Supplementary Material
Gatekeepers at Work: An Empirical Analysis of a Maternity Unit

Michael Freeman
Judge Business School, University of Cambridge 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 s.scholtes@jbs.cam.ac.uk

---

1. Introduction

This appendix contains supporting material designed to accompany the investigation presented in the main paper (Freeman et al. 2016). As such, it should not be read in isolation, but rather it serves as reference material that provides further background and support for the methods and techniques employed, presents additional empirical results that confirm the robustness of the headline findings, and lays out the mathematical derivation of effect size calculations that underly a number of the results.

In §2 we consider a number of alternative workload specifications, such as those including the focal patient, calculated using alternative standardization options, different averaging windows, and considering the possibility of non-linearity. In §3 we provide results based on alternative model specifications, as well as interaction effects models. In §4 we present results calculated for a number of additional outcomes, such as the incidence of perineal tears, the Apgar score, and the length of stay on the delivery unit. In §5 we discuss a number of other potential explanations for the results reported in the main paper, and show why these are less likely than those hypothesized. In §6 we present results from empirical tests designed to test for weak or invalid instruments. Finally, in §7 we show how to calculate effect sizes for a Heckman treatment effects (HeckTreat) model.

2. Measuring Workload

In the main paper we are primarily interested in identifying the effect of delivery unit (DU) workload on midwife rationing and referral behavior. When specifying workload in §5.1 of the main paper

we briefly discussed the workload measure that we employ, the time-weighted average number of patients per midwife over the 3-hour period prior to the time of delivery when *excluding* the focal patient from the calculation. In this section we discuss the rationale behind the workload measure, including the level at which workload is measured, the issue of endogeneity bias and how we avoid this by excluding the focal patient, and the consequence of excluding the focal patient on the properties of the workload measure. We also demonstrate that the headline findings in the main paper are robust against alternative specifications of workload.

2.1. Unit- vs. server-level workload

There are two levels at which workload in the DU could have been measured, (i) the unit level, and (ii) the midwife level. Unit-level workload measures load under the assumption that midwifes constitute a common resource pool which shares workload, while midwife-level workload measures the load of a particular midwife as determined by the number of patients to who she was specifically assigned. In the main paper we selected to take the approach of measuring workload at the unit-level, and this section lays out the justification for that decision.

- *Data issues:* In the available data, information is recorded on which midwife was assigned to which patient at the time of delivery. It is not always the case, however, that the midwife who performed the delivery will be the same midwife as the one who took care of the patient throughout the entirety of that patient's DU stay. The long length of labor, for example, means that patients may be handed over from one midwife to another at some point during their stay. This may happen, for example, when the initial midwife ends her shift on the unit or when patients are reallocated amongst midwives to better cope with the arrival of a new patient who may require particular midwife expertise. Since information is not available on when each midwife was attached to each patient, instead only the midwife assigned at the time of birth, it is not possible to accurately identify the workload of a particular midwife at a specific time.
- *Work sharing:* While there is a main midwife who is responsible for the care of each patient, when a particular midwife is busy then it may become necessary for other midwives in the unit to assume temporary responsibility for the care of her other patients. Sharing of work means that while midwives working in the DU at the same time may appear to have varying workload levels when measured at the midwife level, they will in reality often face similar levels of workload. This is because those midwives who are relatively less busy will tend to pick up additional work from those who are more busy.
- *Patient heterogeneity:* Not all patients are equal when it comes to the demand that they place on a midwife's time. When work is allocated amongst midwives in the DU, assignment takes into account the likely resource intensity of patients based on observable factors such as patient history,

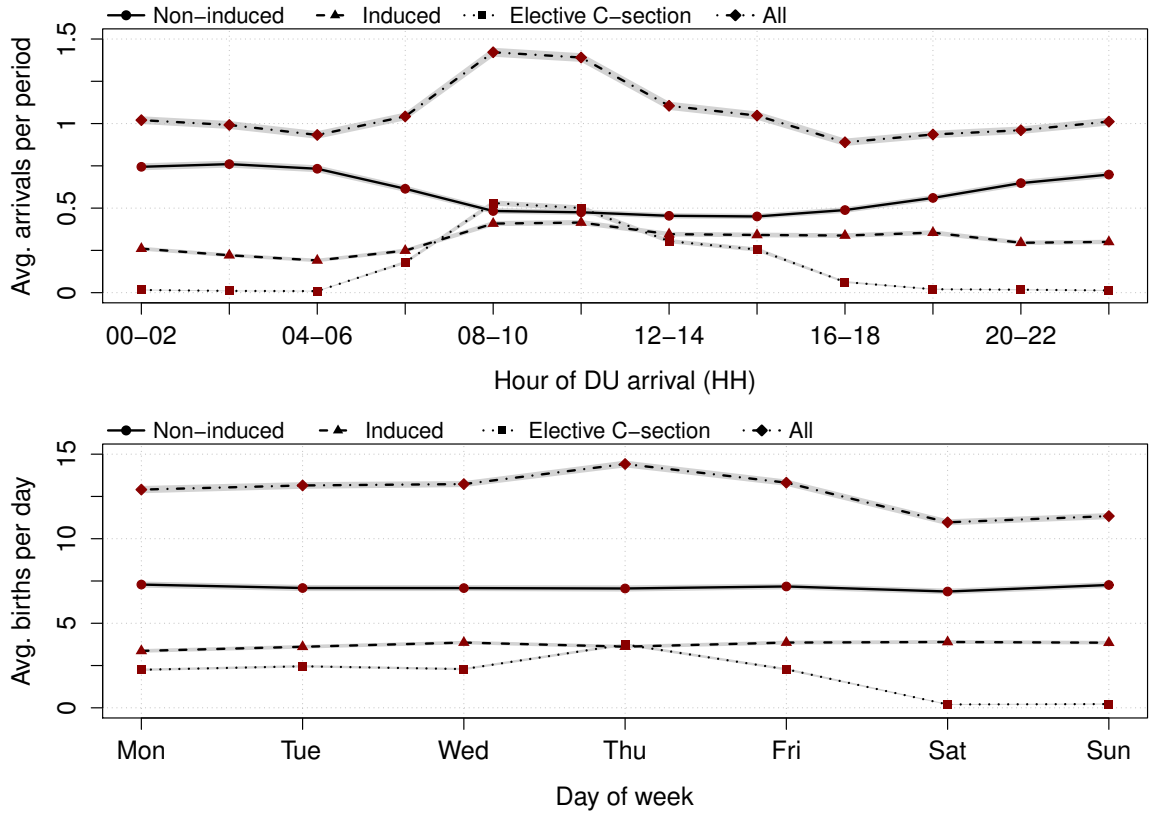
type of onset of labor, whether it is the first or a subsequent delivery, etc. The number of patients to whom a particular midwife is assigned will therefore often be a poor measure of the actual demands placed on that midwife. For example, a midwife allocated to one complex patient may face a similar level of work as a midwife assigned to two non-complex patients, yet a workload measure calculated at the midwife level would suggest that the former is only half as busy as the latter.

- *Staff experience:* Midwives on the DU may have different levels of experience and skill, and more experienced/skilled midwives may be better at parallel processing than their more junior colleagues. A more senior midwife may, therefore, be more likely to be assigned to multiple patients, or to more complicated patients. This means that the number of patients assigned to a particular midwife may poorly reflect their workload since it does not account for differences in a midwife's capacity to cope with multiple or more resource intensive patients.

Based on the reasons outlined above, in this setting it would not be appropriate to use the number of patients assigned to a particular midwife as a measure of server busyness. In particular, to do so we would have to assume that patients are allocated randomly to midwives, regardless of patient needs or midwife experience, that midwives work in isolation without assisting each other or sharing responsibilities, and that midwives never end their shift until all patients under their supervision have delivered. In this setting these assumptions are unrealistic, though for servers in other service settings (such as waiting staff in restaurants (Tan and Netessine 2014) or physicians in the emergency department (KC 2013)) this may not be the case. The approach that we take, that more accurately reflects the nature of work and worker behavior in the DU, is to assume that work is allocated out between midwives in the unit so that each is approximately equally busy, and that if one becomes more busy then other midwives respond by sharing some of the load. This means that a workload variable that is measured at the unit level, calculated by dividing the number of patients by the number of midwives at any point in time, will provide a good snapshot of the relative busyness of the staff.

While unit-level workload accounts for patient heterogeneity and staff experience at a specified point in time, it is possible that these factors may cause workload to differ in unobservable ways *over* time. In particular, if the DU is equally loaded, with e.g. the same number of staff and number of patients present on any two days, but on one of those days the cases are more simple and the other they are more complex, and/or on one of the days the staff are more experienced and on the other they are less skilled, then the DU will be reported as being equally busy when workload is measured at the unit level, but the actual busyness of the midwives may be significantly different. Unlike for a midwife-level measure where patient-staff assignment can be non-random, however, so long as the underlying distributions of patient risk and of staff experience are equivalent at

Figure 1 **Average time of arrival to DU by hour of day (top) and average number of birth by day of the week (bottom) with 95% confidence intervals, broken down by patient type.**



different workload levels (i.e. there is no correlation between unobservable factors that might affect the rationing or referral decision and the workload measure) then this will not cause a bias to the estimated coefficients of interest. On the staffing side, the "establishment", which defines the number and experience level of the midwives who work a particular shift, is stable and is set weeks in advance of the shift to be worked, and so will be uncorrelated with the patient case-mix. On the patient side, arrivals to the DU by patients with spontaneous onset of labor occur at random, following a homogenous Poisson distribution (see §5 of the main paper). Moreover, elective C-sections are removed from the analysis, while Figure 1 shows that there is no pattern in the arrivals of induced patients (who are scheduled to arrive at the antenatal unit for the induction) to the DU across hours of the day or days of the week. This indicates that patient numbers will also be uncorrelated with the risk profile of the analysis sample.

Together the evidence presented here suggests that it is more appropriate to use a workload measure calculated at the unit level rather than the midwife level, since this better reflects the nature of work sharing in the DU, and will also not be affected by heterogeneity in the patient risk profile or differences in staff experience which may bias the estimated coefficients of interest.

2.2. Reverse causation and confounding factors

In the main paper we hypothesize and show that an increase in workload leads to a change in the rate of provision of epidural analgesia (rationing) and the rate of physician-led deliveries (referrals). One possibility that had to be accounted for was that the decision to provide an epidural or to refer a patient for a physician-led delivery might have itself brought about a change in workload. Under such a scenario it could have been argued that instead of workload affecting rationing or referral behavior, the change in behavior may have itself brought about the change in workload. This empirical phenomenon is known as reverse causation, and in such a situation the identified effects can be biased.

In the study context, a patient who receives an epidural is more likely to stay longer in the DU as the epidural slows the process of labor (see e.g. Anim-Somuah et al. 2011). On the other hand, a patient who receives a physician-led delivery has their labor prematurely shortened, resulting in them staying in the DU for a shorter period of time than if they had delivered naturally. Since the workload measure is calculated as a time-weighted average over the period from three hours prior until the time of birth, any change in the length of time that the patient spends in the DU may affect the measure of workload. To see this by way of example, suppose that there are 10 patients and 10 midwives present in the DU. Suppose one patient has just arrived in the DU, while the other nine have been in the DU for 3+ hours and will stay in the DU for another 3+ hours, with no other admissions or discharges from the DU in this period. Suppose the latest arrival will give birth in: (i) two hours without an epidural or physician-led delivery, (ii) three hours with an epidural, (iii) one hour with a physician-led delivery. Then the time-weighted workload measured over the three hour period leading up to the time of delivery would equal

$$\frac{(3\times 9)+(2\times 1)}{3\times 10}=0.967$$
$$\frac{(3\times 9)+(3\times 1)}{3\times 10}=1.000$$
$$\frac{(3\times 9)+(1\times 1)}{3\times 10}=0.933$$

for each case, respectively. Observe that the difference here is attributable solely to the length of time that the focal patient (the most recent arrival) stays in the DU prior to delivery. As can be seen, the increase in the length of labor and resultant increase in time spent in the DU caused by the epidural results in an increase in workload ($1.000 > 0.967$), while the shortening of labor and time in the DU caused by the physician-led delivery brings about a decrease in workload ($0.933 < 0.967$).

