Received: 8 September 2020 | Revised: 6 October 2021 | Accepted: 3 November 2021

DOI: 10.1002/hec.4460

RESEARCH ARTICLE

Health Economics WILEY

Forgetting-by-not-doing: The case of surgeons and cesarean sections

Gabriel Facchini

Department of Applied Economics, Universitat Autònoma de Barcelona, Bellaterra, Spain

Correspondence
Gabriel Facchini, Department of Applied Economics, Universitat Autònoma de Barcelona, Bellaterra, Spain.
Email: gabriel.facchini@uab.cat

Funding information
Ministerio de Educación y Formación Profesional, Grant/Award Number: PID2019-104619RB-C43
General Secretariat for Research-Government of Catalonia, Grant/Award Number: SGR2017-1301

Abstract

This paper provides new evidence on the link between patient outcome and physician experience. Using birth certificates data from a large hospital in Italy, I analyze whether cesarean section surgeons who have performed more procedures in the recent past observe an improvement in performance. By using data from the Italian health care system, where patients are not allowed to choose their physician, I lower concerns of potential reverse causality (selective referral). I find evidence indicating a strong learning-by-doing effect: for emergent cases, a one standard deviation increase in recent experience reduces the likelihood of neonatal intensive care unit admission by nearly 3.2 percentage points (13.8%) and of being born with a low Apgar Score by about 1.9 percentage points (13.2%), all else equal. This effect is not present for the case of elective C-sections.

KEYWORDS
cesarean section, experience, human capital, learning-by-doing, productivity, volume

JEL CLASSIFICATION
J24, I10, I18

1 | INTRODUCTION

Since the seminal report by (Luft et al., 1979), there has been growing evidence of a positive association between volume and quality in the provision of health services for a wide variety of procedures, time periods, and locations.[1] Nevertheless, the debate about the causal direction of this relationship is far from settled (Halm et al., 2002; Ho, 2014).

Two principal hypothesis have been put forward to explain this association: (i) "learning-by-doing" (or "practice-makes-perfect") and (ii) "selective referral" (Luft et al., 1987).[2] Under "learning-by-doing," increased experience leads to improvement in skills which in turn results in better quality as measured by patient outcomes. "Selective referral," instead, occurs when providers with higher quality attract a larger volume of patients. The importance of identifying which one is driving the correlation between volume and outcome stems from the fact that they have opposite policy implications. If volume causes outcome, as learning-by-doing suggests, then the concentration of procedures in fewer and bigger providers would raise quality. However, if causality runs from outcome to volume, then those benefits are not present anymore, and concentration would only lead to reduced competition between providers and lower geographical coverage.

This paper aims at causally identifying whether learning-by-doing is present at the individual level in the healthcare sector, more specifically, for surgeons performing cesarean sections (C-sections). In particular, I look at whether a surgeon's recent procedure volume affects patient outcomes.[3] In order to establish a causal relationship, I benefit from the fact that, due to state regulation, most pregnant women in Italy do not choose the gynecologist that will help them give birth within the public system. This institutional feature creates a setup where selective referral is not possible.

I make use of a census of birth certificates from a large public hospital in Italy for the period 2011–2014 that contains surgeon identifier for each surgery. Even though patients cannot choose a particular physician, the hospital may assign physicians with higher skills to patients with a higher health risk (selective allocation). To address this concern I use a fixed effect model and rely on changes in volume within surgeon for the estimation. I find strong evidence of learning-by-doing for C-section surgeons: operations performed by physicians with a higher recent experience result in better newborn health. More specifically, I find that a one standard deviation increase in surgeon's experience in the previous 4 weeks lowers a newborn's probability of having a low Apgar score by 13.2%[4] and of being admitted to a neonatal intensive care unit (NICU) by 13.8%. These effects are only present for emergent C-sections (not for elective C-sections), meaning, cases in which the surgeon has to make crucial decisions against the clock.

One important assumption for these results to hold is the absence of any form of dynamic matching between physicians and patients. If, for example, hospitals aware of depreciating skills may assign healthier patients to physicians with a low recent activity, which would biased the estimates toward zero. In this case the results should be considered a lower bound of the true effect. To alleviate this concern, I first show that pre-treatment pregnancy and mother characteristics are uncorrelated with physician's recent experience. Second, one would expect that, if there is some form of dynamic matching between physicians and patients, emergent cases should get the more experienced physicians—biasing my results downwards. However, as mentioned before, I see the strongest effects of recent experience on these non-elective cases. Third, I also implement a sensitivity checks using a bounding approach following Oster (2019) and find that unobservables are unlikely to explain these results. Finally, I perform a robustness check by using different windows for recent experience, from 4 to 52 weeks. Results show that only very recent experience matters providing further evidence for the human capital depreciation hypothesis.

Cesarean sections are an attractive procedure to analyze the presence of surgeon's "learning-by-doing" hypothesis. Unlike other highly studied procedures that are performed by a team of surgeons, C-sections are executed by only one surgeon, allowing for better estimates of the individual surgeon's learning curve. In addition, for many developed countries, C-sections have become the most common surgical procedure.[5] Furthermore, the discussion on volume-outcome effects become all the more relevant in view of the recent wave of closures of maternity services in various countries (e.g., US, Canada, UK, Japan, France, the Netherlands, and others).[6] To the best of my knowledge, this is the first paper to obtain causal estimates of learning-by-doing for the case of cesarean section surgeons.

1.1 | Literature review

There are hundreds of papers in the medical literature finding an association between higher hospital or surgeon procedure volume and better health outcomes (Birkmeyer et al., 2003; Chowdhury et al., 2007; Halm et al., 2002). However, these studies are mostly observational and tend to neglect the potentially endogenous nature of provider volume. Few studies have attempted to translate the association between volume and outcome into a causal relationship, and most rigorous econometric analysis have failed to identify learning-by-doing (Ho, 2014).

At the hospital-level, studies on learning-by-doing typically use lagged or cumulative volume as covariates of interest, and find no support for the learning-by-doing hypothesis (Gaynor et al., 2005; Ho, 2002; Sfekas, 2009). One exception is Avdic et al. (2019), who find a positive effect of hospital operation volume on patients' survival using Swedish register data on advanced cancer surgery procedures. They exploit the closures and openings of entire cancer clinics as an exogenous variation for volume in an instrumental variable set up. Importantly, they provide suggestive evidence that the effect on outcome is mainly due to increases in individual surgeon's experience.[7]

However, the literature testing for volume-outcome effects using individual (surgeon) level data is much more limited and it finds mixed results. On the one hand, Huesch (2009) and Contreras et al. (2011) fail to find any association between cumulative surgeon procedure volume and patient's health. Using longitudinal data for a specific eye surgery (LASIK) in one clinic in Colombia, Contreras et al. (2011) find no effect of cumulative volume on outcome. They exploit a quasi-random allocation of surgeons to patients which makes selective allocation less of a problem. Similarly, Huesch (2009)

fails to find any effect of cumulative volume on outcome for a panel of surgeons performing coronary artery bypass grafts (CABG) in Florida in the period 1998–2006. Moreover, he finds that almost all prior experience is depreciated from one quarter to the next. The author uses a choice model and predicted volume to mitigate potential issues of selective referral, although he does not reject exogeneity of volume.

