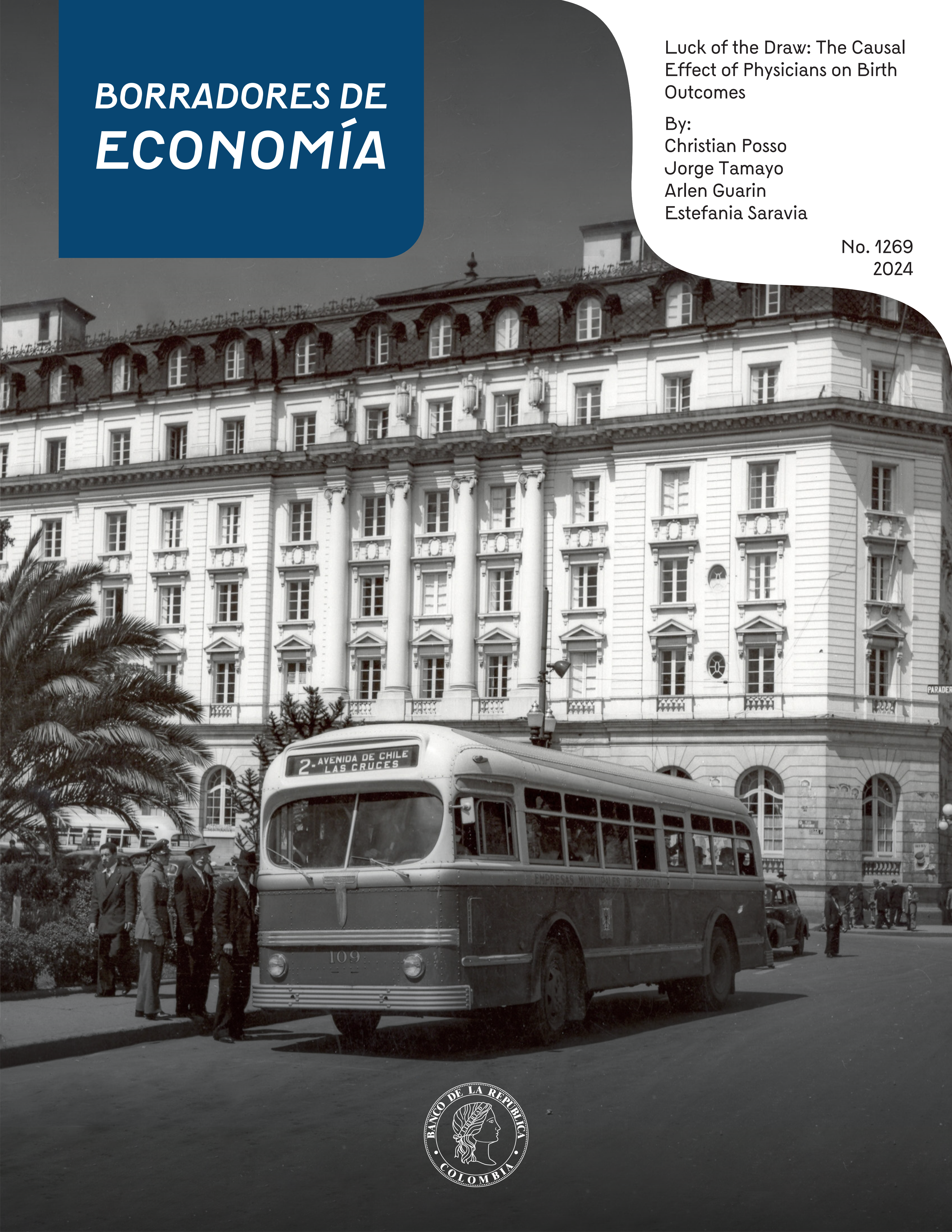
BORRADORES DE ECONOMÍA

Luck of the Draw: The Causal Effect of Physicians on Birth Outcomes

By:
Christian Posso
Jorge Tamayo
Arlen Guarin
Estefania Saravia

No. 1269
2024



BANCO DE LA REPUBLICA
COLOMBIA

Luck of the Draw:
The Causal Effect of Physicians on Birth Outcomes

Christian Posso[1], Jorge Tamayo[2], Arlen Guarin[3], and Estefania Saravia[*4]

[1]Banco de la República (cpossosu@banrep.gov.co)
[2]Harvard University, Harvard Business School (jtamayo@hbs.edu)
[3]Development Impact Evaluation, World Bank (aguarin@worldbank.org)
[4]University of California, Los Angeles (esaravia@ucla.edu)

Abstract

Identifying the impact of physicians on health outcomes is a challenging task due to the nonrandom sorting between physicians, hospitals, and patients. We overcome this challenge by exploiting a Colombian government program that randomly assigned 2,126 physicians to 618 small hospitals. We estimate the impact on the 256,806 children whose mothers received care in those hospitals during their pregnancy, using administrative data from the program, hospitals' vital statistics records, and physicians' records from mandatory health-specific graduation exams. We find that more-skilled physicians improve health at birth outcomes. That is, being assigned a physician with a one standard deviation higher performance in the health graduation exam scores decreases the probability of giving birth to an unhealthy baby by 6.31 percent. We present evidence that an underlying mechanism includes improving the targeting of care toward the more vulnerable mothers.

Keywords: Physicians' health skills, health at birth outcomes, experimental evidence.

JEL Classification: H51, I14, I15, I18.

[*]The authors' names are listed in random order. We thank Manuela Cardona, Silvia Granados, Nicolas Mancera, Brayan Pineda, Daniel Marquez, Gabriel Suarez, Santiago Velasquez, and Carolina Velez for excellent research assistance. We thank Achyuta Adhvaryu, Maria Aristizabal, Carolina Arteaga, Francesco Bogliacino, Leonardo Bonilla, David Card, Maíra Coube, Janet Currie, Kaveh Danesh, Margarita Gafaro, Robert Gonzalez, Hilary Hoynes, Raymond Kluender, Rem Koning, Juliana Londoño-Vélez, Edward Miguel, Anant Nyshadham, Paul Rodriguez, Emmanuel Saez, Molly Schnell, Benjamin Scuderi, Jesse Shapiro, Christopher Walters, Danny Yagan and seminar participants at Banco de la Republica, Harvard, NBER SI Children, ESSEN Health Conference 2020, EEA 2020, Eafit, RIDGE LACEA's Health Economics Network, Universidad del Rosario, UC Berkeley (Development and Labor Lunch), EALE 2022, American-European Health Economics Study Group Barcelona 2022, DIME-KDIS conference 2023 for insightful comments. Arlen gratefully acknowledges financial support from the UC Berkeley Opportunity Lab.

La suerte del sorteo:
El Efecto Causal de los Médicos en la Salud al Nacer

Christian Posso, Jorge Tamayo, Arlen Guarín, Estefanía Saravia

Resumen

La asignación de médicos, hospitales y pacientes no suele ser aleatorio, por lo que identificar la importancia de los médicos en la salud de sus pacientes no resulta trivial. En este documento aprovechamos la asignación aleatoria de 2.126 médicos a 618 hospitales pequeños, que se realizó en Colombia a través del Servicio Social Obligatorio (SSO), para estimar el impacto de las habilidades de los médicos en la salud al nacer de los 256.806 bebés cuyas madres recibieron atención en dichos hospitales durante su embarazo. Para este análisis, utilizamos datos administrativos del programa SSO de los años 2013 y 2014, registros de estadísticas vitales y los puntajes en la prueba de medicina de la prueba estandarizada Saber Pro. El resultado principal muestra que los médicos más hábiles mejoran la salud de los niños al nacer. Específicamente, la asignación de un médico que obtuvo un puntaje en la prueba de medicina (Saber Pro) una desviación estándar superior a la media, disminuye la probabilidad de dar a luz a un bebé no sano en un 6,31 %. Adicionalmente, encontramos que este resultado es explicado por una mejor focalización de los controles prenatales hacía las madres más vulnerables.