If reverse causality were to operate in the manner outlined above, as is expected in the DU setting under study, then workload would be higher when a higher rate of epidurals are provided, and

workload would be lower when patients are more likely to be referred for a physician-led delivery. This would cause the estimate of the workload effect in the epidural models to be biased upwards, and in the physician-led delivery models to be biased downwards. It is important to note that, if anything, this would make it *less likely* for the effects identified in the analysis to be observed, rather than causing these effects. In particular, the signs of the estimated coefficients reported in §6.2 of the main paper are in the opposite direction to those that would be expected if reverse causality were driving the results. Specifically, we identify a negative coefficient for the effect of workload on epidural analgesia, rather than positive, and a positive coefficient for the effect of workload on physician-led deliveries, rather than negative.

The reverse causation problem will exist for any measure of workload that is calculated over a time window which extends to a period sufficiently far prior to the time of birth so that the rationing or referral decision might affect the proportion of time during that window during which a patient was present in the DU. To tackle this, two approaches could have been taken: (1) Exclude the focal patient from the workload calculation; (2) Measure workload over a shorter time window during which it will be unaffected by the rationing/referral behavior of the midwife. To see why the former would work, observe that in the earlier example excluding the focal patient would have resulted in workload being equal to $\frac{3\times9}{3\times10} = 0.9$ in each of the three scenarios. In particular, reverse causality is not a concern when excluding the focal patient from the workload variable since any decisions relating to the care of the focal patient will have no affect on the workload measure. To see why the latter would work, note that if instead we were to measure workload over e.g. only the one hour period prior to the time of birth then the resultant workload would have been equal to $\frac{(1\times9)+(1\times1)}{1\times10} = 1.0$ in each case. In this example, since the patient would have stayed at least one hour regardless of the decisions made with respect to their care, there is no opportunity for reverse causality to arise. This would be the case for the majority of patients in the analysis sample, for which 99.9% stay in the DU for at least one hour.

In addition to the reverse causation problem, these approaches also help to overcome an issue arising from the fact that patients who spend longer in the DU, potentially impacting on workload, may also be more or less likely to receive an epidural or be referred for a physician-led delivery. In this case the length of stay (LOS) in the DU may have been a confounding factor driving a spurious relationship between workload and the rationing and/or referral decisions. For epidural analgesia, clinical guidelines indicate that patients with prolonged labor (and hence who are more likely to stay longer in the DU) will be more likely to receive this form of pain relief (NHS 2015). Omitting DU LOS as a control variable could therefore create a spurious positive relationship between workload and the rate of epidural analgesia, further confounding the bias caused by reverse causation. For patient referrals to physicians, evidence indicates that patients who have a longer

Table 1 Average partial effects with and without correction of workload for endogeneity bias.

	Epidural		Phys.	
Model type	Probit	Probit	BiProbit	Probit
Complexity	*nC*	*C*	*nC*	*C*
Original workload	-0.025***	-0.006	-0.000	0.015*
Workload incl. focal (3h)	-0.011*	0.006	0.001	0.016**
Workload incl. focal (1h)	-0.018***	0.003	-0.000	0.015*

nC and *C* refer to non-complex and complex patient episodes, respectively; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

length of labor will be more likely to be referred, since this is an indication that natural labor is not progressing sufficiently (NICE 2012). In this case a positive relationship between workload and the referral decision may arise if DU LOS is not controlled for, potentially counteracting the effect caused by reverse causation. Excluding the focal patient from the workload variable or shortening the time window over which workload is calculated, however, ensures that workload is uncorrelated with the length of time that the focal patient spends in the DU, and so removes any potential for such spurious relationships.

In the main paper the approach that was taken was to exclude the focal patient from the workload calculation. This choice was made since in order to identify the causal effect of workload on the epidural decision it is important for the measure of workload to overlap with the time period during which the epidural decision is most likely to be taken. Since epidural analgesia is rarely administered close to the time of birth, when it is often contraindicated, a sufficiently long window over which to perform the time-weighting is required. To see the potential effect of reverse causality it is helpful, though, to re-run the analysis from the paper under two further scenarios. In both of these the focal patient is included in the workload calculation, but the first is measured over the same three-hour time window used in the main paper while the second is calculated over a shorter, one-hour time window. Based on the discussion above, it should be expected that reverse causality will be more of a concern for the former than the latter. Reported in Table 1 are the estimated average partial (marginal) effects (APEs) under these two scenarios, along with the original effects observed in the main paper.

Inspecting Table 1 we see evidence of potential reverse causation: When including the focal patient in the estimation for the epidural analgesia decision for the non-complex patient episodes (PEs), the coefficient estimate is biased upwards, as anticipated, in value from -0.025 when excluding the focal patient (or -0.018 when including the focal patient but averaging over the shorter time window of one hour) to the higher value of -0.011. There is no effect, however, when including the focal patient in the workload measure when estimating the physician-led delivery rate for the complex PEs (all APEs approx. 0.014). This may be because by not controlling for the confounding

factor, DU LOS, this serves to counteract the reverse causality issue: patients who stay longer in the DU are inherently more likely to receive a physician-led delivery due to the fact that a longer stay indicates that labor may not be progressing appropriately, with the increased stay contributing to an increase in workload.

Based on the discussion and results presented above, we believe that it is necessary to address the problems of reverse causality and omitted variable bias caused by not controlling for DU LOS, a potential confounding factor. The most appropriate way of achieving this was through the exclusion of the focal patient from the workload calculation. Other methods exist, such as shortening the time window over which workload is calculated, and these produce consistent results (within 25% of the original estimates). Even when the reverse causality problem is not addressed, the headline results reported in the main paper hold up in terms of direction, though the coefficient estimates are biased and underestimate the true effects.

2.3. Excluding the focal patient with variable staffing

An unintended consequence of excluding the focal patient from the calculation of workload is that at different points in time the system may be equally loaded and yet it is possible for the values of workload to differ. This problem arises because the number of midwives present in the DU is non-constant. To see this, consider a system that is "perfectly staffed", in the sense that there is a 1:1 ratio between the number of patients and the number of midwives. If there are 10 patients and 10 midwives present at time t, then assuming that focal patient i is in the DU the (instantaneous) workload for patient i, $LOAD_i(t)$, will be equal to $(10-1)/10 = 0.9$. On the other hand, if there were 5 patients and 5 midwives then workload instead would equal $(5-1)/5 = 0.8$. More generally, if the ratio of patients to midwives at times $t = 1, 2$ are equivalent and given by p, but the number of midwives present at times $t = 1$ and $t = 2$ is given by m_1 and m_2, respectively, then instantaneous workload at time $t = 1$ will be equal to $\frac{p*m_1-1}{m_1}$ and at time $t = 2$ will equal $\frac{p*m_2-1}{m_2}$. There will therefore be a difference in workload of $\frac{1}{m_1} - \frac{1}{m_2}$, despite the fact that the systems are equally loaded (i.e. have the same proportion of patients to midwives, p).

We argue in this section that so long as the workload variable used in the main paper (workload excluding the focal patient) is highly correlated and similar in scale to the workload variable of interest (an unbiased but (partially) unobserved estimate of workload including the focal patient) then the issue highlighted above will have almost no effect on the results. We estimate the correlation to be > 0.99 and the scales to be within $\pm 2.5\%$, indicating that the estimated coefficients will be almost identical to the true coefficients of interest.

To present the argument above in a technically rigorous way, recall that $N_i(t)$ gives the number

of patients besides focal patient i in the DU at time t and $MW(t)$ the number of midwives. The instantaneous workload at time t was then expressed as

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

with

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

the time-weighted average load for a patient i who gives birth at time b_i, 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. Instantaneous workload *including* the focal patient at time t can therefore be expressed as

$$LOAD_i^+(t) = \frac{N_i(t) + \mathbb{1}[i \in P(t)]}{MW(t)} = LOAD_i(t) + \frac{\mathbb{1}[i \in P(t)]}{MW(t)}\,,$$

where $P(t)$ is the set of patients in the DU at time t so that $\mathbb{1}[i \in P(t)]$ is equal to one if patient i is in the DU at time t and zero otherwise. The corresponding time-weighted average load for patient i is equal to

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

where $L^+(\underline{b}_i, b_i)$ is the set of all observed values of $LOAD_i^+(t)$ between $t = \underline{b}_i$ and $t = b_i$. This can be re-expressed as

$$LOAD_i^+ = LOAD_i + \sum_{k \in M(\underline{b}_i, b_i)} \frac{k}{b_i - \underline{b}_i} \int_{\underline{b}_i}^{b_i} \mathbb{1}\left[\frac{\mathbb{1}[i \in P(t)]}{MW(t)} = k\right] dt = LOAD_i + x_i\,,$$

where $M(\underline{b}_i, b_i)$ is the set of all observed values of $\frac{\mathbb{1}[i \in P(t)]}{MW(t)}$ between $t = \underline{b}_i$ and $t = b_i$. Note that x_i is independent of the number of other patients in the unit and only depends on the number of midwives.

Now consider the simple linear regression equations expressed below. In (1) workload includes the focal patient, while in (2) the focal patient is removed:

$$EPI_i = \alpha_0 + \alpha_1 LOAD_i^+ + u_1 \tag{1}$$

$$\begin{aligned} EPI_i &= \beta_0 + \beta_1 LOAD_i + u_2 \\ &= (\beta_0 - \beta_1 x_i) + \beta_1 LOAD_i^+ + u_2\,. \end{aligned} \tag{2}$$

Assume for now that endogeneity/reverse causality is not a concern and that the true effect of workload on the epidural rate is given by α_1, to be estimated in (1). Observe that if x_i is equal to some constant, C say, for all i then we have that $\alpha_0 = \beta_0 - \beta_1 C$ and $\alpha_1 = \beta_1$. In other words,

the difference between the workload variable that arises due to the removal of the focal patient is absorbed in to the intercept term, leaving the estimated slope coefficient unaffected. On the other hand, if x_i is non-constant then $cor\left(LOAD_i, LOAD_i^+\right) \neq 1$ and $SD\left(LOAD_i\right) \neq SD\left(LOAD_i^+\right)$, where *cor* denotes Pearson's correlation coefficient and SD the standard deviation. This will result in slope estimates $\alpha_1 \neq \beta_1$. If, though, (i) $cor\left(LOAD_i, LOAD_i^+\right) \approx 1$ and (ii) $SD\left(LOAD_i\right) \approx SD\left(LOAD_i^+\right)$ then it turns out that the coefficient estimates α_1 and β_1 will be very similar in value. To see why this is the case for simple linear regression, note that the slope terms in equations (1) and (2) can be expressed as $\alpha_1 = cor\left(EPI_i, LOAD_i^+\right) \cdot \frac{SD(EPI_i)}{SD(LOAD_i^+)}$ and $\beta_1 = cor\left(EPI_i, LOAD_i\right) \cdot \frac{SD(EPI_i)}{SD(LOAD_i)}$, respectively (e.g. Duncan 1975, p. 11). Then if conditions (i) and (ii) given above are satisfied it is easy to see that $\alpha_1 \approx \beta_1$. This finding can be extended, approximately, to the multiple linear regression case and similarly to binary choice models such as logit and probit (see e.g. Kim and Ferree 1981, Winship and Mare 1984).