On the other hand, Ramanarayanan (2008) and Huckman and Pisano (2006) find evidence of strong learning-by-doing effects at the physician level when using a measure of recent experience as their covariate of interest—instead of cumulative volume. Ramanarayanan (2008) studies the same dataset for CABG surgeons as Huesch (2009) but uses the departure of a surgeon as an exogenous shock to the yearly volume of the remaining physicians. Instead, Huckman and Pisano (2006) do not discuss potential bias to a great extent and confine themselves to using surgeon risk-adjusted mortality as quality controls. They also focus on CABG cases—although their data comes from Pennsylvania for 1994 and 1995—and find that the mortality rate of patients decreases significantly with increases in the surgeon's experience in the previous calendar quarter.[8]

The current paper contributes to the existing literature in several ways. First, it provides new evidence of the causal link between patient outcomes and surgeon experience. As clearly showed before, the literature on volume-outcome at the individual level is in its early steps and more research is needed. Second, previous studies looking at the causal effects of volume on outcome rely mostly on instrumental variable estimates. This paper serves as a complement to previous studies by exploiting a set up where selective referral is not possible, together with a dataset that allows to estimate the effect from within surgeon variation in volume. In addition, most previous studies use health care data from the United States, and focus almost exclusively on coronary artery disease procedures. Finally, the data employed allows me to make more precise estimates about the volume of patients seen by each surgeon, while previous studies have relied mainly on yearly or quarterly data.

2 | CLINICAL AND INSTITUTIONAL SETTING

2.1 | The performance and organization of cesarean sections

A cesarean section (C-section) is a major surgical procedure in which a fetus is delivered through an incision in the mother's abdomen and uterus (American College of Obstetricians and Gynecologists, 2010). The procedure typically takes 45 min to an hour, and most mothers and babies stay in the hospital for 2 to 3 days.

Based on their degree of urgency, C-sections are typically classified in two groups: elective (or scheduled), and emergent (Lucas et al., 2000). The first group includes all C-sections scheduled in advance to occur before labor begins on the basis of an obstetrical or medical indication -although there is no immediate maternal or fetal compromise. The second group of C-sections includes all cases where the patient attempts to have a vaginal delivery (either through the natural onset of labor or medical induction) but end up delivering by C-section instead. This occurs when the patient develops complications that put in danger the health of the infant and/or the mother and thus the physician recommends to change delivery method toward surgery.

2.2 | The Italian health care system and C-sections

Italian health care is a universal, public-private insurance system. The public part is the national health service—Sistema Sanitario Nazionale (SSN)—which is administered on a regional basis. According to the World Health Organization, in 2000 the Italian system provided the second best overall health care in the world—the first one being France (WHO, 2000). Furthermore it has the lowest maternal mortality rate worldwide at 1.94 for every 100,000 births (WHO et al., 2019).

Under this system, a pregnant woman cannot choose the physician or midwife that will assist her for the delivery unless she pays. Furthermore, given the well functioning of the system, the grand majority (89%) of women choose to use the public service (Ministero della Salute, 2019).[9] This institutional feature eliminates the risk of selective referral, where institutions or surgeons with better performance attract higher volumes of patients—a common endogeneity issue in studies of learning-by-doing.

In the year 2016, Italy had an overall C-section rate of 36.8%, with great disparities between the public (31.7%) and the private sector (50.9%) (Ministero della Salute, 2019). According to the Health Statistics 2019 from the Or-

ganisation for Economic Co-operation and Development (OECD), Italy has the seventh largest C-section rate among OECD countries, with Turkey showing the highest rate at 53.1% and Israel the lowest one at 14.8% (the OECD average was 28.1%). These cross-country variation can be linked not only to patient characteristics, but also to non-medical reasons. Previous work has found that higher C-sections can be the result of higher fees in comparison to vaginal deliveries (Alexander, 2013; Allin et al., 2015; Gruber et al., 1999; Gruber & Owings, 1996), defensive medicine (Bertoli & Grembi, 2019; Currie & MacLeod, 2008; Dranove & Watanabe, 2010) and physician's scheduling convenience (Facchini, 2020; Lefèvre, 2014).[10]

Physicians working in the delivery room in public hospitals are paid a fixed salary, which means they have no personal financial incentive to recommend any particular treatment. They have full-time contracts and are hence not allowed to work in other hospitals, either public or private—although they may take leaves to visit other institutions.

3 | EMPIRICAL METHODOLOGY

3.1 | Empirical model

The main question addressed in this paper is whether there is learning-by-doing in cesarean section surgeons. I test this by looking at whether surgeon's recent experience (e_{st}) has an impact on the next surgery's outcome. Thus I estimate a reduced-form model of the following type:

$$y_{ist} = \alpha + \beta e_{st} + \delta d_{st} + \mathbf{x}_{it}' \theta + \phi_t + \eta_s + \epsilon_{ist} \quad (1)$$

where y_{ist} is a health indicator for patient i whose procedure was performed by surgeon s at time t. Surgeon's recent experience is defined as the number of C-sections performed in the 4 weeks leading up to and including the procedure on the patient surgeon s operated on just before operating on patient i.[11] d is a control for the number of days since the prior cesarean section surgeon s performed.[12] $\mathbf{x}_{it}$ contains individual-level control variables for mother and pregnancy characteristics.[13] ϕ is a vector of indicators for year, month and day of the week of delivery.

Individual surgeon fixed effects (η_s) are included to mitigate concerns that the captured relationship between outcomes and recent experience is driven by composition effects. Surgeon fixed effects ensures that the recent experience parameter in Equation (1) is identified from changes in volume *within* surgeon.[14] As discussed above, if physicians skills improve with recent repetition, then β should be negative: since outcomes are defined as adverse, a higher recent volume of surgeries would help (partially) avoid the loss of skills. On the contrary, a coefficient close to zero would imply that there is full depreciation and recent experience does not affect current outcome.

One important assumption for the previous model to obtain causal effects is the lack of any compositional effect of patients between physicians with different levels of recent experience. To test whether selection bias affects my estimates, I regress each pre-treatment characteristic on the treatment—that is, the number of C-sections performed in the last 4 weeks. If observed characteristics were associated with recent experience, it would be a sign of patient selection. The results for these estimations are reported graphically in Figure 1. After controlling for physician and time fixed effects, the treatment does not predict any of the observed maternal and pregnancy characteristics.[15] This provides further evidence that mothers undergoing surgery with a physician with higher or lower recent experience are similar in observable characteristics.

Even if physician fixed effects help alleviate issues of selection of patients based on physician's skills, there could still be problems of endogeneity if some sort of dynamic matching exists or "selective allocation" (Huesch & Sakakibara, 2009). For instance, hospitals aware of depreciating skills may assign healthier patients to physicians coming back from a period of low activity. In this case my estimates on the impact of recent experience on patients' health would suffer from a downward bias.[16] To mitigate these concerns, I estimate a separate coefficient for different types of C-sections depending on their emergency status. For patients admitted emergently, given the unexpected nature of these cases, surgeons need to make fast decisions under pressure and skill depreciation should be of particular relevance.[17] On the other hand, if selective allocation exist, we would expect these difficult cases to be taken by more experienced surgeons, which would bias my estimates downwards.