Palabras clave: Habilidades de los médicos, salud al nacer, evidencia de experimentos aleatorios.

Clasificación JEL: H51, I14, I15, I18.

1 Introduction

Origins of inequality can be found as early as the nine months that infants are in utero. These critical months shape children's endowments at birth, which have been shown to predict future abilities and health trajectories that genetics cannot explain (Almond, Chay, and Lee, 2005; Currie, 2011; Currie and Almond, 2011). While much research has focused on factors such as maternal health and families' socioeconomic conditions (Currie, 2011; Currie and Schwandt, 2016b), recent evidence suggests that healthcare providers may also play a significant role in determining these outcomes (Okeke, 2023). This evidence has shown that physicians have differential effects on infants compared to other substantially less trained practitioners. Nonetheless, a remaining question is if similarly trained physicians have differential effects on infants' health outcomes.

In this paper, we break new ground by providing causal evidence on the role that skilled physicians play in newborns' health at birth. Physicians significantly impact patients' health (Chan Jr, Gentzkow, and Yu, 2019; Chen, 2021; Currie and MacLeod, 2017,2; Das and Hammer, 2005). Under the hypothesis that more-skilled physicians have differential impacts on the health of infants relative to less-skilled physicians, policies that improve medical skills are a mechanism to boost both the health of infants and their future outcomes.[1]

The lack of causal evidence regarding the impact of skilled physicians on birth outcomes is not surprising since answering this question poses substantial empirical challenges: it requires both accounting for the selection bias associated with the match between physicians and hospitals (Doyle, Ewer, and Wagner, 2010) and measures of physicians' medical skills.[2] We overcome this challenge by exploiting a policy experiment conducted in Colombia. In this national-level program, 2,126 recently graduated physicians were randomly assigned to 618 Local Health Centers (LHCs) through a within-state random match, a setting that enables us to bypass the selection bias issue and isolate the impact of physicians on infant health at birth.[3] Further, we match all physicians with the scores of the mandatory health-specific physicians' graduation exams they took just before graduating, a measure we use as a proxy for medical skills.[4]

[1]There is a large literature that shows that poor health at birth has long-lasting adverse impacts on future outcomes (and the outcomes of the next generation) such as earnings, education, and disability (Adhvaryu et al., 2018; Almond et al., 2018; Currie, 2011; Persson and Rossin-Slater, 2018).

[2]There is an extensive literature on positive assortative matching (PAM) that shows that companies and high-productivity workers match together (for example, Abowd, Kramarz, and Margolis, 1999; Becker, 1973; Kremer, 1993; Roy, 1951; Shimer and Smith, 2000; Woodcock, 2008).

[3]While these LHCs are equipped to provide primary care and emergency attention, and outpatient and inpatient care, including childbirth, they are typically referred to as hospitals despite their relatively smaller size and lower capacity compared to hospitals in more developed urban settings.

[4]Throughout the paper we will use the words skilled, higher quality, and knowledgeable, interchangeably to refer to physicians with a higher score in the graduation exams.

Several features of our context are conducive to accomplishing our study's goals. First, Colombian regulations mandate that medical school graduates must dedicate the first year of their career to the National Mandatory Social Service (SSO, as per its acronym in Spanish). This program randomly assigns new physicians to LHCs across the state where they apply. Thus, LHCs and physicians are assigned without regard to their characteristics, so physicians with different skills are expected to face a similar set of facilities, administrative resources, and health staff. Then, by comparing across LHCs, we can estimate the causal effect of more skilled physicians on newborns' health outcomes.

Second, we combine several rich and granular administrative records in Colombia, allowing us to observe the LHC where physicians were assigned, a proxy for their skills, and hospital outcomes as performance measures. Specifically, we collect data on the reports published by Colombia's Ministry of Health after the SSO lotteries that took place between 2013 and the third quarter of 2014. Further, we use individual records from mandatory health-specific physicians' graduation exams to identify the more skilled, recently graduated physicians. Finally, we link the LHCs to which doctors were randomly assigned to the National Vital Statistics Records (VSR), from which we obtain birth outcomes and maternal sociodemographic characteristics.

Our main finding is that skilled physicians positively affect birth outcomes. We find that being assigned a physician with one standard deviation higher performance in the medical graduation exam scores decreases the probability of giving birth to an *unhealthy* baby by 6.31%. A child is defined as unhealthy if at least one of these three conditions is satisfied: has a low birth weight, was an early-term infant (prematurity), or has a low Apgar score. These effects are consistent across each health measure at birth: we find a negative impact of 7.71% on low birth weight, 7.97% on prematurity, and a 7.16% decrease in the probability of low Apgar scores.[5,6] Our results are similar to the findings

[5]Low birth weight has been one of the key measures of health at birth studied in the literature (Currie, 2011). According to WHO (2016), Almond et al. (2005) and Gonzalez and Gilleskie (2017), prematurity is highly correlated with low birth weight and mortality. The Apgar score has also been used in the literature as an indicator of health at birth; see for example, Almond, Doyle Jr, Kowalski, and Williams (2010) and Lin (2009).

[6]Unfortunately, during our analysis period, we cannot test the impact of physicians on mortality due to data issues. Firstly, the variable indicating the number of gestation weeks, available in birth records, is missing for a significant portion of the fetal and neonatal deaths. This omission prevents us from determining the gestation period's start for these deaths, thereby hindering our ability to precisely identify exposure to physician cohorts as we can for births. Additionally, fetal and neonatal records frequently lack information for the LHC, as well as for mothers' and children's covariates. Given these limitations, we conduct a cohort-level exercise rather than a child level exercise. We quantify the number of fetal and combined fetal plus neonatal deaths during the time the cohort was assigned to a LHC. This quantification is regardless of the gestation period's length of exposure to the cohort and is based on the data with all aforementioned limitations. While these results are expected to be subject to measurement error attenuation bias, we still observe a negative, albeit not statistically significant, point estimate, which goes in line with our main results.

in shared work experience between surgeons and healthcare physicians (Chen, 2021) or eligibility for Medicaid for pregnant women (Currie and Gruber, 1996).[7]

Focusing on test scores may overstate the importance of this proxy as a benchmark for physicians' skills and quality of performance while understating the relevance of other correlated physicians' characteristics. We take advantage of the random assignment to get an unbiased estimate of the physicians' relative value-added (VA) (Angrist, Hull, Pathak, and Walters, 2017; Chetty, Friedman, and Rockoff, 2014; Kane and Staiger, 2008). Following Jackson (2018) and Fletcher, Horwitz, and Bradley (2014), we show that the standard deviation of the shrunken VA is 0.06, so, for instance, having a physician with VA at the 85th versus 50th percentile would decrease the probability of being unhealthy by roughly 0.06 standard deviations. Using our VA estimates, we study the relationship between the physicians' VA and several observable characteristics, such as the scores of the mandatory health-specific physicians' graduation exams (our skill measure), some proxies for the quality of the medicine program attended, family socioeconomic characteristics, and gender. The results suggest that test scores are the variables with the highest predictive power of the physicians' performance measured as the relative value-added. In contrast, the other characteristics have no significant relationship.

To shed light on the potential mechanisms through which physicians impact a child's outcomes, we first analyze several heterogeneous effects across different mothers' characteristics. Although the effects are slightly more pronounced among first-time mothers, teenage mothers, mothers with low education, and single women, the differences between groups are not statistically different. We then estimate effects separately for male and female newborns. One commonly observed phenomenon in literature is that males tend to be more susceptible to health shocks during the prenatal period compared to females (Eriksson, Kajantie, Osmond, Thornburg, and Barker, 2010; Kraemer, 2000; Naeye, Burt, Wright, Blanc, and Tatter, 1971; Pongou, Kuate Defo, and Tsala Dimbuene, 2017). To the extent that more-skilled physicians improve children's health outcomes at birth, they may help mitigate such adverse shocks in utero. The reduction in the probability of giving birth to an unhealthy child is particularly pronounced among male newborns, but it is not statistically different for male and female newborns.