In the above we assumed endogeneity/reverse causality not to be a concern, meaning that α_1 was assumed to be unbiased. Recall, though, that in the paper we *exclude* the focal patient from the workload measure specifically because if they were to be *included*, as in (1), then this will result in some estimate for $\alpha_1^B \neq \alpha_1$, where α_1^B is the biased and α_1 the true, unobserved, effect. To identify α_1 precisely we would need to calculate $LOAD_i^+$ in (1) using unobserved information: specifically, the value of $\mathbb{1}[i \in P(t)]$ at every t and for every i had all patients received identical treatment (to see why this is the case see §2.2). Without this information – and without an appropriate instrumental variable to perform the analysis using suitable sample selection methods – then based on the discussion above it is possible instead to perform this estimation using some other *unbiased* variable if it is *highly correlated* with and *similar in scale* to the variable of interest.

To see that this is the case for the load measure (calculated *excluding* the focal patient) used in the paper, $LOAD_i$, suppose that the actual variable of interest is $LOAD_i^A$, the (partially) unobserved workload level if the focal patient neither received an epidural or a physician-led delivery.[1] Then we want to show that $cor(LOAD_i, LOAD_i^A) \approx 1$ and $SD(LOAD_i) \approx SD(LOAD_i^A)$. To do this, observe that the workload variable we estimate when *including* the focal patient is $LOAD_i^+ = LOAD_i^A + e_i\mathbb{1}[EPI = 1, PHYS = 0] + p_i\mathbb{1}[EPI = 0, PHYS = 1] + b_i\mathbb{1}[EPI = 1, PHYS = 1]$. Here $\mathbb{1}[\cdot]$ is the indicator function, making e_i the (observed or unobserved) change in the workload variable if the focal patient only received an epidural, p_i the change if the patient received only a physician-led delivery, and b_i the change if the patient received both an epidural and physician-led delivery.

[1] Note that the discussion in this paragraph extends to the standardized workload measure used in the main paper since $\mu(\cdot)$ and $\sigma(\cdot)$, the values of the mean and standard deviation used in the calculation of standardized workload, are unaffected by whether or not the focal patient is included in the calculation of $LOAD_i$.

Then note that $LOAD_i^A = LOAD_i^+$ when the patient does not receive an epidural or a physician-led delivery. If we can show, therefore, that for this subset of patients $cor(LOAD_i, LOAD_i^+) = cor(LOAD_i, LOAD_i^A) \approx 1$ and $\frac{SD(LOAD_i)}{SD(LOAD_i^+)} = \frac{SD(LOAD_i)}{SD(LOAD_i^A)} \approx 1$, then this would suggest that the unbiased estimator $LOAD_i$ is highly correlated with and similar in scale to the variable of interest, $LOAD_i^A$. For our data we calculate a correlation of 0.994 and the ratio of standard deviations to be 0.977. Based on these results, we would thus expect any effect on the estimated coefficient of workload caused by excluding the focal patient to be very small.

In addition to the above, to investigate this further we have re-run the analysis from the paper using a subset of the data for patients who faced similar staffing levels. When the number of midwives is constant or near-constant then x_i in (2) will be near-constant also, meaning that the effect of excluding the focal patient will be almost entirely absorbed in the intercept term rather than affecting the estimate of the slope, as described earlier.Therefore, we subset our data to include all patients for who the average number of midwives present in the three hour window prior to the time of birth takes values between 7 and 9, resulting in a sample of 8288 non-complex (82.1% of final sample) and 5205 complex (83.1% of final sample) PEs. In this case, the maximum difference between the workload of two equally loaded systems that is caused by removing the focal patient is equal to $\frac{1}{m_1} - \frac{1}{m_2} = \frac{1}{7} - \frac{1}{9} = 0.032$. Note that this effect is small when compared to the standard deviation of (non-standardized) workload for our sample, which equals 0.375. As such, any impact of removing the focal patient on the results calculated using this reduced sample will be negligible.

Based on the discussion above, if the results are significantly affected by the exclusion of the focal patient from the workload specification then it should be expected that the coefficients estimated using this reduced sample will be significantly different from those calculated using the original sample. Results are presented in Table 2. As is shown in Table 2, reducing the sample to include only those patients for who midwife numbers are similar has almost no impact on the results. Specifically, the APE of workload on the epidural rate for the non-complex PEs decreases from -0.025 to -0.026, while the APE of workload on the rate of physician referrals for the complex PEs stays the same at -0.014. Therefore, putting all of the evidence presented in this section together, it can be concluded that the results are highly insensitive to the exclusion of the focal patient from the workload measure.

2.4. Workload specification

Thus far we have shown that the way in which workload is measured overcomes potential endogeneity issues and is robust to the removal of the focal patient. This subsection continues the discussion on the robustness of the results with respect to workload by reproducing the results from the main paper under alternative workload specifications. These including allowing workload to enter the models in non-linear form, measuring workload over different time-windows, and using alternative time frames over which to standardize workload.

Table 2 Average partial effects when constraining variability in midwife numbers.

	Epidural		Phys.	
Model type	Probit	Probit	BiProbit	Probit
Complexity	*nC*	*C*	*nC*	*C*
Original sample ($n = 16{,}355$)	-0.025***	-0.006	-0.000	0.014*
Reduced sample ($n = 13{,}493$)	-0.026***	-0.008	0.006	0.014*

nC and *C* refer to non-complex and complex patient episodes, respectively; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

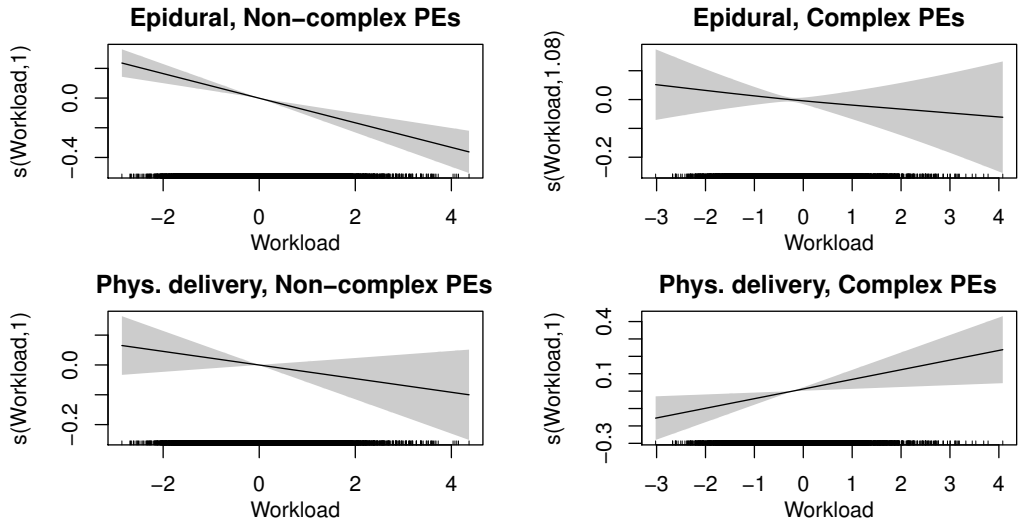
2.4.1. Non-linear workload Other research has shown that the effect of workload in service settings such as hospitals and restaurants may be non-linear in the tails (see e.g., Kuntz et al. 2014, Tan and Netessine 2014). This effect is often attributed to differences in worker behavior when faced with different levels of workload, e.g. a reasonable level of busyness may be required to increase worker motivation, beyond which competing effects such as fatigue and stress may lead to a deterioration in quality. Here we set out to determine whether there is evidence of the presence of a non-linear effect of workload on midwives' rationing and referral behavior in the DU setting.

There are a number of methods that could be used to investigate non-linearity, such as dividing the workload variable into a number of binary variables with cuts made at arbitrary percentiles or fitting a piecewise-linear (segmented) regression model with knots (or breakpoints) as additional parameters for estimation. Since our dataset is relatively small, however, we eschew these methods in favour of an approach using generalized additive models (GAMs). These models are similar to the standard OLS or probit specification, except that the dependent variable is allowed to depend on some (unknown) smooth function of one or more of the predictors. This allows us to estimate a non-linear effect of workload on each of the outcome equations by applying a smoothing function to this variable during estimation. To avoid going into the technical details, we mention only that these models are estimated using a local scoring algorithm by iteratively fitting weighted additive models by backfitting using the Gauss–Seidel method (Wood 2011). This is implemented by the `gam` function in the `mgcv` package within the statistical software R version 3.2.2 (Wood 2013).

The plots in Figure 2 show the component smooth function from the fitted GAMs on the scale of the linear predictor, with two standard error bounds (approx. 95% confidence intervals) indicated by the shaded region. The estimated degrees of freedom (EDF), reported in the vertical-axis labels, approximately correspond to the highest order of a polynomial transformation of the workload variable that would need to be included in the standard OLS/Probit model in order to capture the non-linear relationship. The EDFs are almost all equal to or close to 1, suggesting that there is little evidence of deviation from linearity.

Despite a lack of evidence of deviation from linearity, it can be seen in the plots that the confidence intervals at more extreme levels of workload are reasonably wide. It is feasible, therefore,

Figure 2 Examining non-linear workload effects.



that the true effect(s) may be non-linear in the tails but that there is insufficient power in the analysis from the data to identify this. Increasing the power with more data might help to (partially) resolve this issue, but in general since we concentrate on reporting the results in the main paper at or about the mean values of workload the main findings will be relatively insensitive to small deviations from non-linearity. In particular, results are compared under low- and high-workload scenarios which fall ± 2 standard deviations about the mean (or between the values -1.95 and 1.78) where it can be seen in Figure 2 that the confidence intervals are relatively narrow. While one should expect that this linear relationship may begin to break at some extreme values of workload, and it would be interesting to identify at what point this begins to take effect, this is not the main focus of this paper and it does not invalidate the main findings presented in the paper.

2.4.2. Alternative workload scenarios In the main paper workload was measured as the time-weighted load calculated over the three hour period prior to the time of birth. As mentioned in §5.1 of the paper, a three hour window was a somewhat arbitrary decision that was chosen because it was aligned with the length of the second stage of labor, the period of active labor when interventions are most likely. Other time windows could, though, have been used. Measuring workload closer to the time of birth, for example, will increase proximity to the time at which the decision of whether to refer a patient for a physician-led delivery is made. At the same time, however, it will then not overlap with the time at which the epidural is most likely to occur. On the other hand, measuring over a period further from the time of birth will capture more epidurals that occur outside of the three-hour window we select, but then will be capture less of the variation that occurs at the time at which the referral decision is made. With no approach being perfect,

Table 3 Average partial effects under using alternative workload averaging windows.