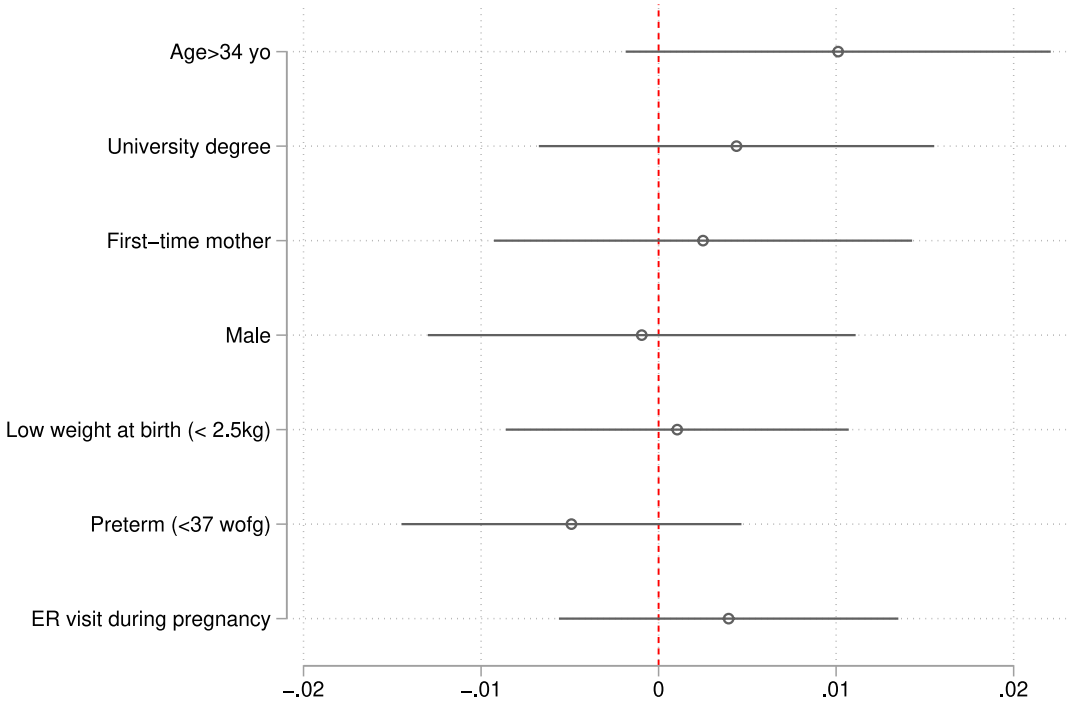

FIGURE 1 Balanced pre-treatment characteristics. The figure represents the coefficients and 95% confidence intervals from separate regressions of each predetermined variable on recent experience, controlling for days since prior C-section, physician fixed effects, and day-of-the-week, month, and year of birth fixed effects [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Data

This study utilizes birth certificates from the maternity ward of a large public university hospital in Tuscany (Italy) for the years 2011 through 2014. The hospital has an average C-section rate of 31% average, close to the national Italian rate of 33% among public hospitals in 2012.[18] Birth certificates constitute a census of all births that took place in the hospital in this period. It contains information on mother characteristics (e.g., community of residence, education, civil status, age, previous deliveries, etc.), pregnancy characteristics (e.g., weeks of gestation, controls, assisted reproduction, etc.), birth characteristics (e.g., time of birth, type of labor, attendant, place, etc.) and indicators on newborns' health (e.g., weight, height, Apgar score, death, etc.). This information is complemented with surgeon's ID.[19]

The richness of this dataset comes at a cost: because the information available corresponds to just one hospital in a 4-year period, the sample size is relatively small. There were approximately 12,343 newborns during the period under study, 4413 (35%) of which were delivered via C-sections. Almost half of these C-sections are planned in advance between the physician and the patient (elective C-sections). From the 4413 cases, I keep only one observation per pregnancy and drop 427 observations from plural births. In addition, I drop from the analysis 86 births that have missing information in at least one of the variables used. Then I restrict the sample to surgeons who have performed, on average, at least 12 C-sections a year. This leaves the sample with 60 surgeons who performed 3599 (92%) surgeries. In addition, since my measure of experience is the number of C-sections in the past 4 weeks by surgeon, I drop all births that take place in the first 28 days from a surgeon's observed first C-section in my data. This leaves a sample of 3467 births performed in the 4-year period. Finally, I restrict the analysis to observations in which the surgeon has performed at least one C-section in the 4 weeks before ("active" surgeons). I do this to avoid using surgeons who have spent sometime practicing in another institution and whose recent experience I cannot observe.[20] The final sample has 2982 births performed by 59 surgeons. Importantly, the empirical results are robust to the sample restriction on days from previous C-section.

Table 1 summarizes the variables used in the analysis. Mean admission to NICU was approximately 21%—including both intensive and sub-intensive units—and mean low Apgar score was 11%. As expected, emergently admitted patients have a higher probability of both having a low Apgar score and of being admitted to NICU than elective patients. The average age of patients is 34.5, and about 41% of them are first-time mothers—although this number is higher (49%) for non-elective procedures. About 21% of all babies are born with less than 37 weeks of gestation.

TABLE 1 Summary statistics

	All CS		Elective CS		Non-elective CS	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Outcomes						
% NICU	21.0	40.7	18.0	38.4	23.3	42.3
% Apgar score<9	11.4	31.7	7.5	26.3	14.4	35.1
Provider characteristics						
(Mean) CS in past 4 weeks	3.0	1.9	2.7	1.8	3.2	1.9
(Mean) days since last CS	8.4	7.9	9.5	8.1	7.6	7.6
Patient characteristics						
(Mean) age	34.5	5.5	35.3	5.4	33.9	5.5
% University degree	31.4	46.4	33.2	47.1	30.0	45.8
% First-time mothers	41.1	49.2	31.0	46.3	49.0	50.0
Pregnancy characteristics						
% Male	51.9	50.0	51.2	50.0	52.5	50.0
(Mean) weight in grams	2993.3	751.4	2989.9	650.7	2995.9	820.7
% Low birthweight (<2500gr)	20.7	40.5	19.9	39.9	21.3	40.9
(Mean) weeks of gestation	37.8	3.0	37.7	2.1	37.8	3.5
% Preterm (<37 wofg)	20.5	40.3	19.1	39.3	21.5	41.1
% with at-least 1 ER visit	19.5	39.6	22.4	41.7	17.2	37.8
Observations	2982		1298		1684	

Note: Table contains variables used in the empirical analysis for the main estimation sample and the restricted sample of physicians who performed at least one C-section (CS) in the past 4 weeks for the period 2011–2014.

Abbreviation: NICU, neonatal intensive care unit.

The mean number of procedures performed in the previous 4 weeks was 3 for the whole sample. Surgeons performing non-elective C-sections have a slightly higher mean recent experience than those performing elective procedures. Figure 2 shows the frequency distribution of the measure of recent experience for the study sample.

3.3 | Outcomes

The most common outcome (almost exclusively) used in the health economics literature analyzing learning-by-doing and forgetting by hospitals and physicians is the death of the patient—both during and after surgery. As mentioned before, one important drawback of the database used here is the small sample size. Both maternal and fetal deaths are rare events, more so in developed countries, hence there are very few observations experiencing either one of these outcomes (e.g., there are only 12 stillbirths in the study sample). This impedes their use as outcomes for this study. However, one may also argue that mortality alone, being an extreme outcome, is an inadequate measure for capturing the full spectrum of the effects of learning-by-doing on patient health and hospital costs (e.g., morbidity or ordered procedures may also be important outcomes).