Furthermore, we study heterogeneous effects between LHCs with high and low incidences of poor newborn health using ex-ante LHC-level health measures. The effects on the probability of giving birth to an unhealthy baby are larger for LHCs with a high

[7] Chen (2021) found that a one standard deviation increase in shared work experience between surgeons and health care physicians reduced patients' 30-day mortality rates by 0.6 and 1.2 percentage points. Currie and Gruber (1996) found that a one standard deviation increase in the eligibility for Medicaid for pregnant women reduced by 2.1 percent the probability of low birth weight and by 9.35 percent the infant mortality rate.

incidence of poor newborn health, which we define as the LHCs in the top quartile of the unhealthy babies' baseline incidence distribution.

We next explore a mechanism through which physicians may improve health at birth: prenatal consultations. According to WHO (2016) and the Colombian government (Gomez, Arevalo, et al., 2013), better and more frequent prenatal care during pregnancy can improve mother's and their newborn's health.[8] We follow the standard recommendations by WHO (2016) in 2013, the first year of the records that we use, and define "adequate prenatal care" as having at least four visits to the doctor during pregnancy.[9] We find that more-skilled doctors, on average, do not schedule more prenatal checkups.[10]

We then test whether the more skilled physicians target prenatal consultations toward the most vulnerable mothers, measured as those with a higher predicted likelihood of giving birth to an unhealthy baby. We use several machine learning techniques to generate two groups of predictions of the probability of giving birth to an unhealthy baby using a set of LHC and mother characteristics, such as indicators of first-time mothers or teenage mothers, that are usually salient to physicians at the time of prenatal care. The results show that regardless of the method we use to predict unhealthiness, more-skilled doctors do not significantly increase the probability of having at least the suggested number of antenatal consultations for mothers with a low predicted probability of giving birth to an unhealthy child. However, physicians do seem to target those prenatal checkups toward more vulnerable mothers, measured as mothers with a higher predicted probability of giving birth to an unhealthy child. Consistent with prenatal care being one of the channels through which more-skilled physicians have an impact on children, we show that the effects of more-skilled physicians on birth outcomes (unhealthy, low birth weight, prematurity, and Apgar score) are particularly pronounced among mothers with an ex-ante high predicted probability of giving birth to an unhealthy child. Altogether, these results are consistent with physicians being time-constrained and unable to increase the average number of prenatal consultations but improving the targeting of care toward the more vulnerable mothers.

To assess the internal validity of our identification strategy, we implement two tests.

[8]This is due to the fact that during a prenatal checkup, pregnant women are screened and treated for risk of complications, avoiding preterm births, and other problems. Also, pregnant women are given critical information on nutrition, diet, and other general mother and child safety practices, which have been shown to play a crucial role in utero infant growth (Amarante, Manacorda, Miguel, and Vigorito, 2016; Kramer, 1987). Furthermore, in Colombia, the Ministry of Health requires that physicians carry out prenatal checkups (Gomez et al., 2013); as such, physicians are responsible for prenatal care, and they are the professionals who attend 98% of deliveries.

[9]The data that we can access only has information on ranges of the number of antenatal consultations, which does not allow for a more flexible usage of this variable. In our sample, 87% of mothers have at least four visits with the doctor.

[10]Carrillo and Feres (2019) found no evidence of increase in prenatal care when physicians were replaced by nurses.

First, we assign a placebo treatment to infants born before the arrival of the physicians in our sample. The random assignments that we use in our main specification took place in 2013 and 2014. We run placebo tests similar to our main specification but using outcomes for the children born in the same LHCs in the period 2009-2012, the four previous years prior the random assignment. We find that the treatment generates precisely estimated zeros. Second, we show evidence on the actual randomness of the assignment by showing that physicians' quality is not correlated with the assigned LHC and municipality's ex-ante characteristics.

This paper contributes to the literature in several ways. Firstly, our identification strategy and the availability of granular administrative records allow us to measure the causal impact of more-skilled physicians on health outcomes. Previous literature has documented the relationships between health outcomes and physicians' diagnosis skills (Currie and MacLeod, 2020), physicians' teams (Chen, 2021), healthcare access (Almond et al., 2010; Finkelstein, Taubman, Wright, Bernstein, Gruber, Neuse, Allen, Baicker, and Group, 2012),[11] healthcare costs (Alsan, Garrick, and Graziani, 2019; Clemens and Gottlieb, 2014; Molitor, 2018), quality of physicians' academic institutions (Doyle et al., 2010), physicians' performance on qualifying examinations (Carrera, Goldman, Joyce, and Sood, 2018; Tamblyn, Abrahamowicz, Dauphinee, Hanley, Norcini, Girard, Grand'Maison, and Brailovsky, 2002; Wenghofer, Klass, Abrahamowicz, Dauphinee, Jacques, Smee, Blackmore, Winslade, Reidel, Bartman, et al., 2009), physicians' competence (Das, Holla, Mohpal, and Muralidharan, 2016)[12], physicians' ability to facilitate adherence to prescription medications (Iizuka, 2012; Simeonova, Skipper, and Thingholm, 2020), physicians' fees and payment for performance (Basinga, Gertler, Binagwaho, Soucat, Sturdy, and Vermeersch, 2011; Ho and Pakes, 2014a,1), general practitioners and specialists (Baicker and Chandra, 2004), and physicians' communication (Curtis, Cai, Wade, Stolshek, Adams, Balasubramanian, Viswanathan, and Kallich, 2013). To our knowledge, this paper is the first to document experimental evidence on the impact of more skilled physicians on health outcomes.

Second, our paper relates to the literature on teacher value-added, where the effect on students from a high-quality (effective) teacher has been proven to be significant (Araujo, Carneiro, Cruz-Aguayo, and Schady, 2016; Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011; Rivkin, Hanushek, and Kain, 2005; Rockoff, 2004). While this literature

[11]See Aron-Dine, Einav, Finkelstein, and Cullen (2015), Bardach, Wang, De Leon, Shih, Boscardin, Goldman, and Dudley (2013), Michalopoulos, Wittenburg, Israel, and Warren (2012), Anderson, Dobkin, and Gross (2012), Anderson, Dobkin, and Gross (2014) for studies related with the effects of healthcare access on population health.

[12]See Das and Hammer (2005), Das and Hammer (2007), Das, Hammer, and Leonard (2008), Das and Sohnesen (2007), Leonard and Masatu (2007), Leonard, Masatu, and Vialou (2007) for literature studying physicians' competence.

estimates that a one standard deviation increase in teacher quality is associated with an increase in students' test scores of 0.19 standard deviations, we find that a physician with one standard deviation higher quality decreases the probability of having a child with an unhealthy condition by 6.3 percent. Our findings highlight the importance of the match between high-quality physicians and the most vulnerable mothers in improving child health outcomes at birth. This result has important implications for marginal optimal assignment and suggests that investing in recruiting and retaining high-quality physicians in areas of greatest need could have significant positive social value, similar to the impact of good teachers on student test scores.

Third, focusing on health at birth is particularly appealing because of its high potential for external validity and relevance. Birth outcomes are usually both objectively measured and almost universally available for healthy and unhealthy children. Endowments at birth have also been proven to be essential for future outcomes —including earnings, education, and health (Adhvaryu, Nyshadham, Molina, and Tamayo, 2018; Currie, 2009; Oreopoulos, Stabile, Walld, and Roos, 2008; Persson and Rossin-Slater, 2018)— and physicians have been shown to play an important role in determining patients' health (Chan Jr et al., 2019; Chen, 2021; Currie and MacLeod, 2017,2; Das and Hammer, 2005; Okeke, 2023), so it is particularly relevant to understand whether physicians can have any effect on children's health at birth.