	Epidural		Phys.	
Model type	Probit	Probit	BiProbit	Probit
Complexity	*nC*	*C*	*nC*	*C*
Workload, 60 minutes	-0.022***	-0.001	0.001	0.015*
Workload, 120 minutes	-0.023***	-0.003	0.000	0.014*
Workload, 180 minutes (original)	-0.025***	-0.006	-0.000	0.015*
Workload, 240 minutes	-0.027***	-0.008	-0.001	0.014*

nC and *C* refer to non-complex and complex patient episodes, respectively;
$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$

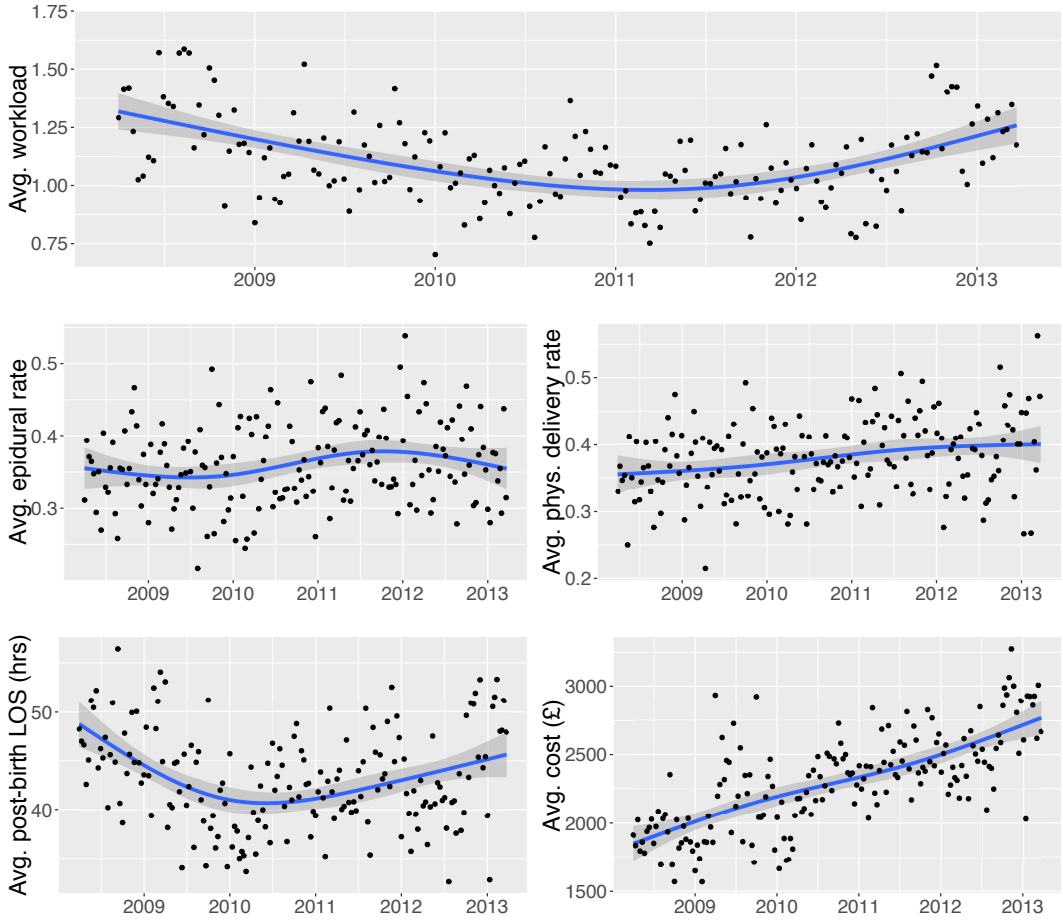
for robustness the analysis has been re-run over three other time windows. These range from one, two and four hours prior to the time of birth up until the time of birth. The resultant coefficient estimates from theses alternative workload scenarios are reported for the rationing and referral decisions in Table 3, alongside the original workload estimates.

As is shown in Table 3, regardless of the time window over which workload is measured the results are entirely consistent with those produced using the three hour averaging window selected for use in the main paper. The APE of workload on the rate of epidural analgesia for the non-complex PEs, given in column 2, ranges from -2.2% to -2.7% for every one standard deviation increase in workload. For the complex PEs, the APE of workload on the rate of referral for physician-led delivery ranges from 1.4% to 1.5% for every one standard deviation increase, as shown in column 4.

2.4.3. Standardization of workload One issue when working with data measured over time, especially a long period such as the five year duration of this study, is that the variables of interest may not be time invariant. If this is the case, then any observed relationship between the dependent and independent variables of interest may be due to them being correlated with time, rather than there being any direct causal relationship, i.e. time may be a confounding factor. To see this, assume that variables Y and X are independent. Now suppose that instead of observing Y and X directly we instead record some Y_t and X_t which evolve over time T in a linear fashion, with $Y_t = Y + aT$ and $X_t = X + bT$, where a, b are multipliers and X, Y are independent of T. Then clearly this introduces a time-related dependence between Y_t and X_t, with $cor(Y_t, X_t) \neq 0$. In a simple linear regression model where the estimated effect of independent variable X on dependent variable Y can be expressed as $cor(Y, X) \cdot \frac{SD(Y)}{SD(X)} = 0$, for example. But neglecting to account for the common response variable, time, can lead to a spurious relationship between Y and X being inferred since that relationship is instead estimated by $cor(Y_t, X_t) \cdot \frac{SD(Y_t)}{SD(X_t)} \neq 0$.

In Figure 3 are plotted mean values for workload, the epidural rate, the rate of referrals for physician-led deliveries, the post-birth LOS and cost. Each are calculated over distinct time windows spanning 10 days. To these have been added smoothing estimators of the relationship between

Figure 3 Time dependence of primary covariates.



each of these variables and time – fitted using natural cubic spines with three degrees of freedom – together with 95% confidence bands. There is evidence of non-randomness in these plot, an indication that over the five year time window there have been gradual changes in the scale or rate of each of these variables. This is not surprising, for example it should be expected that cost will increase over time, at least in line with inflation. Examining the plot for the primary independent variable of interest, workload, it can be seen that, approximately, this decreases slowly for the first two years, before stabilizing for a further two years, and then increasing again in the final year. Given that there is a clear pattern to workload over time, it is therefore important that steps were taken to ensure that the relationships to be identified in the main paper were not spurious and driven by time dependence.

There are two ways in which the time dependence highlighted above can be counteracted. The first is to include control variables in the models that sufficiently account for changes in the dependent and independent variables over time. This method relies on the relationship between the

covariates and time being accurately captured. To see how this would work, suppose that the time dependence is linear as in the example above, and that the relationship of interest is the one between X and Y. Assume that instead of observing X and Y directly, however, instead we observe X_t, Y_t and T, as defined before. Then to recover an unbiased estimate of the coefficient α_1 in the model

$$Y = \alpha_0 + \alpha_1 X + u_1 \,,$$

it is possible to instead estimate this by

$$\begin{aligned} & Y_t = \beta_0 + \beta_1 X_t + \beta_2 T + u_2 \\ \iff \quad & Y + aT = \beta_0 + \beta_1 (X + bT) + \beta_2 T + u_2 \\ \iff \quad & Y = \beta_0 + \beta_1 X + (\beta_2 + \beta_1 b - a) T + u_2 \,. \end{aligned}$$

Since Y is assumed to be independent of T it follows that the sum of the square errors will be minimized when $\beta_2 = a - \beta_1 b$, and so it can be shown that β_1 is an unbiased estimate for the coefficient α_1. Similarly, higher order polynomial could be used to capture a non-linear change in the dependent or independent variables over time. If the time dependence can not be sufficiently captured with a trend/polynomial trend term(s), which may e.g. occur if the covariates evolve in a highly unstable way, then it is possible instead to include fixed effects that act to alter the intercept for the dependent variable at the specified points in time.

Another method that can be used to break the time dependence between a dependent and an independent variable is to de-trend either or both of the covariates. To see this, suppose that bT is subtracted from $X_t = X + bT$. Then $X_t = X$ and $cor(Y_t, X_t) = cor(Y + aT, X) = cor(Y, X) + cor(aT, X) = 0$ when X, Y and T are independent. Returning to the simple linear regression example, this would mean that the estimate of the coefficient would be equal to $cor(Y_t, X_t) \cdot \frac{SD(Y_t)}{SD(X_t)} = 0$, i.e. the true null relationship between the two variables would be identified.

In the main paper both of the approaches described above are employed to ensure that the relationships identified between workload and the dependent variables of interest are not spuriously driven by time invariance. Specifically, a daily trend term, along with its square, together with year-quarter dummies that pick up shocks or larger deviations not accounted for by the trend terms are included as controls. In addition, the workload variable it standardized by subtracting the average workload observed over a one year period centered at the time of birth of the focal patient and dividing through by the standard deviation of workload calculated over that period (for further detail see the discussion on the calculation of the independent variable in §5.1 of the main paper). The standardization serves two purposes: (1) Firstly, it makes interpretation of the results more straightforward, meaning that the reported APE can be interpreted as the change

Table 4 Average partial effects using alternative standardization periods.

	Epidural		Phys.	
Model type	Probit	Probit	BiProbit	Probit
Complexity	*nC*	*C*	*nC*	*C*
Unstandardized workload	-0.062***	-0.010	-0.002	0.039*
1-year de-meaned workload	-0.061***	-0.009	-0.002	0.039*
90-day standardized workload	-0.024***	-0.007	0.001	0.014*
1-year standardized workload (original)	-0.025***	-0.006	-0.000	0.015*

nC and *C* refer to non-complex and complex patient episodes, respectively; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

in the dependent variable that results from a one standard deviation increase in workload, where the standard deviation of workload is updated over time to account for changes in it's value; (2) Secondly, it serves to smooth the measure of workload prior to estimation, making it easier for time controls to capture any residual differences over times.

To ensure that the results are not sensitive to the process of standardization, however, the analysis from the main paper has been re-run under three other workload specifications, one in which no standardization is performed, one where the series is de-meaned but not standardized by dividing through by the standard deviation, and one where the series is standardized instead over a 90 day window. The results are reported in Table 4.

When examining Table 4 is is important to note that the APEs reported in rows 1 and 2, which correspond to workload which is unstandardized and workload which is de-meaned only, are not comparable to the APEs reported in rows 3 and 4, since the latter have been scaled by dividing through by rolling estimates of the standard deviation. Comparing unstandardized and de-meaned workload it can be seen that the de-meaning process has little effect on the coefficient estimates (both APE approx. -0.062 for epidural rate of non-complex PEs, both APE $= 0.039$ for referral rate of complex PEs). This suggests that the regressors in the models that account for time are controlling sufficiently for any time dependence. Similarly, the results in Table 4 show that standardizing over a 90 day window or a 1-year window has little impact on the results.

Based on the above, in the main paper the decision was taken to standardize workload over a 1-year time window so as to improve interpretability of results. The 1-year window was preferred over shorter time frames since it does not smooth potential differences across different seasons of the year (e.g. summer vs. winter), as all 12 months of a year are included when performing the standardization. As such, the resultant workload measure corrects only for slow moving changes in workload over time. It is important to note, however, that the results are insensitive to different windows of standardization and hold up even when standardization is not performed.

3. Robustness to Alternative Model types

In this section we confirm the robustness of the results from the main paper by comparing them against results obtained under alternative model specifications/forms. In 3.1 we discuss how complexity is operationalized in more detail and show that the results hold up under an expanded definition of complexity that includes other factors that might materially complicate the labor and delivery process. In 3.2 we introduce variables that control for differences in the rationing and referral rates across different midwives. In 3.3 we explain that workload on the antenatal unit is not correlated with DU workload and show that the results are not affected by it's inclusion as an additional regressor. Finally, in 3.4 we discuss why the decision was taken to run the regressions separately on the two samples of non-complex and complex PEs, and show that commensurate findings would have been obtained had we instead used a single model that included an interaction between workload the complexity of the PE.

3.1. Measuring complexity

In the main paper it is hypothesized that the effect of workload on rationing and referral decisions is moderated by, following Shumsky and Pinker (2003), the complexity of the problem/service needs posed by the customer to the server. In the context of the DU, the 'problem' that the midwife must manage is the process of labor and the delivery of the baby. As such, in differentiating between non-complex and complex observations in the DU context it was necessary to identify ex-ante observable factors that contribute towards making the service episode, i.e. the labor and delivery process, more challenging for the midwife to manage. Since it is the complexity of the patient episode that is being studied, the induced and non-induced sub-samples are an excellent fit for this purpose, and far more suitable than a measure based on the risk profile of the mother. In particular, as described in the paper:

> "Women with spontaneous onset of labor tend to have less complex deliveries 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 readily observable by the midwife, and inductions are sufficiently frequent to provide the requisite statistical power."

Specifically, inductions change the birth process itself and, through this, serve to increase the complexity of the service episode (in this case the complexity of labor and the delivery).

It is important to recognize why a measure of complexity defined using other factors, such as those known to be associated with patient risk, would not be appropriate. There are, for example, maternal factors and comorbidities that are associated with greater risk during pregnancy, such as first pregnancy over the age of 35, gestational diabetes, eclampsia, etc. Yet while it is certainly the case that risk factors must be managed appropriately during the pregnancy, they do not materially affect the labor and delivery process itself. In particular, the mechanics of labor are unlikely to be materially different for two otherwise identical women if, say, one has gestational diabetes and the other does not. It is certainly the case that such risk factors must be managed appropriately during the pregnancy. However, on arrival in the DU their is no reason to expect that patients with a higher *risk profile* will be treated any differently or be more or less susceptible to workload than others since their *service needs*, as they relate to labor and the delivery process, will be essentially identical. This is not the case, though, for induced patients who, as described above, experience a more complex process of labor than their non-induced counterparts.