The data in hand contains other potential outcomes for patients' health beyond death that can be affected by surgeons skills. As proxies for newborns' health, this study uses the probability of needing to be transferred to a NICU and the probability of having a low APGAR score (at 5 min). The first one measures whether the newborn had to be transferred to a NICU. All else equal, a newborn that is transferred to the NICU is likely to have worse health than one that is not. Furthermore, NICU admissions are among the most expensive treatments in regular hospitals, with one day cost being above $3000. Finally, there are also psychological costs for the parents of the infant. The second outcome is based on a total score of 1 to 10, where the higher the score the better the baby is doing after birth. This test is done to determine whether a newborn needs help breathing or is having heart trouble. Any score lower than 7 is a sign that the baby needs medical attention. In this study, there are only 72 newborns with score below 7. For this reason a new measure was constructed setting the bar higher and all births with a score lower than 9 will be considered of "relatively" lower health. This doesn't

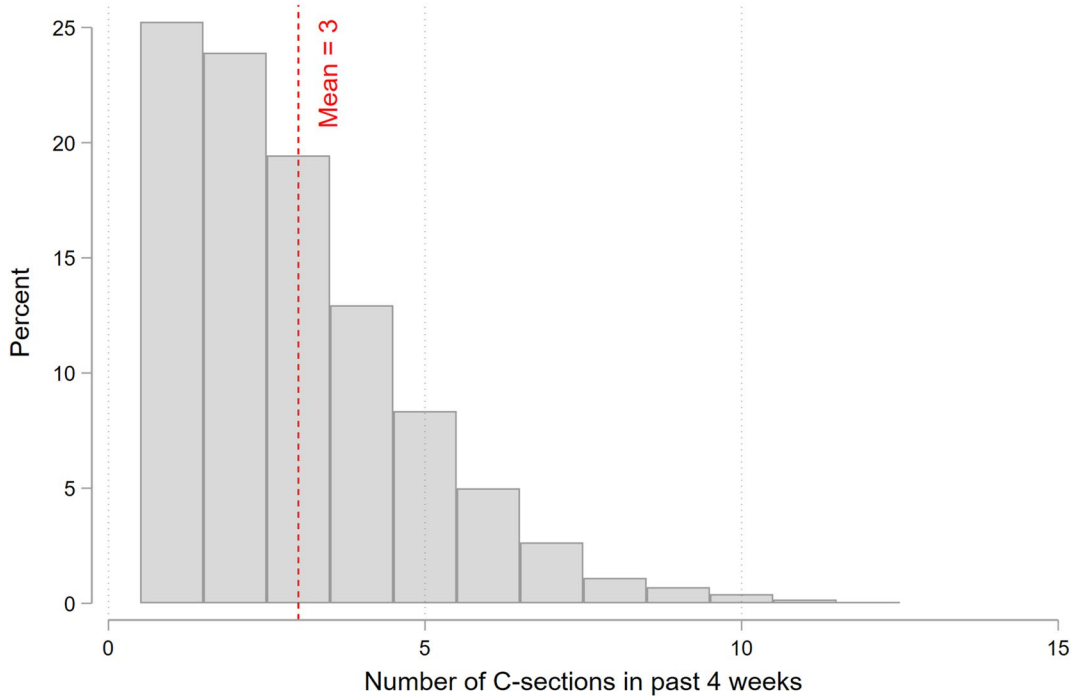

FIGURE 2 Frequency distribution of recent experience. The figure represents the distribution of recent experience measured at 4 weeks. The red dashed-line is the mean average [Colour figure can be viewed at wileyonlinelibrary.com]

necessarily mean a bad score that doctors should act on, but it can be argue that a newborn with an APGAR score below 9 is in relatively worse health condition than a newborn with a score of 9 or 10.

4 | RESULTS

4.1 | The effect of recent practice on patient health

Figure 3 provides the most direct illustration of the effect of experience on outcomes by plotting this bivariate relationship. The dots are 50 equally sized bins plotting the mean of the y-variable (probability of neonatal ICU or probability of an Apgar score below 9) against the mean of the x-variable (surgeon's experience in the last 4 weeks). The dashed line, instead, are the fitted values of a linear regression. The graphs on the left hand-side correspond to elective C-sections, while the ones on the right use only non-elective C-sections. For the elective cases, we cannot observe any clear relationship between experience and either outcome. Instead, when looking at emergent cases, we observe a negative effect of experience on the both outcomes.

Table 2 shows the results of estimating Equation (1) using a linear probability model for each outcome.[21] Panel A shows results using the whole sample, while Panel B and Panel C repeat the exercise for elective and non-elective C-sections. For each outcome, column one shows results of a model with controls, and physician and additive time fixed effects. Column two adds a control for surgeon's number of days since last C-section, and column four adds surgeon's specific time-trends. Standard errors are clustered at the surgeon level in all specifications.

Panel (A) shows that an increase in recent experience is associated with a decrease in the likelihood of a NICU admission. Specifically, one additional C-section in the previous 4 weeks reduces the probability of being transferred to NICU by about 0.8 percentage points. To put the estimate in context, this result implies that a one standard deviation increase in the average number of C-section performed in the last 4 weeks, an increase of two C-sections, is associated with a 1.6 percentage-point reduction in need for NICU, or 7.6% of the sample mean use of NICU. The estimates are consistent and stable across specifications. I fail to find any significant effect on the likelihood of having an Apgar score below 9. Results are qualitatively the same when using the whole sample of cases (including surgeons who have not performed any CS in the last 4 weeks), although coefficients are of a slightly smaller magnitude (see Table A4 in the Appendix). Finally,