Finally, we contribute to the literature showing differential effects on the variation in the healthcare personnel expertise. Previous papers have found wide variations in the treatment rates across LHCs due to allocative inefficiencies and variations in treatment expertise (Abaluck, Agha, Kabrhel, Raja, and Venkatesh, 2016; Chandra and Staiger, 2020; Currie and MacLeod, 2017). We benefit from recent advances in machine learning techniques to show that more skilled physicians target prenatal consultations toward mothers with the highest risk of giving birth to an unhealthy child. Our results suggest observable risk factors receive more attention from more skilled physicians.

The remainder of this paper is organized as follows: In Section 2, we describe the Colombian health system and the SSO program, the setting we exploit to identify parameters of interest. Section 3 describes the rich administrative data we derive from the doctors' college exit exams and patients' outcomes at birth. In Section 4, we introduce our empirical strategy, show evidence on the randomness of physicians' assignment to LHCs, and present our main estimated effects. Section 5 discusses the frequency of prescribed prenatal consultations as a potential mechanism for determining physicians' skill levels. Then, after presenting several robustness checks in Section 6, we conclude in Section 7.

2 Institutional Background, Experimental Setting, and Physicians

2.1 Institutional Background

According to the Political Constitution of Colombia of 1991, access to health services is an individual basic right. The principles of the system are based on progressivity and equity in the distribution of subsidies and access to health services (Law 100, Congress of Colombia, 1993). Law 100 of 1993 introduced two types of health insurance: subsidized and contributive. The contributive regime is inclusive of formal employees (and their families) who contribute a fixed share of their employment income to the system. The subsidized regime is inclusive of poor household members who do not have formal employment.[13] By 2011, access to healthcare was close to universal; indeed, even in the poorest population, the coverage was 87%, while in rural areas it was about 88% (Páez, Jaramillo, Franco, and Arregoces, 2007).

One of the main characteristics of high coverage is the greater use of reproductive-health-related services, an essential aspect of reducing risks associated with pregnancy, childbirth, and infant mortality (WHO, 2016). For our period of analysis, the percentage of women with at least four prenatal examinations[14] in Colombia was 87.7%, while the percentages of newborns with low birth weight and prematurity were 8.8% and 9.3%, respectively. Still, the system faces important challenges. In 2017, according to the United Nations Statistics Division database, the neonatal mortality rate (deaths per 1,000 live births) was 7.8 and the infant mortality rate (infant deaths per 1,000 live births) was 12.2.[15]

To become a physician in Colombia, one must study in an undergraduate medical program. Like college programs in nursing, bacteriology, and dentistry, medicine is considered a health program. Students accepted into health programs earn a BA after five to six years of education. According to Colombian law, all professionals graduating from health programs are social servants; directly after graduation, they must provide professional services in urban and rural areas with limited access to health services for one year before practicing as professionals. This service is provided under the Mandatory Social Service (SSO). The current SSO program was created by Law 1164/2007 (Congress of Colombia, 2007), but it was only adopted in 2010 when its implementation was legislated

[13] The eligibility for the subsidized regime is defined by the SISBEN score, a household-level wealth score used to target public program beneficiaries in Colombia.

[14] WHO (2016) defines "adequate prenatal care" as having at least four visits to the doctor during pregnancy.

[15] https://data.un.org/, consulted in May 2020.

by Resolution 1058/2010 (Ministry of Health, 2010). The main objective of the SSO is to improve the quality of and the access to health services in depressed urban and rural areas (or those with difficult access to health services), as well as to stimulate an adequate geographic distribution of human talent in health. The SSO also promotes spaces for the personal and professional development of those beginning their careers in the health sector.[16]

Physicians also play a key role in maternal medicine in the Colombian health system. The Ministry of Health (2013), in resolution 1441 of 2013, states that any physician in Colombia can perform low-complexity surgeries and procedures, including childbirth, C-sections, medical care to newborns, and early detection activities like antenatal consultations. An important characteristic of the Colombian health system is that physicians must carry out prenatal examinations. According to the practical guide for preventing, detecting, and treating pregnancy complications by the Colombian Ministry of Health (Gomez et al., 2013), prenatal visits should be carried out by physicians or nurses specializing in maternal-perinatal care. In fact, calculations from the VSR show that physicians are responsible for all prenatal check-ups, and physicians attend 98% of all deliveries.[17]

2.2 The experimental setting: SSO program

By 2007, as the number of people getting medical training in Colombia increased, the available positions for SSO physicians were fewer than the number of applicants. Therefore, how the applicants were chosen and the LHCs to which they were assigned became one of the program's most critical decisions. Law 1164/2007 (Congress of Colombia, 2007) required that an assignment was to be "guided by the principles of transparency and equal conditions for all applicants." In concordance, Resolution 1058/2010 established that decisions regarding who is selected and for which locations must be made through state-level random draws.

At the end of 2012, a more organized approach to running the random assignments was introduced for the SSO program. The first two years of implementing the new program showed that the directions (Resolution 1058/2010) were not robust enough to guarantee a transparent and organized assignment of physicians. Consequently, Resolution 566/2012 (Ministry of Health, 2012b) was introduced, which mandated that there would be four yearly (state level) waves of SSO draws starting in January 2013. [18] Professionals would choose the state, but be randomly assigned to available positions in that state..

[16] See resolution 1058/2010 (Ministry of Health, 2010).

[17] Nurses who have just graduated from college cannot perform prenatal examinations in Colombia.

[18] Taking place in January, April, July, and October in each of the 32 states in which the country is divided.

Resolution 4503/2012 (Ministry of Health, 2012a) was also introduced to provide clearer and more organized guidance on how the random draws should be conducted. To prevent strategic application behavior and to take advantage of the fact that the number of newly graduated physicians was about twice the number of available positions, Resolution 4503/2012 established that physicians could apply only to one state and only when the number of applicants for that state was lower than twice the number of available spots. This process feature guaranteed an excess of demand for spots in each state and cohort.

After the application process closes, each state runs a public, random assignment of the available spots for each profession, according to the following steps: First, an oversight board consisting of one civil servant from the state secretariat of health, and four health professionals are chosen. The civil servant then publicly announces the number of spots available and who registered for each profession. At this point, she also states the rules for the lotteries, typically by using ballots. If a health professional gets a white ballot, they are exempt from the social service and receive a certificate that allows them to work in Colombia as a professional (i.e. medical license). Otherwise, the professional gets a red ballot with the randomly assigned code of the specific LHC where they will provide their services as a professional. If there are fewer professionals than spots available, all professionals registered are assigned to a LHC. Still, the specific LHC is assigned through the lotteries. Finally, the civil servant of the secretariat of health prepares a report stating the SSO-assigned physicians and their assigned LHCs, as well as the professionals who are exempt from the SSO program.

The social service at the assigned LHCs begins typically between one to three months after the draw and lasts for 12 months. The starting date is defined before the random assignment and, therefore, is orthogonal to the physicians' characteristics as well. If a health professional refuses to work in the place to which they were assigned or unilaterally quits before the official end of their service, they are given a six-month sanction where they cannot work as a health professional. After that period, they must apply to the SSO program again. This sanction imposes strong costs for quitters and has proved to be a good deterrence for dropping the program. The system of assigning professionals to LHCs randomly lasted for seven draws.[19] Since October 2014, a new centralized system giving more weight to professionals stating preferences and a list of prioritizations has replaced the random assignment.

The random assignment period is a perfect setting to estimate causal relationships that would otherwise be difficult to identify. The SSO assignment has implications for both the professionals who are selected randomly and the communities that get assigned doctors with different qualities. The latter are the focus of the present paper; the implications

[19] All four of the 2013 cohorts and the first three of 2014.

for the professionals are studied in Guarin, Posso, Saravia, and Tamayo (2023). In this paper, we use the exogenous rule of assignment to compare the birth outcomes of patients in LHCs that were assigned professionals with different medical skills but are otherwise comparable.