While maternal risk factors do not capture 'problem' complexity in this setting, it is worth considering whether there are any factors related to the pregnancy that can be used to add additional patients into the subsample of complex PEs. Doing so serves two purposes: First, it addresses the concern that the differential effect of workload on the non-complex and complex PEs may be more to do with the induction process itself, rather than complexity; Second, it increases the power of the subsample of complex PEs and so makes the empirical identification of an effect more likely, if one exists. Factors that have a material impact on the complexity of the labor and delivery process from the midwife's perspective are rare and occur with low frequency, but two others that have been identified are:

1. Breech presentation of the baby, and
2. A multiple birth.

A breech occurs when the baby exits the pelvis with buttocks or feet first, rather than with the normal head first presentation. This occurs at rate of 2.15% in the final analysis sample of 16,355 patients. A multiple birth occurs when the mother delivers two or more babies, and occurs at rate of 1.33% in the final analysis sample. For these complications of pregnancy, both the labor *and* delivery process/problem become significantly more complex for a midwife to manage (Ball and Washbrook 1996). Therefore, as was the case for induced patients, there should be expected to be a difference in the rationing (of epidural analgesia) and referral (for a physician-led delivery) behavior of the midwife for these patients, relative to their non-complex counterparts. Specifically, there will be less discretion in the provision of epidural analgesia to a breech or multiple birth patient, and

Table 5 Average partial effects using different measures of complexity.

	Epidural		Phys.	
Model type	Probit	Probit	BiProbit	Probit
Complexity	*nC*	*C*	*nC*	*C*
Original sample	-0.025***	-0.006	-0.000	0.015*
Expanded complex subsample	-0.027***	-0.004	-0.000	0.014*

nC and *C* refer to non-complex and complex patient episodes, respectively; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

so the decision should be expected to be unaffected by midwife workload. In addition, owing to the increased demands they place on a midwife and the greater their likelihood of eventually ending up requiring an physician-led delivery, these patients are expected to be referred at a higher rate when midwife workload increases. Results using this expanded sample of complex (and reduced sample of non-complex) PEs are presented in Table 5, together with results using the original definition of complexity used in the main paper.

As can be seen in Table 5, expanding the definition of complexity has almost no impact on the results. Given this, and the additional insight provided earlier as to how the definition of complexity in this context relates to the complexity of the labor/delivery process rather than the risk profile of the patient, it follows that operationalizing complexity by way of the distinction between induced and non-induced patients enables us to effectively capture how complex a service episode was from the midwife's perspective. While the subsample of complex PEs could be expanded to include breech and multiple births, this would have a negligible impact on the results.

3.2. Accounting for midwife effects

It is clear that different midwives will have different skill sets and that these might impact on the rate at which particular midwives ration the provision of epidural analgesia or refer patients for a physician-led delivery. If in addition midwives work regular shift patterns and workload during particular shifts is, on average, slightly higher or lower than the across shift mean, then one could argue that factors related to the midwife might act as correlated omitted variables (i.e. correlated with both the dependent variables of interest and the independent variable of interest, workload, and so bias the coefficient estimate of workload). In §2.1 of this document the fact that midwife working schedules are established weeks in advance of the shifts to be worked is discussed, making it unlikely that there should be any such bias. However, since information is available on the midwifes present in the DU, it is possible to confirm this by introducing fixed effects for each midwife as control variables in the models.

One problem with including midwife fixed effects is that there are a large number of midwives who delivered only a small number of babies during the sample period. This may be because they were e.g., bank staff, agency staff (i.e., temporary staff who work for an agency and are brought

Table 6 Average Partial Effects Controlling for Midwife Effects.

	Probit				BiProbit	
Complexity	(1) Epidural *nC*	(2) Phys. *nC*	(3) Epidural *C*	(4) Phys. *C*	(1) Epidural *nC*	(2) Phys. *nC*
Std. workload	-0.020***	-0.005	-0.002	0.014*	-0.018***	-0.001
	(0.005)	(0.004)	(0.007)	(0.006)	(0.004)	(0.004)
Epidural	–	0.205***	–	0.207***	–	0.196***
	–	(0.010)	–	(0.011)	–	(0.010)
Dist. to home	-0.017***	-0.007	0.004	-0.003	-0.017***	–
	(0.005)	(0.004)	(0.007)	(0.006)	(0.004)	–
2–4h op. tht. use	-0.035*	0.010	0.031	-0.022	-0.029†	–
	(0.017)	(0.016)	(0.025)	(0.020)	(0.015)	–
Inst. op. tht. use	-0.004	-0.087***	-0.020	-0.092***	-0.003	-0.085***
	(0.009)	(0.009)	(0.013)	(0.012)	(0.009)	(0.009)
N	10,091	10,091	6,264	6,264	10,091	
Log-lik	-5,213.91	-4,493.87	-3,779.73	-2,845.59	-9,701.88	
Pseudo-R^2	0.115	0.323	0.129	0.323	–	
ρ	–	–	–	–	-0.431*** (0.094)	

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, 1,000 simulations; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

in to fill vacancies at short notice, often at great expense), midwives on a temporary trial, student midwives, recent hires, etc. Therefore, in order to add fixed midwife effects the 325 unique midwives that are observed must be reduced to a smaller subset of those who performed deliveries in the DU most frequently. A cutoff of 164 deliveries (1% of the sample) is used and an "Other" category introduced in which all of those midwives who delivered fewer than 164 babies are collected. To augment this we introduce dummy variables that capture (a) the grade band of the midwives, (b) the age of the midwives, and (c) the experience of the midwife, as measured by the amount of time that they have been a registered midwife. Note that these covariates are not included in the models reported in the paper as the effects are, for the most part, insignificant or only weakly significant.

Updated results for the effects of workload on the rationing (epidural) and referral (physician-led delivery) decisions are presented in Table 6. Probits (1)-(2) and BiProbits (1)-(2) give the APE of workload on the epidural and physician-led delivery rate for the non-complex PEs, after controlling for midwife fixed effects. Probits (3)-(4) gives the equivalent for the complex PEs. Since the results do not change in any material way, in the main paper we opted to use the simpler models.

3.3. Controlling for antenatal unit workload

In the main paper the measure of workload that is used accounts for the number of patients in the DU relative to the number of midwives available. It is plausible that workload in other units of the hospital – particular the number of patients present in the antenatal unit and waiting in the waiting lounge – might also have an impact on rationing and referral behavior. Batt and Terwiesch (2015) show, for example, that service times in an emergency department are affected by congestion in

Table 7 Average Partial Effects Controlling for Antenatal Unit Occupancy.

	Probit			BiProbit	
Complexity	(1) Epidural *nC*	(2) Epidural *C*	(3) Phys. *C*	(1) Epidural *nC*	(2) Phys. *nC*
Std. workload	-0.026***	-0.006	0.013*	-0.023***	-0.001
	(0.005)	(0.007)	(0.006)	(0.004)	(0.005)
Antenatal occupancy	0.006	0.001	0.007	0.005	0.006
	(0.004)	(0.006)	(0.005)	(0.004)	(0.004)
Dist. to home	-0.017***	0.004	-0.002	-0.017***	–
	(0.005)	(0.007)	(0.006)	(0.004)	–
2–4h op. tht. use	-0.035*	0.025	-0.019	-0.028†	–
	(0.018)	(0.025)	(0.021)	(0.015)	–
Inst. op. tht. use	-0.006	-0.017	-0.098***	-0.005	-0.087***
	(0.009)	(0.014)	(0.012)	(0.008)	(0.009)
N	10,091	6,264	6,264	10,091	
Log-lik	-5,317.82	-3,835.31	-2,927.05	-9,884.88	
Pseudo-R^2	0.097	0.117	0.304	–	
ρ	–	–	–	-0.448*** (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, 1,000 simulations; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

the waiting room. In the DU setting it is possible that if staff know that there are many expectant mothers building up in other units – who will be coming to the DU soon – then they may "feel the pressure". In particular, they may elect not to engage in high resource activities recognizing that it is possible that those resources will be needed by other patients very soon. If occupancy in the DU and on the antenatal unit are correlated, also, then omitting to control for antenatal unit workload might bias the estimated coefficient for DU workload.

In order to test whether there is any evidence of an effect as described above, a variable that captures the effect of antenatal workload is introduced into the model specification. The results are presented in Table 7. It is important to note that as information is not available on the number of staff members in the antenatal unit it is necessary to instead use occupancy, as measured by the number of patients present in the antenatal unit, as a proxy for workload, recognizing that this does not also account for staff (resource) availability. Other than the lack of an adjustment for staffing, antenatal workload is measured in the same way as for DU workload, taking the time-weighted occupancy (excluding the focal patient) over the three hour period prior to the time of delivery by the focal patient, and de-trending the series by subtracting by the 1-year rolling mean antenatal unit occupancy and dividing through by the standard deviation.

The results reported in Table 7 show that antenatal unit occupancy is not statistically significant (even at the 10% level) in all of the model estimations, and that DU workload (Std. workload) does not change significantly with the inclusion of this additional regressor, changing from the original value of -0.025 to that of -0.026 in Probit (1) and from 0.014 to 0.013 in Probit (3), while remaining insignificant in Probit (2) and BiProbit (2).

It is likely that the reason for the lack of an effect of antenatal unit workload is that the majority of the arrivals to the DU occur randomly from the community when labor starts spontaneously. For the remainder, it is very difficult to predict how long it will take for a patient induced in the antenatal unit to begin labor and be transferred to the DU – Figure 1 in §2.1, for example, shows no apparent pattern in the arrival times of induced patients to the DU. In fact, in regressing DU occupancy against antenatal unit occupancy (not shown) we find antenatal workload not to have any explanatory power. Putting this together with the results in Table 7, this suggests that the midwives are not affected by antenatal workload when making a decision on the DU. It is important to recognize, though, that in other settings where there is more pressure to service "waiting" customers as soon as possible (e.g. when a queue builds up for a table in a restaurant) then this pressure might affect the gatekeeping behavior of the server.

3.4. Interaction effect models

In the main paper separate models were estimated for the subsets of non-complex and complex PEs. If the independent variables were to act in a similar way on both segments then this method would be inefficient, reducing the sample size (by splitting it in to two subsamples) and increasing the confidence intervals about the coefficient estimates. An alternative to this specification would have been to run a single model estimating coefficients jointly for both segments. In order to identify a differential effect of workload for the non-complex and complex PEs in such an estimation an interaction could be introduced between workload and complexity. By way of example, for the service configuration decision this model would be written:

$$\begin{aligned} EPI_i^* &= \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{W}_i + \alpha_2 ZLOAD_i + \alpha_3 COMPLEX_i + \alpha_4 \left(ZLOAD_i \times COMPLEX_i\right) + \delta_i \\ EPI_i &= \mathbb{1}[EPI_i^* > 0], \end{aligned}$$

where $\delta_i \sim \mathcal{N}(0,1)$, EPI_i^* is a latent variable, the vector $\mathbf{W}_i$ contains the set of controls, $COMPLEX_i$ is a dichotomous variable taking value 1 if the PE is complex and 0 otherwise, EPI_i is the observed dichotomous variable indicating epidural administration, and $\mathbb{1}[\cdot]$ is the indicator function. The coefficient α_4 then identifies the extent to which the workload effect for a complex PE differs from that of a non-complex PE, given by α_2.

While the model detailed above would have made it easier to interpret and compare the effects of workload across the two subsamples – and in general interaction effect models are preferable to subset analysis – for the purposes of the analysis performed in the paper this would have meant that it would not be possible to estimate the endogeneity corrected models since the following would have to be assumed:

1. Homoskedasticity of errors: more specifically that the correlation coefficient is identical for both the non-complex and complex sub-samples, which is shown not to be the case in the main paper; and

2. That the same instrumental variables could be used (i.e. that they were relevant and valid) for *both* the non-complex and complex samples, which we know not to be the case since we select particular IVs for use in particular models.