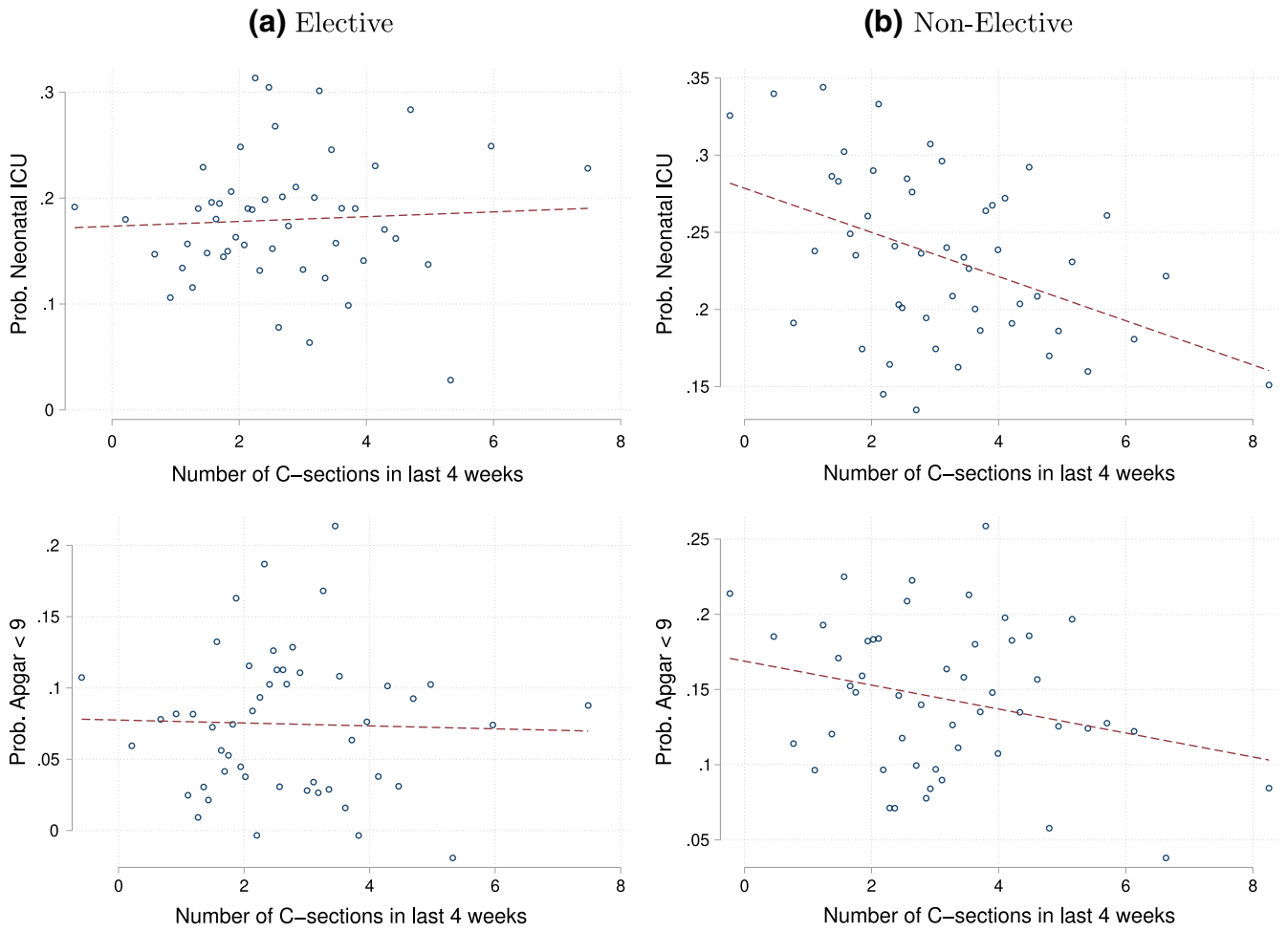

FIGURE 3 Visualization of the relationship between experience and outcomes. (a) Elective, (b) non-elective. The dots are the mean outcome as a function of experience (after controlling for fixed effects for year, month and day-of-the-week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects) using a binned scatter plot with 50 equally sized bins (using "binscatter" in Stata [Stepner, 2013]). The dashed line shows the fitted values of a linear regression. The left hand-side graphs use only the sample of elective C-sections while the right hand-side graphs use the sample of non-elective C-sections [Colour figure can be viewed at wileyonlinelibrary.com]

estimates of the association between days since last C-section and outcome, although sometimes statistically different from zero, are quantitatively very close to zero.[22]

Looking at Panels B and C, the first thing to notice is that the effect of recent experience on outcome is present only for non-elective procedures (as already shown in Figure 3). This is what we would expect since emergent procedures are, on average, more difficult for the surgeon. Estimates for elective surgeries show precisely estimated zero effects of experience on outcome. On the other hand, the effect on emergency procedures are about double the size of those using whole sample. Specifically, a one standard deviation increase in the average number of C-sections performed in the previous 4 weeks, an increase of 1.9 C-sections, is associated with a 3.2 percentage-point reduction in need for NICU, or 13.8% of the sample mean use of NICU. Furthermore, the effect on the likelihood of a low Apgar score are also statistically significant. A one standard deviation increase in recent experience implies a 1.9 percentage-point reduction in the likelihood of having an Apgar score below 9, or 13.2% of the sample mean.

As a robustness check, I use the statistics develop by Oster (2019) on the selection of observables and unobservables to establish the degree of omitted variable bias. Table A5 in the Appendix shows that omitted factors would need to be between 1.3 and 2.3 times more strongly correlated with NICU than all controls accounted for in order to explain the estimated effect of experience on NICU (depending on the assumed R_{max}). Results are qualitatively similar for the probability of having an Apgar score below 9, although smaller correlations between omitted factors and outcome could explain the results.

TABLE 2 Effect of recent experience on birth outcomes

	Neonatal ICU			Apgar < 9		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (A): All C-sections						
Experience (4w)	−0.008***	−0.010***	−0.008**	−0.005	−0.005	−0.005
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Days since last CS		−0.002**	−0.001*		−0.000	−0.001
		(0.001)	(0.001)		(0.001)	(0.001)
Observations	2982	2982	2982	2982	2982	2982
Mean dep.	0.210	0.210	0.210	0.114	0.114	0.114
Panel (B): Elective C-sections						
Experience (4w)	0.004	0.000	0.002	0.001	0.001	0.000
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.006)
Days since last CS		−0.002*	−0.002		0.000	0.000
		(0.001)	(0.001)		(0.001)	(0.001)
Observations	1297	1297	1297	1297	1297	1297
Mean dep.	0.180	0.180	0.180	0.075	0.075	0.075
Panel (C): Non-Elective C-sections						
Experience (4w)	−0.015***	−0.017***	−0.015***	−0.009**	−0.010**	−0.010*
	(0.003)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)
Days since last CS		−0.001	−0.001		−0.001	−0.001
		(0.001)	(0.001)		(0.001)	(0.001)
Observations	1681	1681	1681	1681	1681	1681
Mean dep.	0.234	0.234	0.234	0.144	0.144	0.144
Surgeon trends			x			x

Note: All models contain fixed effects for year, month and day of the week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects. Standard errors, clustered by surgeon, are in parentheses.

Abbreviation: CS, C-section.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

4.2 | Different time-windows of recent experience

A priori, there is no clear criteria to choose a specific time period for my measure of recent experience. If one were to choose a very long period, it could happen that the effect of the further away surgeries have little impact on today's one. On the other hand, choose a period too short and maybe there is not enough variation in the amount of experience. In this paper I decided to measure recent experience within the last 4 weeks. To test how sensitive my results are to this decision, I run a set of regressions for different time spams (from 4 weeks up to 52 weeks) for the two outcomes. Figure 4 shows the results for the case of non-elective C-sections using the full sample of surgeons. For both NICU admission and low Apgar Score, the effect of the number of previous C-sections gets monotonically smaller the longer the measurement period is. This provides further evidence for the human capital depreciation hypothesis, where procedures performed further back in time have little effect on surgeon's ability today—after controlling for her average ability.