Despite the SSO being mandatory for health graduates in different fields,[20] in this paper we focus on physicians for three reasons. First, it was the profession for which the excess demand for the state-level draws was mandatory, creating perfect conditions for lotteries. Second, in Colombia, prenatal examinations must be carried out by physicians (Gomez et al., 2013). Finally, these professionals arguably make the greatest contribution to the health of the patient (Das and Hammer, 2005), specifically to birth outcomes.

The Ministry of Health (1990, 2001) specifies that the responsibilities of a physician during their SSO period are: developing health prevention programs (such as vaccinations, family planning, antenatal controls, control of chronic diseases, buccal and visual health); providing primary care and diagnosis; assigning treatment and therapies; creating and improving medical records; making a health plan and the epidemiological profile for the local community; and performing any other duty stated in their contract. Moreover, LHCs explicitly mention attending and performing surgical procedures, including C-sections and childbirth, as part of the functions and activities of SSO physicians.[21]

3 Data

We use data from five main sets of administrative records. The primary dataset comes from the reports written and published by the Ministry of Health for each state-level draw implemented in January, April, July, and October 2013 and January, April, and July 2014 (Ministry of Health, 2014). From this data, we obtained individual identifications, the draw date, the state to which physician applied, whether the physician was selected by the lottery or not, and importantly the LHC to which each student was randomly assigned and the proposed start date. For our period of analysis, 45% of the LHCs in the program show up in only one draw, while 29% of the LHCs appear in two draws and 26% of the LHCs appear between three to five times.

The second administrative dataset comes from the Colombian Institute for Educational Evaluation (Spanish acronym, ICFES). The ICFES is the institution that administers the mandatory college exit exam (called SABER PRO) that all professionals, including physicians, must take before graduation (Colombian Institute for Educational Evaluation,

[20] It is mandatory for newly graduated professionals from medicine, nursing, bacteriology, and dentistry.

[21] We reviewed the manual of functions for five LHCs included in our sample. The reviewed institutions were LHC Salazar de Villeta, LHC Francisco Valderrama, Subred de Servicios de Salud sur, Red de servicios del primer nivel, and Guaviare.

2014). Using national ID numbers, we are able to link the physicians participating in the SSO program to the ICFES records and recover their information from their field-specific medical exams (SABER PRO). From the SABER PRO, we gleaned physicians' individual performance on two health-related fields, *health care* and *disease prevention*, plus detailed sociodemographic information about each professional.[22]

Our estimations use the scores in the two health-related exams (health care and disease prevention) as proxies of physicians' quality before the SSO program.[23] The objective of the health-related tests in the SABER PRO is to measure the skills and knowledge of medical professionals. According to ICFES, the health care module assesses whether the physician has the competence to provide care that integrates both disease prevention and proper diagnosis with medical treatment and patient rehabilitation at all levels of complexity. In addition, the module on disease prevention evaluates the physicians' competence to apply basic concepts of health promotion and disease prevention that allow the prioritization of actions according to the individuals' health conditions.

Furthermore, ICFES ranks students into one of four categories of quality. For example, the lowest level in the health care module includes students who only understand basic concepts and elements of epidemiology and public health. On the other hand, the highest level includes students who understand public health concepts (actions aimed at mitigating health problems of communities), can assess patients' health conditions, and can analyze social, cultural, and economic factors that may influence differences across patients' health. Similarly, for the disease prevention module, ICFES groups the lowest level individuals who understand basic concepts of biosafety and occupational risk. The highest level includes professionals who can analyze complex health situations in a given context and select appropriate actions following current regulations and standards in medicine. Because the SSO program is the physicians' first real work experience, and the SABER PRO is taken just before graduation, we consider their scores a good measure of the their general and medical skills at the time they start their SSO service and their professional career.[24]

In Colombia, as in many other developing countries, there is high heterogeneity in the

[22]We also get the individual performance on two other fields: reading (comprehension) and quantitative (reasoning). Test scores are only available for the newly-appointed physicians (i.e., we do not have the test scores of physicians that graduated before 2009).

[23]The correlation between the physician's medical skills and their test performance has been documented previously in the literature. For example, Norcini, Lipner, and Kimball (2002) and Norcini, Boulet, Opalek, and Dauphinee (2014) showed a strong correlation between mortality and physicians' certifying examinations performances. Similarly, Tamblyn et al. (2002) found a relationship between examination scores and the primary care practice of doctors in Quebec. Wenghofer et al. (2009) found an association between medical examination scores and quality of health care in Canada, while Tamblyn, Abrahamowicz, Dauphinee, Wenghofer, Jacques, Klass, Smee, Blackmore, Winslade, Girard, et al. (2007) found a relationship between physicians' exam scores and patients' complaints to the medical regulatory authorities.

[24]Schnell and Currie (2018) provided evidence on the important link between physicians' education and their professional performance.

quality of education in medicine. In 2009, only 30% of medicine programs in Colombia had been accredited as high-quality programs by the Ministry of Education (Fernández Ávila, Mancipe García, Fernández Ávila, Reyes Sanmiguel, Díaz, and Gutiérrez, 2011). Figure 1 shows high heterogeneity on the average score of the health-specific SABER PRO test scores between and within programs (and universities) for the physicians in our sample.[25] The figure shows the mean score for each university/program and an interval of one standard deviation to each side of the mean. Notice that there is a difference of almost two standard deviations between the averages of the best and the worst programs. This high heterogeneity plays in our favor, because it allows us to compare the outcomes of patients who were randomly exposed to physicians with very different baseline knowledge and skills.[26]

Using the scores and demographic characteristics from the SABER PRO, Guarin et al. (2023) showed that the SSO lotteries in our sample are well balanced between SSO-assigned physicians and those who were randomly exempt. They use individual regressions correlating physicians' characteristics and lottery status as well as machine learning techniques and a classification permutation test to provide evidence of equality of multivariate distributions between treatment and control groups (Gagnon-Bartsch, Shem-Tov, et al., 2019) and the randomness of the selection into the program in general.

The third administrative dataset comes from Vital Statistics Records (VSR) collected by the Administrative Department of Statistics - DANE (Administrative Department of Statistics, 2018). The VSR records have rich information for all birth certificates filed in LHCs within Colombia's 1,120 municipalities, which are sub-divisions of the 32 states, from 1998 to 2018. Using LHCs' identification codes, we are able to link physicians and the birth records of the LHCs to which they were assigned. Using the birth date and number of gestation weeks from VSR, we are able to identify children born between 2013 and 2016 who were exposed to each team of physicians. We also use the VSR data from 2009 to 2012 to create mother and LHC-level controls to provide evidence of the covariate balance at the LHC level and to run placebo tests.

The fourth administrative data set comes from the 2005 National Census, also collected by DANE (Administrative Department of Statistics, 2005). From the census, we get the population and other variables at the municipality level that we use to test the randomization of the program and as controls in the robustness checks.

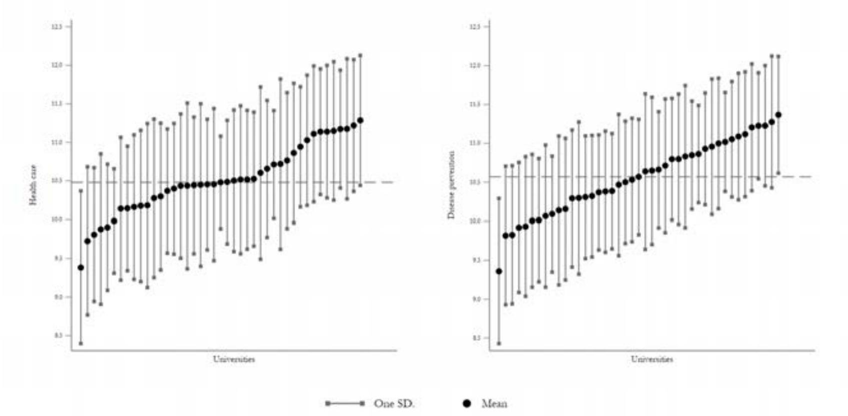
Finally, we collect information from the National Registry of Human Resources in Health, known as RETHUS. The Ministry of Health designed RETHUS through Law 1164 of 2007 (Congress of Colombia, 2007). RETHUS registers all individuals authorized to

[25]In the particular case of Colombia, each university, at most, has one medicine program.