Despite the inappropriateness of the interaction modeling approach for the purposes of the main paper, though, to check for consistency it is helpful to re-run the analysis for the simpler Probit models in which endogeneity is not accounted for. The estimated effects are reported in Table 8 for both the rationing and referral decisions under a probit model specification.

The results in Table 8 support the findings from the main paper. Probit (1) shows that higher workload reduces the epidural rate for the non-complex PEs (APE $= -0.026$, p-value < 0.001), but the positive interaction between workload and the complexity dummy (APE $= 0.022$, p-value $= 0.003$) cancels out the effect of workload for the complex PEs, indicating that workload has no effect on the epidural rate during these episodes. In Probit (2) it can be seen that workload reduces the rate of referrals for physician-led deliveries for the non-complex PEs (APE $= -0.011$, p-value $= 0.013$), while the positive interaction effect between workload and the complexity dummy (APE $= 0.021$, p-value < 0.001) indicates that higher workload leads to a net increase in the referral rate for the complex PEs. Results are similar when adding the rationing (epidural) decision as an additional regressor in the referral equation in Probit (3). Note that while the results are consistent, the fact that it is not possible to account for endogeneity prohibits the use of interaction-type models in the main paper.

4. Other Outcome Measures

In §7.3 of the main paper three other outcome measures that were investigated were discussed. These were:

1. The incidence of third and fourth degree perineal tears;
2. The Apgar score recorded for the baby 5 minutes after birth; and
3. The length of time the patient spent in the DU.

The estimation results for these models are presented in this section. In addition, results are reported when estimating post-birth LOS using an alternative measure that is calculated as the number of full nights spent in the hospital post-delivery, rather than the number of hours used in the main paper. Summary statistics for these four additional outcome measures and their correlations with the primary covariates reported in the main paper are presented in Table 9.

Table 8 Average Partial Effects in Interaction Effect Models.

	Probit		
	(1) Epidural	(2) Phys.	(3) Phys.
Complexity	*All*	*All*	*All*
Std. workload	-0.026***	-0.011*	-0.007
	(0.005)	(0.004)	(0.004)
Workload×High complexity	0.022**	0.025***	0.021**
	(0.008)	(0.007)	(0.007)
High complexity	0.239***	0.021*	-0.026**
	(0.009)	(0.008)	(0.008)
Epidural	–	–	0.196***
	–	–	(0.007)
Dist. to home	-0.008*	-0.007†	-0.005
	(0.004)	(0.004)	(0.004)
2–4h op. tht. use	-0.016	0.001	0.004
	(0.015)	(0.013)	(0.012)
Inst. op. tht. use	-0.007	-0.097***	-0.095***
	(0.008)	(0.007)	(0.007)
N	16,355	16,355	16,355
Log-lik	-9,283.93	-7,994.19	-7,611.50
Pseudo-R^2	0.130	0.263	0.298

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

Table 9 Descriptive Statistics and Correlation Table for the Additional Outcome Measures.

	Descriptive statistics				Correlation table					
Variable	Mean	SD	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)
(7) 3rd/4th degree perineal tear	0.04	0.19	0.00	1.00	0.02*	-0.02*	-0.01	-0.01	0.01	0.00
(8) Apgar score < 9	0.03	0.18	0.00	1.00	-0.01	0.00	0.03***	0.10***	-0.02**	-0.01
(9) DU LOS (hours)	17.69	14.63	2.26	99.22	0.01	0.25***	0.32***	0.22***	-0.02*	0.03***
(10) Post-birth LOS (nights)	1.62	1.66	0.00	11.00	-0.01	0.05***	0.14***	0.32***	-0.04***	0.04***

(1) Std. workload, (2) Complex PE, (3) Epidural, (4) Phys.-led delivery, (5) Inst. op. tht. use, (6) 2–4h op. tht. use
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

4.1. Perineal tears

The rate of perineal tearing is commonly used as a quality indicator for cross-hospital comparison and features on the list of 26 hospital safety indicators compiled by the US Department of Health & Human Services Agency for Healthcare Research and Quality (AHRQ 2013). Precautionary action against tearing is difficult owing to the challenge in predicting if and how it will occur prior to labor (Williams et al. 2005); however, there is evidence that appropriate midwifery care during labor can reduce the likelihood of a tear (Aasheim et al. 2011). As such, if there is an effect of workload on the incidence of tearing then it should be expected that this will be positive in direction.

In order to examine the effect of workload on the tear rate of patients, it is necessary to exclude from the analysis sample any patient who gives birth by way of an elective C-section. This is because it only possible for a perineal tear to occur in patients who give birth vaginally. The results of the estimations using these reduced samples are provided in Table 10. Evidence suggests that higher levels of workload leads to an increase in the tear rate for complex PEs (APE = 0.011,

Table 10 Average Partial Effects for Perineal Tears.

	Probit			
Complexity	(1) Tear *nC*	(2) Tear *C*	(3) Tear *nC*	(4) Tear *C*
Std. workload	0.001	0.011***	0.001	0.011**
	(0.003)	(0.003)	(0.003)	(0.003)
Epidural	–	–	-0.008	-0.018**
	–	–	(0.005)	(0.006)
Phys. delivery	–	–	–	0.019*
	–	–	–	(0.008)
N	8,061	4,993	8,061	4,993
Log-lik	-1,506.26	-750.37	-1,505.29	-743.61
Pseudo-R^2	0.074	0.126	0.075	0.134

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; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

p-value $= 0.001$ in Probit (4)), but not for the non-complex PEs (APE $= 0.001$, p-value $= 0.715$ in Probit (3)).

Unfortunately, due to the relative infrequency of tearing, occurring in only 4.1% of births for non-complex PEs and 3.3% for complex PEs, it is difficult to use a bivariate probit model as it becomes challenging empirically to accurately estimate the correlation coefficient, ρ. In particular, in running BiProbit models the estimated 95% confidence intervals for ρ for the non-complex PEs is $\rho \in [-0.559, 0.585]$, while for the complex PEs $\rho \in [-0.371, 0.647]$. The wide confidence intervals for ρ make it hard to estimate the effect of midwife rationing and referral decisions on the tear rate, as the lack of precision in ρ leads to a great deal of uncertainty in the coefficient estimates of these covariates. This means that it is not possible to accurately identify the direct and indirect effects of workload. Given this, in Table 10 we report only basic probit models for the non-complex and complex PEs, and cannot comment specifically on how workload induced changes in midwife behavior affect this outcome. However, it is clear that when introducing the epidural decision as a control in the model for the non-complex PEs, i.e., when going from Probit (1) to Probit (3), there is no change in the direct workload effect, and similarly when adding the physician-led delivery decision in the model for the complex PEs, i.e., when going from Probit (2) to Probit (4). This indicates that even if there were an effect of workload through the rationing or referral decisions on the tear rate it is likely to be small.

4.2. Apgar score

The Apgar score is a number between 0 and 10 that is used to quickly summarize the health of newborn babies immediately after birth. Five factors are used to evaluate a baby's condition, each scored between 0 and 2, with 2 being the best score. The five factors are the appearance, pulse, grimace response, activity, and respiration of the baby. The higher the total score, the better the

Table 11 Average Partial Effects for Apgar Score.

	Probit			
Complexity	(1) Apgar *nC*	(2) Apgar *C*	(3) Apgar *nC*	(4) Apgar *C*
Std. workload	-0.003	0.002	-0.003	0.001
	(0.002)	(0.002)	(0.002)	(0.003)
Epidural	–	0.017***	0.004	0.013**
	–	(0.005)	(0.004)	(0.005)
Phys. delivery	–	–	–	0.015*
	–	–	–	(0.006)
N	9,976	6,195	9,976	6,195
Log-lik	-1,351.15	-821.75	-1,350.67	-817.89
Pseudo-R^2	0.068	0.087	0.068	0.092

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; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

indication that the newborn is in good health (Apgar 1953). The Apgar score reported in our data is the score taken 5 minutes after birth, at which point most babies have a near normal score. Due to the relative infrequency of low Apgar scores we have converted the score from a continuous to a binary form, with any score less than 9 assigned a value of 1, and any score of 9 or 10 (in most cases the reasons newborns score 9 instead of 10 is due transient cyanosis, which is prevalent, and causes many newborns to lose 1 point – see e.g. Montgomery 2000) assigned a value of 0. 184 observations are dropped from this analysis as Apgar score information was not available.

As was the case for perineal tears, only a small proportion of babies are assigned a low Apgar score (3.3% for both the non-complex and complex PEs) making it empirically challenging to estimate a bivariate probit model due to the difficulty in identifying the correlation coefficient, ρ, and so leading to poor specificity for the independent variables of interest (i.e., the rationing and referral decisions). However, there is no evidence – see Table 11 – of a direct workload effect, and, as was the case for tears, the fact that the coefficient estimates of workload do not change when including epidural and physician-led deliveries as additional regressors indicates that even if there were a workload induced effect of rationing/referral behavior on the Apgar score, the effect size would be small.

4.3. DU length of stay

The delivery unit length of stay (DU LOS) is a measure of throughput, capturing the length of time that a patient spends in the DU during birth before being discharged to the postnatal unit. The longer that the patient spends in the DU, the more of the midwife's expensive time and resource that they consume. In order to identify if there is an effect of workload on throughput through a change in rationing of referral behavior by the midwife, the results of estimations examining this effect are reported in Table 12, where DU LOS is measured as the natural logarithm of the number of hours that the patient spent in the DU.

Table 12 Coefficient Estimates in Statistical Models for Delivery Unit LOS.

	OLS				HeckTreat			
Complexity	(1) DU LOS *nC*	(2) DU LOS *C*	(3) DU LOS *nC*	(4) DU LOS *C*	(1) Epidural *nC*	(2) DU LOS *nC*	(3) Phys. *C*	(4) DU LOS *C*
Std. workload	-0.009	0.010	0.004	0.009	-0.083***	0.027***	0.064**	0.017*
	(0.006)	(0.007)	(0.006)	(0.007)	(0.015)	(0.007)	(0.023)	(0.008)
Epidural	–	0.406***	0.501***	0.388***	–	1.342***	0.689***	0.446***
	–	(0.013)	(0.011)	(0.013)	–	(0.023)	(0.041)	(0.029)
Phys. delivery	–	–	–	0.089***	–	–	–	-0.195
	–	–	–	(0.015)	–	–	–	(0.122)
2–4h op. tht. use	0.009	0.022	0.027	0.024	-0.050	–	–	–
	(0.023)	(0.027)	(0.021)	(0.027)	(0.044)	–	–	–
Inst. op. tht. use	0.004	-0.007	0.007	0.001	0.001	0.014	-0.332***	–
	(0.012)	(0.014)	(0.011)	(0.014)	(0.029)	(0.013)	(0.052)	–
N	10,091	6,264	10,091	6,264	10,091		6,264	
Log-lik	-7,868.29	-3,745.04	-7,064.91	-3,728.39	-12,067.48		-6,393.67	
Adj-R^2	0.147	0.352	0.276	0.356	–		–	
ρ	–	–	–	–	-0.819*** (0.012)		0.360* (0.143)	

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; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

As can be seen, before running endogeneity corrected (HeckTreat) models there appears to be a significant effect of the epidural analgesia decision (coef $= 0.501$, p-value < 0.001 in Probit (3)) and the physician-led delivery decision (coef $= 0.089$, p-value < 0.001 in Probit (4)) on DU LOS for both the non-complex and the complex PEs. In particular, patients who are rationed less or referred more experience an increase in the time that they spend in the DU. As was the case for post-birth LOS and cost, as described in the main paper, there is a risk that these effects are driven by reverse causality and/or omitted variable bias: Patients may receive an epidural or physician-led delivery because they are expected to have a longer labor and stay longer, and unobserved risk factors that may increase/decrease the patients likelihood of receiving an epidural or physician-led delivery may also affect their DU LOS. To try to account for the endogeneity it is necessary to estimate HeckTreat models. As reported in §7.3 of the main paper, however, there are problems in using the IVs in these estimations due to the fact that the time at which the IVs are measured overlaps with the time over which DU LOS is being measured. As such, we do not comment on the results here, but report them for the sake of completeness.

