and the future. *Journal of Obstetric, Gynecologic, & Neonatal Nursing* **35**(2) 295–303.

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.

Mas, A., E. Moretti. 2009. Peers at work. *The American Economic Review* **99**(1) 112–145.

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.

NICE. 2012. Caesarean section. Tech. Rep. Clinical guideline 132, National Institute for Health and Clinical Excellence.

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. The Wharton School, Working paper.

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. London Business School, Working paper.

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.

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.

Zhang, Z.G., H.P. Luh, C.-H. Wang. 2011. Modeling security-check queues. *Management Science* **57**(11) 1979–1995.