[26]Similarly, Appendix Figure A.4 shows substantial heterogeneity for the quantitative and reading test scores in the universities the physicians in our sample attended.

practice a profession or occupation in health. This data contains detailed information on the date of degrees, the date on which the medical license was granted, and postgraduate degrees. We also collected additional data at the LHC level from the Colombian Ministry of Health.

Figure 1: Heterogeneity in SABER PRO scores in medicine programs



Notes: This figure reports the health care and prevention diseases test scores for the universities that the physicians in our sample attended. Data accounts for 44 different universities. The figure shows the mean score for each university/program and an interval of one standard deviation. The dashed horizontal line represents the overall median. The figure shows substantial heterogeneity both within and between programs. For all the fields reported, there is a difference of almost two standard deviations between the averages of the best and the worst programs and almost a one standard deviation difference between the averages of the worst and the median program and the averages of the median program and the best program.

3.1 Main sample

Our primary data source comes from the draws implemented in January, April, July, and October 2013 and January, April, and July 2014, which are the ones that were done at random. We constrain our sample to LHCs with at least one physician assigned in the seven draws and at least one birth certificate filed from 2013 through 2016. We also restrict the sample to 600 municipalities located outside of the main 23 Colombian metropolitan areas, which defined by the National Administrative Department of Statistics (DANE) based on the number of inhabitants and the degree of integration of the urban centers with the surrounding municipalities. Our sample of 600 municipalities covers around 58% of the Colombian population.

We exclude metropolitan areas for three reasons. First, the program's objective is to provide professional services in mostly rural areas with difficult access to health services (Resolution 1058/2010, Ministry of Health, 2010). Between 2013 and 2014, 77.3% of the available positions for assigned physicians within the LHCs were located in small cities beyond the metropolitan areas. Second, Colombia classifies LHCs by three levels of complexity. Local municipalities usually manage level 1 LHCs. LHCs at this level are institutions with low complexity technology, simple and easy to use in outpatient, hospitalization, emergency, and support services for diagnosis and treatment of minor health problems. Moreover, care is provided primarily by health care professionals. All LHCs in our main sample are level 1.[27] Finally, mothers located in metropolitan areas have access to a large supply of LHCs at all levels, where they can easily substitute among LHCs, while mothers in small cities outside of the metropolitan areas usually only have access to one level 1 LHC. We thus expect assigned physicians to play a less pivotal role in metropolitan areas.

We observe the birth certificate for each newborn, which includes information on low birth weight, Apgar score, weeks of gestation, antenatal consultations and demographic information for mothers and newborns. For each physician, we observe the four scores in health care, disease prevention, reading, and quantitative fields, plus some socio-demographic information provided by the physician at the time of the test. Our final sample contains 256,806 newborns and 2,126 physicians.

Table 1 provides the basic descriptive statistics for the main health outcomes used from the VSR.[28] It also shows how our sample changes as we add the restrictions used in our main estimations. Columns 1 and 2 show the mean and standard deviation, respectively, for newborns in LHCs where at least one SSO physician was assigned (SSO sample); columns 3 and 4 show the same statistics when we constrain the sample to the municipalities outside of the main metropolitan areas (i.e., rural areas). The last two columns (3 and 4) correspond to our final main sample. In our main sample, 4.26% of births were low birth weight (LBW), 4.09% were early-term infants (prematurity), and 3.74% of births had an Apgar score below 7. Our main outcome, *unhealthy*, takes the value of 1 if at least one of the previous outcomes (LBW, prematurity, or Apgar) is one.[29] also shows that newborn outcomes tend to be lower in metropolitan areas than in the rest of the country. This is a regularity for our study period, and for all years, the VSR is available. In our sample, 9.5% of the births experienced at least one of these three medical conditions. The share of female newborns is 48.85%. Moreover, 16.28% of the

[27]LHCs classified as levels 2 and 3 are administered by the departmental governments or co-administered between departments and municipalities. Level 2 LHCs have medium-level technology and offer specialized health professionals for outpatient care, diagnostic services, and treatment of medium severity pathologies. Level 3 LHCs are located in metropolitan areas and offer the highest level of technology and care by specialized and subspecialized health professionals at all levels of care.

[28]Unfortunately, we do not have the continuous measures for birthweight, weeks of gestation, and the level of the APGAR score.

[29]Table 1

mothers had insufficient prenatal consultations which is an indicator variable that takes the value of 1 if the mother received less than 4 prenatal visits. Finally, teenage pregnancy is 28.46% of total births in the main sample.

Table 1: Descriptive Vital Statistics Registers main sample 2013-2016

Covariate	Description	SSO sample		SSO Rural	
		Mean (1)	SD (2)	Mean (3)	SD (4)
Low birth weight	1(Weight < 2500)	0.0601	0.2377	0.0426	0.2019
Prematurity	1(Gestational weeks < 37)	0.0623	0.2417	0.0409	0.1982
Low Apgar	1(Apgar Score < 7)	0.0378	0.1908	0.0374	0.1897
Unhealthy	max $(LBW, Premature, APGAR)$	0.1183	0.3230	0.0950	0.2932
Female newborns		0.4877	0.4998	0.4885	0.4999
Insufficient prenatal visits	1(Prenatal visits < 4)	0.1798	0.3840	0.1628	0.3692
Teenage mother	1(Mother's age at birth $\leq$ 19)	0.2840	0.4509	0.2846	0.4512
Number of observations		372,609		256,806	

Notes: This table presents the mean and standard deviation (SD) for the main birth statistics of the newborns affected by the SSO program. The data comes from the 2013-2016 DANE VSR, which collects information about all births and deaths in Colombia. The "SSO sample" contains information on all the LHCs that had a physician assigned through the SSO program in our sample, while the "SSO rural" refers to our main sample, which restricts to municipalities outside metropolitan areas. "Low birth weight" is the proportion of newborns with low birth weight (weight <2,500 grams); "Prematurity" is the proportion of newborns who were premature (fewer than 37 weeks of gestation); "Low Apgar" is the proportion of newborns whose Apgar score is lower than 7; "Unhealthy" is the proportion of newborns with at least one of the three previous conditions (Low Birth Weight, prematurity, or low Apgar); "Female newborns" is the proportion of female newborns; "Insufficient prenatal visits" is the proportion of mothers who had less than four visits; and "Teenage mother" is the proportion of mothers that are 19 years old or younger at the time of giving birth.

3.1.1 Municipalities

As aforementioned, we keep those municipalities in rural areas—outside of the main 23 Colombian metropolitan areas—where we expect fewer physicians per municipality. There are 600 municipalities included in our sample (see Appendix Figure A.2). The median number of people living in each municipality is 14,049 (the mean is 22,042). The average share of people living with unsatisfied basic needs (UBN) is almost 50%, with municipalities where the whole population live under UBN.[30] This reassures that SSO physicians provide their services for one year in LHCs located in underserved areas.

We obtain the number of physicians per municipality from RETHUS.[31] From the 600 municipalities included in our sample, only 10 have more than two LHCs per municipality. The median number of physicians per LHC is three, and around 94% of the LHCs have less than 20 physicians per LHC.[32]

[30] As a reference, the average UBN for the 23 and 7 largest cities and their metropolitan areas is 21.5% and 17.4%, respectively.

[31] Unfortunately, from RETHUS, we have information at the municipality level, and we cannot match every physician (except for the SSO physicians) to the LHC at which they work.

[32] Figure A.3 shows the distribution of physicians per municipality for the sample of 590 municipalities with one LHC per municipality.