4.4. Post-birth LOS measured by number of nights

The measure of post-birth LOS used in the main paper was calculated as the number of hours between the time at which the patient gives birth and the time at which they are discharged from the hospital. One potential problem with this is that the time of day or day of week at which the patient gives birth may introduce a degree of noise into the post-birth LOS measure. For example, suppose that the hospital policy is that a patient should stay 2 nights after giving birth. Then a mother who e.g. gives birth naturally at 12:01am on a Monday will spend all day Monday, Tuesday and part of Wednesday in the hospital. Whereas a mother who gives birth at 11:58pm on that

same Monday will likewise be discharged on the same Wednesday, perhaps at the same time (after the physician conducts their discharge rounds). Yet this second patient will experience a post-birth LOS almost 24 hours shorter, due to the (random) time at which the baby was born.

The problem highlighted above is particular the case in the U.S., where most insurance companies stipulate that patients may stay two nights in the hospital after a vaginal delivery and three nights after a C-section. The guidelines in the UK, however, where the study DU is located, are not as rigid. The latest guidelines from the National Institute of Clinical Excellence, for example, stipulate (Ritchie et al. 1996, p. 54): "Length of stay in a maternity unit should be discussed between the individual woman and her healthcare professional, taking into account the health and well-being of the woman and her baby and the level of support available following discharge." Moreover, in the econometric specification are time-of birth fixed effects measured in two-hour blocks. These account for any systematic difference in the post-birth LOS of mothers delivering at different times of the day. Roughly speaking, with this control in place the estimation compares the post-birth LOS of a patient delivering between 12:01am and 02:00am, say, on a busy day with that of another patient delivering at the same time-window but on a non-busy day.

An alternative way of calculating post-birth LOS, however, would have been to look at the number of nights that the patient stayed in the hospital after delivery. This can be measured as the number of times after delivery that the patient experienced midnight in the hospital during their stay. For example, if a patient gave birth at 20:00 on 1st January and was discharged at 08:00 on 3rd January then the number of nights they would be considered to have stayed would be 2, once for passing midnight on 1st January and the other for crossing midnight on 2nd January. In Table 13 are results from OLS models (Poisson models produce similar results – not reported here) in which post-birth LOS is measured as the number of nights instead of the number of hours.

As is shown in Table 13, the results are qualitatively similar to those reported in the paper. For example, we find that epidural analgesia and physician-led delivery both increase the number of nights stayed, consistent with the results presented in the main paper. However, possibly due to the insufficient variation in the nightly measure of post-birth LOS, we do not identify a statistically significant workload effect at the nightly level. This finding is unsurprising given that the effect of workload on post-birth LOS reported in the main paper was shown to be in the magnitude of hours rather than days.

5. Alternative Explanations

In the main paper the increase in the rate of referrals to physicians at higher levels of workload was described as a lever used by midwives to relieve workload pressure and free up capacity to tend to the needs of other (non-referred) patients. In this section we discuss two other potential

Table 13 Coefficient Estimates in Statistical Models for Post-birth LOS Measured in Nights.

	OLS			
Complexity	(1) PbLOS *nC*	(2) PbLOS *C*	(3) PbLOS *nC*	(4) PbLOS *C*
Std. workload	-0.013	-0.010	-0.005	-0.021
	(0.016)	(0.022)	(0.016)	(0.021)
Epidural	–	0.414***	0.337***	0.282***
	–	(0.039)	(0.032)	(0.039)
Phys. delivery	–	–	–	0.638***
	–	–	–	(0.046)
2–4h op. tht. use	-0.045	-0.016	-0.034	-0.002
	(0.059)	(0.084)	(0.059)	(0.082)
Inst. op. tht. use	-0.078**	-0.073†	-0.076**	-0.015
	(0.029)	(0.041)	(0.029)	(0.041)
N	10,091	6,264	10,091	6,264
Log-lik	-17,655.40	-11,059.56	-17,603.00	-10,958.07
Adj-R^2	0.366	0.237	0.372	0.261

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; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

explanations for the increase in the rate of referrals at higher levels of load – one being that this is instead attributable to changes in physician (specialist) rather than midwife (gatekeeper) behavior, and the second being that there is an unobserved deterioration in quality at higher levels of workload that create a need for additional referrals – and discuss why these are less likely than the hypothesized behavior.

5.1. Interdependence of upstream and downstream workload

In the service setting that 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 is dependent also on the availability of the physicians and on whether the physician judges the referred patient to require their care. This means that if upstream (midwife) and downstream (physician) workload are interdependent 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. Therefore, it is important that, as far as possible, we are able to isolate the effect of workload specific to the midwives. This can be achieved by controlling for the workload of the physicians.

However, accounting for physician workload fully is difficult as demand is not readily observed. In particular, unlike demand for midwives – which can be inferred directly using the number of patients in the DU (since every patient must be assigned a midwife) – for the physicians, the number of patients in the DU will instead be a poor measure of load since not every patient interacts with a physician. Moreover, obstetricians and physicians working in the DU may also

perform other duties and care for patients not physically present in the DU, such as working in gynecology, performing prenatal checks, conducting postnatal examinations and discharges, etc. This means that workload as measured by the number of patients in the DU will only be weakly correlated with the actual workload of the physicians.

If we assume that the workload of physicians that takes place outside of the DU is exogenous and uncorrelated with DU workload – which is not unreasonable given the random nature of arrivals to the DU, with any residual correlation accounted for with time-fixed effects – then the component of a physician's workload that needs to be accounted for is that which is caused by the patients present in the DU. However, data is unfortunately not available on the time that a physician spent with each patient. Therefore, it is necessary to account for this instead using control variables that we know to be highly correlated with this work.

To identify these control variables it is helpful to consider the nature of the work performed by physicians in the DU. For a simple natural birth a physician will have limited involvement with the patient since labor and the delivery will be midwife-led. It is only when responsibility for the birth is passed from a midwife to a physician in more difficult cases that a physician becomes significant involved and so will experience a non-negligible increase in her workload. Therefore, the main demand placed on physicians by patients in the DU can be approximated by the number of instrumental deliveries or C-section that are performed. By controlling for the busyness of the operating theater at the times at which the rationing and referral decisions are made, as we do, we therefore are able to control for a large portion of the variation in physician workload caused by patients in the DU. Any residual differences in physician availability across different times of the day are also accounted for with time-of-day fixed effects. This means that the residual effect of workload that we estimate in the analysis in the main paper can be primarily attributed to the midwife, rather than the physician, as desired.

Finally, it is worth noting that while it is possible that downstream workload may increase with upstream workload in ways that we have not been able to capture using the controls described above, there is no reason to believe that physicians should respond to an *increase* in their workload by offering to perform *more* assisted deliveries, since this would only serve to make them even busier. Therefore, if anything, it is more likely that any change in physician behavior would serve to dampen the effect that we observe and are attributing to gatekeeping behavior, rendering our estimate a conservative estimate of the effect of workload on the midwives' referral behaviour.

5.2. Unobserved deterioration in service quality

Workload has been shown, both in the operations management and the medical literature, to have an adverse effect on service quality and clinical outcomes (Needleman et al. 2011, KC and

Terwiesch 2009, Kuntz et al. 2014). An alternative explanation for the increase in referrals at higher load, therefore, may be that as the midwives become busier they are unable to monitor patients as closely or provide as high quality care. If this causes an increase in risk to the patients, then the physicians may have to intervene in order to maintain high quality patient outcomes. Then, as a result of model specification error, the observed effect of workload on the referral rate may instead be explained by a common response variable, the unobserved health status of the patient (or baby), which is the true causal factor. While this does not invalidate the empirical results, it does bring into question the interpretation of the results. Why does workload result in more physician referrals? Is it because of the midwife pushing work to the physician when she becomes too busy? Or is it because the midwife's quality of service degrades and so the physician must actively intervene?

If patients were put more at risk at higher workload then one should also expect there to be a deterioration in outcomes correlated with that risk. If this were not the case then there would be no need for the physician to intervene. Evidence from two measures of service quality – that we report on in the main paper and elsewhere in the supplementary material – suggests that this is not the case. In particular:

- Post-birth LOS: In §7.2.1 of the paper we show that there is no evidence that increased workload in the DU leads to an increase in post-birth LOS of the complex cases, neither directly or indirectly. This indicates that patients with complex needs who are in the DU when it is busy stay no longer in the hospital after delivery than those that give birth when the DU is quieter. If the risk profile of the patients were increased at higher workload, however, we would expect them to stay longer in the hospital after delivery since they require more careful observation, further tests, and more time to recover.
- Apgar score: In §4.2 of the supplementary material we show that there is no evidence that workload affects the likelihood of a baby being born with a lower Apgar score (a summary measure of the health status of newborn babies immediately after birth). If the risk profile of the babies were increased at higher workload, however, then one would expect it to be reflected in the Apgar score, which measures factors that include the appearance, pulse, grimace response, activity, and respiration of the baby.

This provides some evidence against the argument that quality of care in the DU deteriorates, in an unobservable way, at higher levels of workload. It is possible, though, that these additional risks may not be substantial enough to manifest themselves in worse outcomes. For example, it is plausible that factors such as increase heart rate for the baby and mother could necessitate physician intervention. While it is not feasible to consider every possible variable that might signal

Table 14 Average Partial Effects in Statistical Models for Fetal Distress

	Probit	
Complexity	(1) Distress *nC*	(2) Distress *C*
Std. workload	-0.007	0.007
	(0.005)	(0.006)
N	10,091	6,264
Log-lik	-5,792.20	-3,541.51
Pseudo-R^2	0.070	0.091

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; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

a deterioration in service quality, we have identified one factor in the data that we believe should be highly correlated with quality of care: the reported incidence of fetal distress in labor.

Fetal distress is reported if there are signs during labor that suggest that the fetus may not be well. Examples of such signs include elevated or decreased fetal heart rate, decreased movement felt by the mother and signs of acidosis in the fetal blood. Fetal distress occurs at a rate of 31.0% in the final analysis sample of 16,355 patients. Care by a certified midwife has been shown to be associated with a lower risk of diagnosis of fetal distress (Butler et al. 1993) and appropriate response to fetal distress by midwives can reduce distress and associated C-sections (Gagnon et al. 1997). Therefore, if at higher levels of workload there is a deterioration in monitoring and service quality that necessitates an increase in referrals to physicians, this would be expected to be reflected by an increase in the observed rate of fetal distress. There is no evidence that this is the case, however: In Table 14 the APE of workload on the reported rates of fetal distress is shown to be insignificant for both the non-complex (APE $= -0.007$, p-value $= 0.134$) and complex (APE $= 0.007$, p-value $= 0.248$) patient episodes.

In summary, we find no evidence in the data that there is any systematic deterioration in the health status of the patients or the babies at higher levels of workload that would be driving the increase in referrals. This suggests instead that the midwife protects against the potentially adverse impact of excess load on patient safety by pulling on the two levels at her disposal: the service configuration decision and the referral decision. Overall, therefore, we claim that the results we observe are more likely to be driven by the hypothesized behavioral change, though we acknowledge that it is possible that a small component of this effect may be driven by an unobserved increase in patient risk.