5 | DISCUSSION

There is a well established positive association in healthcare between providers' volume and health outcomes, yet our current understanding on the drivers behind this correlation are limited. The two leading explanatory mechanisms are "learning-by-doing" and "selective referral." In this paper I use a feature of the Italian health care system—patients are not allowed to choose physician—to investigate whether there is evidence of "learning-by-doing" in cesarean section

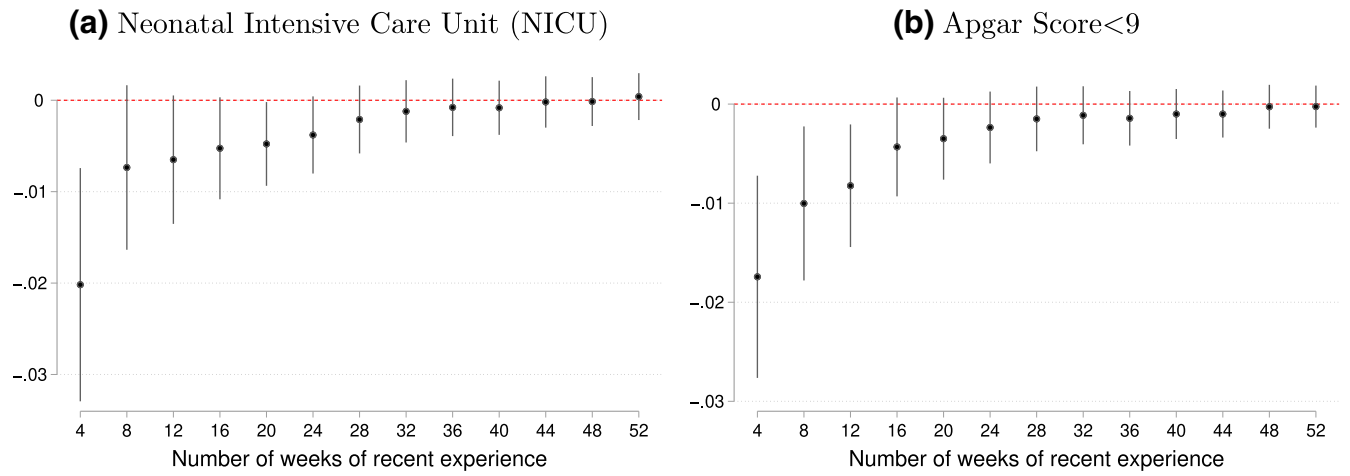

FIGURE 4 Effect of recent experience using different time windows. (a) Neonatal intensive care unit (NICU). (b) Apgar Score<9. Each black dot is the average marginal effect of recent experience using the time window specified in the *x*-axis on the probability of needing NICU or having an Apgar Score below 9 (controlling for year, month and day-of-the-week of birth, and surgeon fixed effects). 90% confidence bounds indicated by the vertical gray lines. In order to keep the sample constant given that I use a lagged independent variable, regressions include only cases for which experience 52 weeks before is not missing [Colour figure can be viewed at wileyonlinelibrary.com]

surgeons. More specifically, I test whether surgeons who have performed more procedures in the recent past observe an improvement in performance. The contribution is threefold: First, my empirical approach rests on an institutional context that allows me to estimate parameters that are free from selective referral bias. Second, I provide evidence that learning-by-doing effects are heterogeneous across procedure types depending on their emergent nature. Finally, I investigate this for C-sections, a procedure that is nowadays very relevant but has been ignored so far in the literature of learning-by-doing.

Using information on birth certificates for one large hospital in Italy between 2011 and 2014, I find that, for emergent cases, a one standard deviation increase in recent experience reduces the likelihood of NICU admission by nearly 3.2 percentage points (13.8%) and of being born with a low Apgar Score by about 1.9 percentage points (13.2%), all else equal. This effect is not present in the case of elective C-sections.

The results of this study would suggest that learning-from-recent-experience effects may be substantive. These findings support recent initiatives to favor higher volume providers when making volume allocation decisions.

ACKNOWLEDGMENTS

I would like to thank the editor, Dr. Jalpa Doshi, and two anonymous referees for their very valuable comments and suggestions. I would like to extend my gratitude to Andrea Ichino and Jérôme Adda for their invaluable support and advice throughout this research project. My gratitude goes to Joseph Doyle and Liertad González as well, their detailed comments have substantially improved this paper. I am extremely grateful to Carlo Dani, Simone Pratesi, Federico Mecacci, Franca Rusconi and Luigi Gagliardi for their clinical expertise, to Tommaso Lanis for performing the confidential merge, and to Francesca Superbi for her assistance in accessing the data. All errors remain my own. I gratefully acknowledge financial support from the General Secretariat for Research-Government of Catalonia (SGR2017-1301) and the Spanish Ministry of Education (PID2019-104619RB-C43).

CONFLICT OF INTEREST

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Maternity ward at the analyzed hospital or the Italian Ministry of Health. Restrictions apply to the availability of these data, which were used under license for this study.

ORCID

Gabriel Facchini https://orcid.org/0000-0002-6382-5486

ENDNOTES

[1] See Halm et al. (2002) and Chowdhury et al. (2007) for a revision of the medical literature.

[2] For a comprehensive analysis of the different channels that can explain the association between volume and outcome in health care see Huesch and Sakakibara (2009).

[3] This type of volume-outcome relationship has been called learning-from-recent-experience by Huesch and Sakakibara (2009).

[4] The Apgar score is a method used to quickly summarize the health of newborn children. The Apgar scale is determined by evaluating the newborn baby on five simple criteria on a scale from zero to two, then summing up the five values thus obtained. The resulting Apgar score ranges from zero to 10.

[5] In the US alone, in 2011 there were almost 1.3 million C-sections (Pfuntner et al., 2013).

[6] Anecdotal evidence: Healthy Debate-Canada, Womens Enews-US, The Guardian-UK.

[7] Although they don't have data on surgeons, Avdic et al. (2019) suggest this is the main mechanism through a process of eliminating other possible alternatives.

[8] A related subset of papers find evidence of substantive human capital depreciation (forgetting) in the medical sector (David & Brachet, 2011; Gaynor et al., 2005; Gowrisankaran et al., 2006; Hockenberry & Helmchen, 2014; Huesch, 2009).

[9] For the data in hand, only a few dozen cases opted to pay, and are consequentially dropped from the study sample.

[10] For an extensive review of the literature on PID in the maternity ward setting see Allin et al. (2015).

[11] This measure is more precise than fixed calendar year as it responds instantaneously to any changes in the recent experience profile. I test for different time windows for recent experience below, from 4 to 52 weeks, and find complete forgetting beyond the last 4 weeks.

[12] Hockenberry and Helmchen (2014) show that a surgeon's number of days since last CABG is positively associated with patients' mortality rate. Since, by construction, temporal distance and surgeon recent volume will be correlated, it may carry a problem of multicollinearity. In my sample, the correlation between these two is −0.40. Furthermore, I show that estimates from models with and without temporal distance are nearly identical.

[13] These include: a quadratic term for mother's age, a dummy for whether the mother has a university degree, a dummy for whether this is her first pregnancy, a dummy for whether the infant is a male, a quadratic term for the number of gestational weeks, a dummy for whether the baby is born with low weight (less than 2500 g), and a dummy for whether the mother had at least one emergency check up during pregnancy (ER visit).

[14] The remaining variation in recent experience comes from career growth (residents perform more procedures as they gain experience), holidays, sick-days, research leaves (e.g., conferences), and natural fluctuations in patients' demand.

[15] Table A1 in the Appendix shows the estimated coefficients. Importantly, coefficients are small in size relative to the mean (at the bottom of the table) and not statistically different from zero.

[16] The opposite case were less active surgeons are assigned patients with worse health conditions is less likely to occur. Under this scenario, my estimates would be an upper bound of the true effect.

[17] One important factor for the success of emergent C-sections is the timing of the cut in relation to the contractions since most of these patients are already in active labor at the time of the surgery.

[18] Hospital and national statistics were obtained from ARS Toscana (2014) and Ministero della Salut (2012).