Moreover, most deliveries are attended by general practitioners and SSO physicians. In fact, approximately 90% (527 out of 588) of the municipalities with one LHC and available data on specialist availability do not have an obstetrician/gynecologist working in their LHCs. While this highlights the limited access to specialists in these areas, it is significant for our study, as we anticipate assigned SSO physicians will play a crucial role in providing maternal healthcare in their LHCs.

3.1.2 SSO Physicians

The analysis includes information for 2,126 physicians who won the lottery for the seven draws implemented between 2013 and 2014. Table A.1 presents the baseline summary statistics of the physicians considered in our sample. Nearly 56% of the physicians are females; 0.29% live in a neighborhood classified with a socioeconomic stratum 1 or 2, whereas 36% of them live in a socioeconomic stratum 3.[33] The average household of the physicians consists of 4 people. 64.4% (63.4%) of fathers (mothers) of the SSO physician have a degree of tertiary education. Almost 45% of these households have a monthly income of less than three monthly minimum wages (22.9% earn less than two). Finally, the average score in the Health care score for the physicians considered in our sample is 10.4, with a maximum of 13.9, and a standard deviation of 1, and the average score in the disease prevention for the physicians considered in our sample is 10.4, with a maximum of 13.4 and a standard deviation of 1.

3.1.3 Compliance

We use the ID numbers of all the physicians in the SSO program between 2013 and 2014 and merge them with RETHUS, which provides the dates on which the physicians graduated and obtained their licenses. We define as compliers those physicians for which the time between their graduation date and the date they obtain their licenses is more than three months and less than two years. The share of compliers is 94%.

4 Empirical Analysis

Our aim is to identify the impact of physicians on health outcomes. We estimate the impact on the 256,806 children whose mothers received care from 2,126 physicians randomly allocated to 618 small LHCs. We first test the internal validity of our identification strategy.

[33]Urban areas in Colombia are split into six socioeconomic strata and rural areas into two socioeconomic strata, in which the first has the lowest income levels (the poorest). Authorities use the strata to spatially target social spending like that in the supply of public services (e.g., water, electricity), health insurance for the poor, housing, among others.

Next, we provide our main results on health at birth outcomes. We explore whether physicians' effects are more pronounced among different subgroups. Finally, we compute a relative measure of value-added and regress it on several physicians' characteristics including the physicians' quality measure.

4.1 Empirical Strategy

Our empirical setting focuses on a health production function that relates health outcomes at birth to physicians' quality. In our setting, multiple teams were assigned randomly to a large number of patients who are associated with a specific LHC. The randomness of the assignment allows us to satisfy the identification assumption that the physician team is mean independent of the unobservable variables. Our main empirical strategy is based on an intent-to-treat (ITT) type that estimates the impact of being assigned to a more-skilled physician on a newborn's health outcome (i.e., unhealthy, low birth weight, prematurity, Apgar), using the following linear specification:

$$Y_{j,i} = \alpha + \gamma_d + \beta Z_{j(i)} + \epsilon_{j,i}, \tag{1}$$

where $Y_{j,i}$ is the outcome of child i exposed to a physician team j (in LHC h). $Z_{j(i)}$ is a score that measures the overall quality of the physicians team j that was assigned randomly to serve in LHC h and whose service period intersects with child i's gestation. Finally, γ_d are draw-by-state fixed effects. The key identifying assumption behind our specification is that conditional on the draw-by-state fixed effects, γ_d, the allocation of physicians to LHC h is independent of potential outcomes, $Y_{j,i}$. Thus, controlling for draw-by-state fixed effects is crucial to our identification strategy; otherwise, variation in physician quality could reflect other regional differences in the assignment of physicians to LHCs.[34] Finally, standard errors are clustered at the LHC level.

The coefficient of interest is β. Under the assumption that teams of doctors within each draw-state were assigned randomly to LHCs, β (estimated by OLS) cleanly identifies the effect of being assigned to a more-skilled team of physicians on children potentially exposed

[34] We check the robustness of our estimates to including a vector of ex-ante LHC, X_h, such as, number of inhabitants in the municipality, number of LHCs per municipality, area, an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise, and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. We also include a vector of sociodemographic information of mother-child i, W_i, including an indicator variable for the sex of the newborn, an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise, an indicator variable that takes the value of 1 if the mother was 19 or younger at the time of the birth and zero otherwise, and marital status. The results, as expected given the random assignment, show that the estimated effects are robust to the inclusion/exclusion of controls.

to their service in the assigned LHC h. To make the interpretation of the estimated coefficient β straightforward, $Z_{j(i)}$ is expressed in standard deviations of the skill measure. Therefore, the final result is interpreted as the percentage points change in the outcome variable associated with being assigned to a physician with one standard deviation higher skill measure. We also estimate heterogeneous effects using demographic characteristics of the newborns, their mothers, and the LHCs in which they were born.

We focus on unhealthy as our principal measure of health at birth, which captures the three main outcomes on which the literature has focused on: low birth weight, prematurity, and the Apgar score. Low birth weight is defined as being born with a birth weight below 2,500 grams and has been one of the principal measures of health at birth studied in the literature (Currie, 2011). Prematurity is defined as being born before the 37th gestational week. Prematurity is highly correlated with low birth weight and mortality (Almond et al., 2005; Gonzalez and Gilleskie, 2017). Children born prematurely are at greater risk of suffering a variety of health problems, some of which can ultimately cause death.[35] For Apgar, we use an indicator of whether the newborn had a score below 7 in Apgar, as the threshold of 7 is commonly used in the literature (Ehrenstein, 2009).[36] Almond et al. (2005) argued that using the Apgar score to evaluate birth outcomes has the same practical advantages as birth weight: (i) it is relatively easy to collect; (ii) it is already available in birth records' data; and (iii) it is a measure that does not depend on a rare event (such as mortality).

We focus on the average score of the two health-related health exams. Nonetheless, we provide robustness results using the first principal component and each score individually.[37]

[35]Complications include immunological, respiratory, central nervous system, gastrointestinal, hearing, and vision problems, as well as cognitive, motor, social-emotional, behavioral, and long-term growth problems (Butler, Behrman, et al., 2007; Currie and Walker, 2011; Taylor, Klein, Minich, and Hack, 2001; Veddovi, Kenny, Gibson, Bowen, and Starte, 2001). Callaghan, MacDorman, Rasmussen, Qin, and Lackritz (2006) reexamined the top 20 causes of infant deaths in 2002 and determined that both low birth weight and prematurity are the most common causes in the US and account for almost a third of infant deaths.

[36]Apgar has been used in the literature as a measure of newborn health status; for example, see Almond et al. (2010) and Lin (2009). Apgar is a measurement of the health of newborns based on breathing, heart rate, color, reflexes, and muscle tone (Moore, Harris, Laurens, Green, Brinkman, Lenroot, and Carr, 2014). Apgar scoring at birth was developed to evaluate the newborn's immediate condition and the potential need for resuscitation. Posterior studies have shown that Apgar scoring is a good predictor of infant death and ventilator use. Low Apgar scores can also predict long-term cognitive outcomes, such as neurological disability, reduced IQ, lower math scores, and low cognitive function (Almond et al., 2005; Moore et al., 2014; Moster, Lie, and Markestad, 2002). Among school-age children, low Apgar scores are also associated with minor language, motor, speech, and developmental impairments (Razaz, Boyce, Brownell, Jutte, Tremlett, Marrie, and Joseph, 2016).

[37]We repeat our main empirical analysis using all test score fields as proxies of physicians' quality before the SSO program. According to ICFES, the reading test measures how well a student understands the meaning of words or phrases, matches the parts of a text to make it global, and reflects on a text and evaluates its content. The quantitative test measures general knowledge in mathematics, statistics, and data analysis.

In addition, when a child is exposed to multiple physicians, a weighted average of the scores is computed, where the number of months exposed to each team of physicians during the pregnancy period is used as a weight.[38] Finally, we focus on municipalities outside of the main metropolitan areas for the entirety of the analysis.