6. Relevance and Validity of the Instruments

One of the principal findings in the main paper is that workload can have a surprisingly large effect on the provision of discretionary components of a service, e.g. epidural analgesia, and through this can, unexpectedly, reduce the likelihood of a customer experiencing an undesirable outcome, such as an unnecessary physician-led delivery. Given that endogeneity of the epidural analgesia variable was a concern when estimating its effect on the rate of physician-led deliveries, this was corrected for in the paper by specifying a recursive bivariate probit (BiProbit) model. When doing so we explained that in order to improve the estimation it was desirable (though not strictly necessary) to find instrumental variables (IVs). These IVs are designed to be included in the estimation of the potential endogeneous regressor, but excluded from the estimation of the outcome. Two instruments that were identified for the epidural/physician-led delivery BiProbit models were (i) the number of operating theaters in use in the 2-4 hour period prior to the time of birth of the focal patient, and (ii) the distance between the patient's place of residence and the hospital. In this section formal testing is performed to determine the relevance and validity of these instruments.

6.1. Tests of under- and weak identification

The underidentification test is a Lagrange multiplier (LM) test to determine whether the equation is identified. Specifically, the test determines whether the excluded instruments are correlated with the potential endogenous regressor, i.e. that the excluded instruments are "relevant" in the selection (first-stage) equation. "Weak identification", on the other hand, arises when the excluded instruments are correlated with the endogenous regressors, but only weakly. Estimators can perform poorly when instruments are weak: estimates may be inconsistent, tests for the significance of coefficients may lead to the wrong conclusions, and confidence intervals are likely to be be incorrect. Here we describe how we test for both of these properties.

First it is important to note that the majority of tests are based on a linear IV regression model where the dependent variable in the outcome equation and the endogeneous variable are continuous. In order to perform formal testing we therefore follow convention and treat the binary epidural analgesia and physician-led delivery variables as continuous. While this means that the true critical values of the tests and significance levels may differ from those that are reported here, we note that differences in estimated parameters that arise from using a continuous rather than binary model specification are often small, and that the estimated coefficients using these models (not shown) are consistent with those reported in the main paper.

In testing for both underidentification and weak identification we use the method of Sanderson and Windmeijer (2015), implemented in and reported by the `ivreg2` command in Stata 12.1 (Baum et al. 2010). The Sanderson-Windmeijer (SW) first-stage chi-squared Wald statistic is distributed

as chi-squared with $(I_E - N_{EN} + 1)$ degrees of freedom under the null that the particular endogenous regressor of interest is underidentified, where I_E is the number of excluded instruments ($= 2$ here) and N_{EN} is the number of endogenous regressors ($= 1$ here). For the model under investigation, the SW Chi-sq statistic is calculated to take a value of 15.91 with 2 d.f., which has corresponding p-value $= 0.0004$. This means that there is strong evidence to reject the null hypothesis of underidentification at e.g. the 1% significance level, and so it is possible to conclude that the excluded instruments are "relevant".

Turning next to the issue of weak identification, the SW first-stage F-statistic is the F form of the SW chi-squared test statistic and can be used as a diagnostic for whether a particular endogenous regressor is "weakly identified". In particular, the F-statistic can be compared against the critical values for the Cragg-Donald F-statistic reported in Stock and Yogo (2005) to determine whether or not the instruments perform poorly. The relevant test has null hypothesis that the maximum bias of the IV estimator relative to the bias of ordinary least squares, i.e. $\left|\frac{\mathbb{E}[\hat{\beta}_{IV}]-\beta}{\mathbb{E}[\hat{\beta}_{OLS}]-\beta}\right|$, is b, where b is some specified value such as 10%. For a single endogenous regressor, assuming the model to be estimated under limited information maximum likelihood, the critical F-values are 8.68, 5.33 and 4.42 for maximum biases of $b = 10\%$, 15%, and 20%, respectively. If the estimated F-statistic is less than a particular critical value then the conclusion is that the instruments are weak for that level of bias. Here, the estimated SW F-statistic is equal to 7.89, indicating a maximal bias of between 10% and 15%. This indicates that there is no evidence to suspect that our models are heavily affected by the problem of weak instruments.

6.2. Testing for overidentification

In addition to the excluded instruments being "relevant", it is also important to check that they are "valid", i.e. (1) uncorrelated with the error term (i.e. orthogonal to epsilon) and (2) correctly excluded from the output equation (i.e. only indirectly influence dependent variable y). A joint null hypothesis test that (1) and (2) are satisfied, i.e. of the overidentifying restrictions, is the Sargan-Hansen (SH) test (Sargan 1958, Hansen 1982). A rejection of the SH test raises doubts about the validitiy of the instruments. Under the null, the test statistic is distributed as chi-squared with degrees of freedom equal to the number of $(I - N)$ overidentifying restrictions, where I is the number of instruments (I_I included and I_E excluded) and N is the number of regressors (N_{EN} endogeneous and N_{EX} exogenous). In the models in the paper $I_I = N_{EX}$ and so the degrees of freedom is equal to $I_I - N_{EN} = 2 - 1 = 1$. When $I = N$ the model is exactly identified and the SH test cannot be performed.

The Sargan-Hansen test statistic is reported in the output when using the `ivreg2` command in Stata 12.1 (Baum et al. 2010), and takes value 2.72, with corresponding p-value $= 0.10$. This is insignificant at the 5% significance level, and so we do not reject the null hypothesis that the overidentifying restrictions are satisfied, i.e. the evidence suggests that the instruments are "valid".

7. Effect Size Calculations for HeckTreat Models

We report below the method for calculating the effect size under a HeckTreat model. While the method for doing this for a BiProbit model is well documented, this is not the case for the HeckTreat model and so the derivation is provided below in order to elucidate this process.

The HeckTreat model can be expressed as:

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta} + \delta z_i + \epsilon_i \\ z_i &= \mathbb{1}[z_i^* > 0] \\ z_i^* &= \mathbf{w}_i\boldsymbol{\gamma} + u_i \\ (\epsilon_i, u_i) &\sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma & \rho\\ \rho & 1\end{pmatrix}\right). \end{aligned}$$

We are interested in the marginal effect of the regressors on y_i in the observed sample. For a particular regressor, x_{ik}, this is equal to

$$\frac{\partial \mathbb{E}\left[y_i | w_i, x_i, z_i\right]}{\partial x_{ik}} .$$

To find this, start by observing that

$$\mathbb{E}\left[y_i \,|\, w_i, x_i, z_i\right] = \mathbb{P}\left[z_i = 0 \,|\, w_i\right] \mathbb{E}\left[y_i \,|\, w_i, x_i, z_i = 0\right] + \mathbb{P}\left[z_i = 1 \,|\, w_i\right] \mathbb{E}\left[y_i \,|\, w_i, x_i, z_i = 1\right] .$$

Solving for the component parts, we get

$$\mathbb{P}\left[z_i = 0 \,|\, w_i\right] = \mathbb{P}\left[z_i^* \leq 0 \,|\, w_i\right] = \mathbb{P}\left[\mathbf{w}_i\boldsymbol{\gamma} + u_i \leq 0 \,|\, w_i\right] = \mathbb{P}\left[u_i \geq -\mathbf{w}_i\boldsymbol{\gamma} \,|\, w_i\right] = 1 - \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right) ,$$

and a similar derivation gives $\mathbb{P}\left[z_i = 1 \,|\, w_i\right] = \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)$. Using the moments of the incidentally truncated bivariate normal distribution (Theorem 19.5 in Greene 2008) gives

$$\begin{aligned} \mathbb{E}\left[y_i \,|\, w_i, x_i, z_i = 0\right] &= \mathbf{x}_i\boldsymbol{\beta} - \sigma\rho\frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{1 - \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)} \\ \mathbb{E}\left[y_i \,|\, w_i, x_i, z_i = 1\right] &= \mathbf{x}_i\boldsymbol{\beta} + \delta + \sigma\rho\frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)} . \end{aligned}$$

Putting this together it follows that

$$\begin{aligned} \mathbb{E}\left[y_i \,|\, w_i, x_i, z_i\right] &= \left(1 - \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right)\left[\mathbf{x}_i\boldsymbol{\beta} - \sigma\rho\frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{1 - \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}\right] + \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\left[\mathbf{x}_i\boldsymbol{\beta} + \delta + \sigma\rho\frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}\right] \\ &= \left[\mathbf{x}_i\boldsymbol{\beta} - \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\mathbf{x}_i\boldsymbol{\beta} - \sigma\rho\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right] + \left[\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\mathbf{x}_i\boldsymbol{\beta} + \Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\delta + \sigma\rho\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right] \\ &= \mathbf{x}_i\boldsymbol{\beta} + \delta\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right) . \end{aligned}$$

This means that the total change of a regressor x_{ik} that appears in both $\mathbf{x}_i$ and $\mathbf{w}_i$ on y is

$$\frac{\partial \mathbb{E}\left[y_i | w_i, x_i, z_i\right]}{\partial x_{ik}} = \beta_k + \delta\gamma_k\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right) , \tag{3}$$

where β_k and γ_k are the coefficient estimates for x_{ik} in the outcome and selection equations respectively, and δ the coefficient for the endogenous regressor z in the outcome equation. The term β_k is the direct effect of workload on y; $\delta\gamma_k\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)$ is the indirect effect of workload through z.

References

Aasheim, V., A. Nilsen, M. Lukasse, L. Reinar. 2011. Perineal techniques during the second stage of labour for reducing perineal trauma. *Cochrane Database of Systematic Reviews* **12** 1–47.

AHRQ. 2013. Patient safety indicators technical specifications. URL `http://www.qualityindicators.ahrq.gov/Modules/PSI_TechSpec.aspx`.

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

Apgar, V. 1953. A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia* **32**(4) 260–267.

Ball, J., M. Washbrook. 1996. *Birthrate Plus*. Books for Midwives Press, Cheshire, England.

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.

Baum, C.F., M.E. Schaffer, S. Stillman. 2010. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. URL `http://ideas.repec.org/c/boc/bocode/s425401.html`.

Butler, J., B. Abrams, J. Parker, J. M. Roberts, R. K. Laros Jr. 1993. Supportive nurse-midwife care is associated with a reduced incidence of cesarean section. *American Journal of Obstetrics and Gynecology* **168**(5) 1407–1413.

Duncan, O.D. 1975. *Introduction to structural equation models*. Academic Press Inc., New York.

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

Gagnon, A.J., K. Waghorn, C. Covell. 1997. A randomized trial of one-to-one nurse support of women in labor. *Birth* **24**(2) 71–77.

Greene, W. 2008. *Econometric Analysis*. 7th ed. Prentice Hall.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

KC, D. 2013. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

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.

Kim, J.O., G.D. Ferree. 1981. Standardization in causal analysis. *Sociological Methods and Research* **10**(2) 187–210.

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

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

Montgomery, K.S. 2000. Apgar scores: Examining the long-term significance. *The Journal of Perinatal Education* **9**(3) 5–9.

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

NHS. 2015. Epidural anaesthesia - what it is used for. *NHS Choices* URL `http://www.nhs.uk/Conditions/Epidural-anaesthesia/Pages/Whatitisusedfor.aspx`. Published: 02/02/2015. Accessed: 2016-01-06.

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

Ritchie, G., N. Turnbull, C. Adams, C. Barry, S. Byrom, D. Elliman, S. Marchant, R. Mccandlish, H. Mellows, C. Neale, M. Parkar, P. Tait, C. Taylor. 1996. *Clinical Guidelines And Evidence Review For Post Natal Care: Routine Post Natal Care Of Recently Delivered Women And Their Babies*. National Collaborating Centre For Primary Care And Royal College Of General Practitioners, London.

Sanderson, E., F. Windmeijer. 2015. A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics* .

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* **26** 393–415.

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

Stock, J., M. Yogo. 2005. Testing for weak instruments in linear IV regression. D. Andrews, J. Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge University Press, 80–108.

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.

Williams, A., D. Tincello, S. White, E. Adams, Z. Alfirevic, D. Richmond. 2005. Risk scoring system for prediction of obstetric anal sphincter injury. *British Journal of Obstetrics and Gynaecology* **112**(8) 1066–1069.

Winship, C., R. Mare. 1984. Regression models with ordinal variables. *American Sociological Review* **49**(4) 512–525.

Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73**(1) 3–36.

Wood, S. 2013. mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. URL `http://cran.r-project.org/package=mgcv`.