[19] The data in hand encompasses only births, hence I am blind to any other activities gynecologists may perform when not doing C-sections. Other surgeries include removal of the uterus, tubes and ovaries in the case of tumors, removal of ovarian cysts, removal of uterine fibroids, removal of "pathologic" tissue in endometriosis, treatment of ectopic pregnancies (i.e., where the fetus develops out from the uterus), and more. Hockenberry and Helmchen (2014) utilize two measures of temporal break, time since the last CABG performed and time since any surgical procedure, and find that the last one affects patient outcomes substantially more than the procedure-specific measure. If that were the case also for surgeons performing C-sections, the estimates reported below would be biased towards zero and constitute a lower bound of the true effect.

[20] Although physicians work exclusively in one hospital at a time, they may take leaves to practice in other institutions. In fact, the distribution of days since last C-section for the whole sample is highly positively skewed, with some surgeons showing gaps of more than 100 days (see Figure A1a in the Appendix). As a result, these surgeons have lower recent experience in the hospital under analysis. However, it would be wrong to assume that these surgeons have not been performing somewhere else.

[21] In Table A3 in the Appendix I test for a non-linear relationship between experience and outcomes. As observed in Figure 3, the linear model is the one that better fits the data, with the lowest Akaike information criterion (AIC). Table A2 in the Appendix estimates the main model using a seemingly unrelated regressions estimator to take into consideration the potential correlation between NICU and Apgar score. Results are virtually the same.

[22] Furthermore, in auxiliary regressions where recent experience is not included the estimates are not statistically significant.

REFERENCES

Alexander, D. (2013). *Does physician compensation impact procedure choice and patient health?* (Working Papers 1475). Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Health and Wellbeing.

Allin, S., Baker, M., Isabelle, M., & Stabile, M. (2015). *Physician incentives and the rise in c-sections: Evidence from Canada*. Technical report. National Bureau of Economic Research.

American College of Obstetricians and Gynecologists. (2010). Faqs: Cesarean birth. Technical report. http://www.acog.org//media/For%20Patients/faq006.pdf?dmc=1&ts=20120731T1617495597

ARS della Toscana. (2014). *Nascere in Toscana 2012*. Osservatorio di Epidemiologia dell'Agenzia Regionale di Sanità (ARS). Technical report. https://www.ars.toscana.it/files/pubblicazioni/in_cifre/2_Incifre_CAP_web.pdf

Avdic, D., Lundborg, P., & Vikström, J. (2019). Estimating returns to hospital volume: Evidence from advanced cancer surgery. *Journal of Health Economics*, *63*, 81–99.

Bertoli, P., & Grembi, V. (2019). Malpractice risk and medical treatment selection. *Journal of Public Economics*, *174*, 22–35.

Birkmeyer, J. D., Stukel, T. A., Siewers, A. E., Goodney, P. P., Wennberg, D. E., & Lucas, F. L. (2003). Surgeon volume and operative mortality in the United States. *New England Journal of Medicine*, *349*(22), 2117–2127.

Chowdhury, M., Dagash, H., & Pierro, A. (2007). A systematic review of the impact of volume of surgery and specialization on patient outcome. *British Journal of Surgery: Incorporating European Journal of Surgery and Swiss Surgery*, *94*(2), 145–161.

Contreras, J. M., Kim, B., & Tristao, I. M. (2011). Does doctors' experience matter in lasik surgeries? *Health Economics*, *20*(6), 699–722.

Currie, J., & MacLeod, W. B. (2008). First do no harm? Tort reform and birth outcomes. *Quarterly Journal of Economics*, *123*(2).

David, G., & Brachet, T. (2011). On the determinants of organizational forgetting. *American Economic Journal: Microeconomics*, *3*(3), 100–123.

Dranove, D., & Watanabe, Y. (2010). Influence and deterrence: How obstetricians respond to litigation against themselves and their colleagues. *American Law and Economics Review*, *12*(1), 69–94.

Facchini, G. (2020). *Low staffing in the maternity ward: Keep calm and call the surgeon* (Working Paper 20.09). Department of Applied Economics - Autonomous University of Barcelona.

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

Gowrisankaran, G., Ho, V., & Town, R. J. (2006). *Causality, learning and forgetting in surgery*. Mimeo.

Gruber, J., Kim, J., & Mayzlin, D. (1999). Physician fees and procedure intensity: The case of cesarean delivery. *Journal of Health Economics*, *18*(4), 473–490.

Gruber, J., & Owings, M. (1996). Physician financial incentives and cesarean section delivery. *The RAND Journal of Economics*, *27*, 99–123.

Halm, E. A., Lee, C., & Chassin, M. R. (2002). Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Annals of Internal Medicine*, *137*(6), 511–520.

Ho, V. (2002). Learning and the evolution of medical technologies: The diffusion of coronary angioplasty. *Journal of Health Economics*, *21*(5), 873–885.

Ho, V. (2014). Learning by doing. In A. J. Culyer (Ed.), *Encyclopedia of health economics* (pp. 141–145). Elsevier. http://www.sciencedirect.com/science/article/pii/B978012375678701110X

Hockenberry, J. M., & Helmchen, L. A. (2014). The nature of surgeon human capital depreciation. *Journal of Health Economics*, *37*, 70–80.

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

Huesch, M. D. (2009). Learning by doing, scale effects, or neither? Cardiac surgeons after residency. *Health Services Research*, *44*(6), 1960–1982.

Huesch, M. D., & Sakakibara, M. (2009). Forgetting the learning curve for a moment: How much performance is unrelated to own experience? *Health Economics*, *18*(7), 855–862.

Lefèvre, M. (2014). Physician induced demand for c-sections: Does the convenience incentive matter? *Health, Econometrics and Data Group (HEDG) Working Papers*, *14/08*, 1–22.

Lucas, D., Yentis, S., Kinsella, S., Holdcroft, A., May, A., Wee, M., & Robinson, P. (2000). Urgency of caesarean section: A new classification. *Journal of the Royal Society of Medicine*, *93*(7), 346–350.

Luft, H. S., Bunker, J. P., & Enthoven, A. C. (1979). Should operations be regionalized? The empirical relation between surgical volume and mortality. *New England Journal of Medicine*, *301*(25), 1364–1369.

Luft, H. S., Hunt, S. S., & Maerki, S. C. (1987). The volume-outcome relationship: Practice-makes-perfect or selective-referral patterns? *Health Services Research*, *22*(2), 157.

Ministero della Salut. (2012). *Certificato di assistenza al parto (cedap). Analisi dell'evento nascita – anno 2012*. Technical report. http://www.salute.gov.it/imgs/C_17_pubblicazioni_2768_allegato.pdf

Ministero della Salute. (2019). *Certificato di assistenza al parto (cedap): Analisi dell'evento nascita – anno 2016*. Technical report. http://www.salute.gov.it/imgs/C_17_pubblicazioni_2881_allegato.pdf

Oster, E. (2013). *Psacalc: Stata module to calculate treatment effects and relative degree of selection under proportional selection of observables and unobservables*. Boston College Department of Economics.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204.

Pfuntner, A., Wier, L., & Stocks, C. (2013). *Most frequent procedures performed in us hospitals, 2011: Statistical brief# 165. Healthcare Cost and Utilization Project (HCUP) statistical briefs*.

Ramanarayanan, S. (2008). *Does practice make perfect: An empirical analysis of learning-by-doing in cardiac surgery*. Mimeo.

Sfekas, A. (2009). Learning, forgetting, and hospital quality: An empirical analysis of cardiac procedures in Maryland and Arizona. *Health Economics*, *18*(6), 697–711.

Stepner, M. (2013). *"BINSCATTER: Stata module to generate binned scatterplots," Statistical Software Components*. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457709.html

WHO, UNICEF, UNFPA, WBG and UNDP. (2019). *Trends in maternal mortality: 2000 to 2017: Estimates by WHO, UNICEF, UNFPA*. World Bank Group, and United Nations Population Division. Technical report. https://www.unfpa.org/sites/default/files/pub-pdf/Maternal_mortality_report.pdf

World Health Organization. (2000). *The world health report 2000: Health systems: Improving performance*. World Health Organization.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Facchini, G. (2022). Forgetting-by-not-doing: The case of surgeons and cesarean sections. *Health Economics*, *31*(3), 481–495. https://doi.org/10.1002/hec.4460

APPENDIX A: OTHER GRAPHS AND TABLES

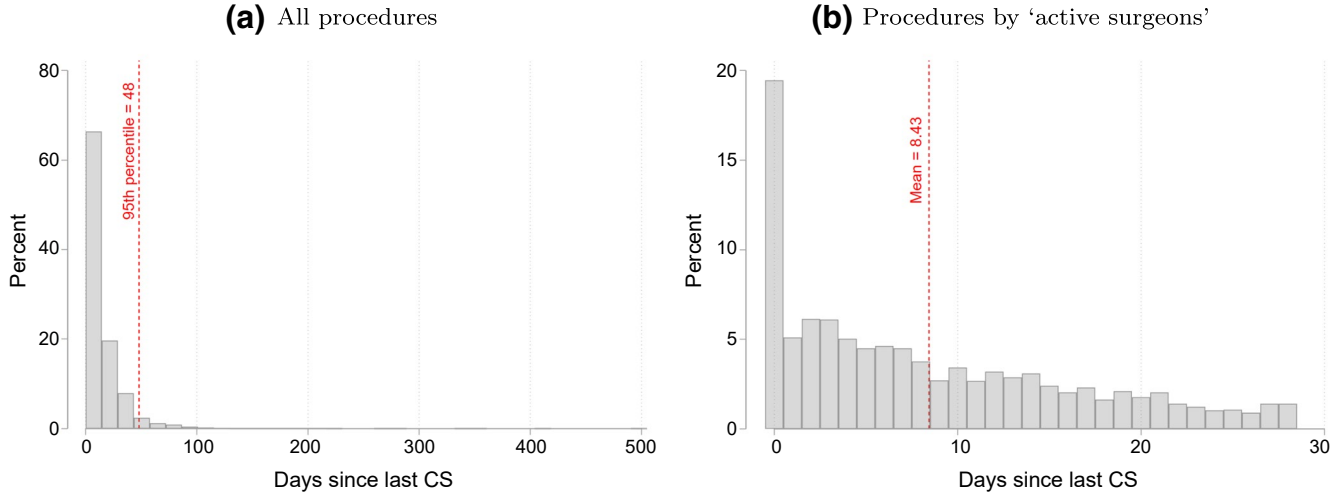

FIGURE A1 Frequency distribution of days since last CS. (a) All procedures (b) Procedures by "active surgeons." The figure represents the distribution of days since previous C-section for two samples. The left hand-side panel uses the whole sample without restrictions on the number of days. The right hand-side panel uses the study sample which is restricted to surgeons performing at least one C-section in the previous 4 weeks [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE A1 Balanced pre-treatment characteristics

	Mother's Age>34	University degree	First-time mother	Male newborn	Low weight at birth	Preterm birth	ER visit
Experience (4w)	0.008	0.007	0.004	−0.000	0.002	−0.004	0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Observations	3466	3466	3466	3466	3466	3466	3466
Mean dep.	0.53	0.31	0.41	0.52	0.21	0.21	0.20

Note: Table contains the coefficients and standard errors from separate regressions of each predetermined variable on the treatment (experience), controlling for surgeon, day-of-the-week, month, and year of birth fixed effects. Standard errors, clustered by surgeon, are in parentheses.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

TABLE A2 Ordinary least squares (OLS) vs seemingly unrelated regressions (SUR)

	OLS		SUR	
	NICU	**Apgar<9**	**NICU**	**Apgar<9**
Experience (4w)	−0.016***	−0.009*	−0.016***	−0.009*
	(0.003)	(0.005)	(0.003)	(0.005)
Days since last CS	−0.001	−0.001	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)

Note: The first two column assumes errors are independent across outcomes, while the last two allow the errors to be correlated across equations. All models contain fixed effects for year, month and day of the week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects. Standard errors, clustered by surgeon, are in parentheses.

Abbreviation: NICU, neonatal intensive care unit.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

TABLE A3 Alternative model specifications

	Neonatal ICU				Apgar < 9			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Experience	−0.017***	−0.017	−0.039		−0.010**	0.002	0.033	
	(0.003)	(0.011)	(0.029)		(0.005)	(0.016)	(0.040)	
Exp. square		0.000	0.006			−0.001	−0.009	
		(0.001)	(0.007)			(0.002)	(0.009)	
Exp. cubic			−0.000				0.001	
			(0.000)				(0.001)	
33-66th percentile				−0.039**				0.022
				(0.016)				(0.029)
>66th percentile				−0.058***				−0.032
				(0.015)				(0.023)
Observations	1681	1681	1681	1681	1681	1681	1681	1681
AIC	591.3	593.3	594.7	596.8	604.9	606.0	606.8	604.6

Note: All models contain fixed effects for year, month and day of the week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects. Standard errors, clustered by surgeon, are in parentheses.

Abbreviations: AIC, Akaike information criterion; ICU, intensive care unit.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

TABLE A4 Effect of experience on outcomes for the whole sample of surgeons

	All C-sections		Elective C-sections		Non-elective C-sections	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	NICU	**Apgar<9**	**NICU**	**Apgar<9**	**NICU**	**Apgar<9**
Experience (4w)	−0.008***	−0.002	0.002	0.002	−0.016***	−0.008*
	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
Days since last CS	−0.001**	−0.000*	−0.000	0.000	−0.001	−0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Observations	3466	3466	1576	1576	1889	1889
Mean dep.	0.208	0.110	0.176	0.076	0.234	0.139

Note: All models contain fixed effects for year, month and day of the week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects. Standard errors, clustered by surgeon, are in parentheses.

Abbreviation: NICU, neonatal intensive care unit.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

TABLE A5 Selection on unobservables

	(1) NICU	(2) Apgar<9
Experience	−0.016***	−0.009*
	(0.003)	(0.005)
Observations	1684	1684
R^2	0.54	0.32
Delta (R max = 1)	1.32	0.43
Delta (R max = 1.5*R^2)	2.32	1.85

Note: Delta is the coefficient of proportionality (δ), which estimates how big the selection on unobservables has to be relative to the selection on observables for the true effect to be zero. Two different values for Delta are estimated using two different R_{max}: 1 and 1.5 times the R^2 from the baseline regression. Delta estimates are obtained using the Stata command *psacalc* (Oster, 2013). The sample is restricted to non-elective C-sections. All models contain fixed effects for year, month and day of the week of birth, mother and pregnancy controls mentioned in Section 3.1 and surgeon fixed effects. Standard errors, clustered by surgeon, are in parentheses.

Abbreviation: NICU, neonatal intensive care unit.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.