4.2 Identification

Identification relies on the independence of the allocation of physicians to LHC h and potential outcomes, $Y_{j,i}$, conditional on the draw-by-state fixed effects, γ_d. To evaluate the internal validity of our identification strategy, we perform two tests: First, we test whether LHCs' birth outcomes and additional covariates measured in years 2010, 2011 and 2012 from VSR are correlated with the physicians' quality who were randomly assigned in 2013 and 2014 and provided medical care to individuals who were born between 2013 to 2016 (four years later). Our results show that there is no correlation between different LHCs' characteristics and our proxy for physicians' quality. Second, we implement different falsification tests, where we assign a "placebo treatment" to the newborns who show up in the VSR of the four years before the program (2009, 2010, 2011, and 2012) instead of years 2013, 2014, 2015, and 2016 used in our main estimation sample. We use the same draw date, proposed start date, and LHC to which each of the physicians was assigned randomly but four years before the actual date. We then run equation (1) under the same conditions used for the main sample.

4.2.1 Characteristics of the LHCs and physicians' quality

To test whether the main health at birth outcomes and additional covariates, measured before our main sample of the SSO program, are correlated with the quality of the physicians assigned to each LHC, we regress each LHC's characteristics three years before our main sample of the SSO program (i.e., from 2010 to 2012) on physicians' quality—proxied by the average of health-related college examination scores.[39] We include draw-by-state fixed effects and cluster the standard errors at the LHC level. Table 2 shows the coefficients and their standard errors from each regression. From Table 2, it follows that there is no significant correlation between physicians' quality (randomly assigned during our main sample period) and the health outcomes, as well as the LHC

[38] When the child is also exposed to cohorts from different draws, the draw-by-state fixed effect for the first cohort is assigned. Furthermore, our results hold when we use an unweighted average of the scores.

[39] We collapsed birth outcomes and other covariates at the LHC level using data for the three years before our main sample of the SSO program (i.e., from 2010 to 2012), and regress health-related college examination scores of each physician against each outcome or covariate.

characteristics measured three years before.[40]

Table 2: Covariate balance at LHC level

Covariate	Coefficient	Standard Error
Unhealthy	-0.0049	0.0170
Low birth weight	0.0011	0.0076
Prematurity	-0.0056	0.0078
Apgar < 7	-0.0084	0.0071
Antenatal consultations < 4	0.0155	0.0131
Female newborn	0.0018	0.0013
Mother with basic education	-0.0032	0.0058
Married mother	0.0042	0.0040
Teenage mother	0.0007	0.0023
Number of LHCs by municipalities	-0.0044	0.0032
Municipality population	-296.58	550.01

Notes: This table presents the results of eleven different LHC-by-cohort level regressions of the LHC-level variables, listed in the first column, on the physicians' quality measure and the draw-by-state fixed effects. The coefficient and the standard error of the physicians' quality variable are reported in the second and third columns, respectively. Standard errors are clustered at the LHC level. LHCs' characteristics come from the 2010-2012 DANE VSR. "Low birth weight" is the proportion of newborns with low birth weight (weight <2,500 grams); "Prematurity" is the proportion of newborns who were premature (fewer than 37 weeks of gestation); "Apgar < 7" is the proportion of newborns whose Apgar score is lower than 7; "Unhealthy" is the proportion of newborns with at least one of the three previous conditions. "Antenatal consultations < 4" is the proportion of mothers who had less than four antenatal visits (Insufficient antenatal consultations); "Female newborn" is the proportion of female newborns; "Mothers with basic education" is the proportion of mothers with at least secondary education at the time of the birth; "Married mother" is the proportion of mothers that are married at the time of the birth; and "Teenage mother" is the proportion of mothers that are 19 years old or younger at the time of the birth. We interpret the non-significance of these estimates as evidence in favor of the randomness of the assignment of physicians.

4.2.2 Placebo Tests

Recall that for our main results, we use data from the physicians randomly assigned in 2013 and 2014 and who provided medical care to individuals who were born between 2013 to 2016. We move the physician's arrival time four years back and run placebo tests similar to our main specification but using data for the four previous years (2009-2012). We then estimate equation (1) using the same outcomes and set of fixed effects.

Because physicians in our sample did not treat children born in 2009, 2010, 2011, and 2012, we would expect a null effect. Table 3 shows that the point estimates are

[40] In Appendix Table A.2, we estimate the same regression using the average of the four fields of the college examination scores as a proxy for physicians' quality. Again, we find no correlation between the proxy used for the quality of physicians that arrived during our main sample period (i.e., in 2013 to 2014) and the different LHC characteristics measured three years before.

precisely estimated zeros for our main outcome, unhealthy, and for each of the other health outcomes (LBW, prematurity, Apgar).[41] Our results are robust to the use of the first principal component as a proxy for skill, as well as to the inclusion of a set of controls such as ex-ante LHC and team characteristics as well as a vector of sociodemographic information of mother-child (Appendix Figure A.5 and Table A.4). Finally, we implement an additional placebo test using a Logit model and compute the average marginal effect associated with an increase in one standard deviation of the skill measure. Appendix Table A.5 shows that the marginal effects (signs and magnitudes) are null, similar to the ones estimated using a linear regression model.

Table 3: Placebo test

	Unhealthy	**LBW**	**Prematurity**	**Apgar < 7**
	Average Health Scores			
	(1)	(2)	(3)	(4)
Coefficient	-0.0009	-0.0008	-0.0020	0.0004
Stand. Err.	(0.0022)	(0.0011)	(0.0016)	(0.0013)
Relative effect	-0.79%	-1.76%	-3.76%	0.78%
Average Dependent Variable	0.119	0.046	0.052	0.047
Number of Observations	262,089			

Notes: This table shows a placebo test where we estimate equation (1) but moving the arrival date of the physician four years back (years 2009-2012). The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature or if the Apgar score of the newborn is lower than 7 and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. We read the results of this placebo test as additional evidence in favor of the randomness of the assignment of the physicians to LHCs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.3 Impacts on health at birth

In this section, we provide our main results on health at birth outcomes. Table 4 presents the estimated coefficient β, in equation (1), using ordinary least squares. We find that our main quality measure has a negative and significant effect on unhealthy as well as each of the health outcomes (i.e., LBW, prematurity, and Apgar). The standard error of the coefficient is presented in parenthesis, and below we present the relative (percent) effect (i.e., we divide the main coefficient by the mean of the dependent variable).

In column (1) of Table 4, we see a significant negative relationship between the quality

[41] In Appendix Table A.3, we repeat the same exercise and present the results for windows of 3.5, 3, 2.5 and 2 years before the start of the SSO program.

of physicians and unhealthy—a decrease in the probability of being born unhealthy of 0.6 percentage points. That is, being assigned a physician with one standard deviation higher performance in the health graduation exam scores decreases the probability of being born unhealthy by 6.31%.[42] Columns (2) to (4) in Table 4 examine each measure of health at birth. The point estimate for the average of the health-specific scores is associated with a decrease in the probability of low birth weight of 0.33 percentage points (7.71%), being premature of 0.33 percentage points (7.97%) and a drop in the probability of being born with an Apgar score below 7 of 0.27 percentage points (7.16%).[43,44]

Our results are similar to Amarante et al. (2016) who explores in utero exposure to a social assistance program in Uruguay to estimate the effects on birth outcomes. They found that participation in the program led to a "sizeable" (19% - 25%) reduction in the incidence of low birth weight. Similarly, Currie and Schwandt (2016a) found that fetal exposure to 9/11 release of toxic dust negatively affected gestation length, prematurity, birth weight, and low birth weight. Barber and Gertler (2010) evaluated the impact of *Progresa/Oportunidades* on birth weight and found a very large reduction in the incidence of low birth weight (44.5% lower among beneficiary mothers).

4.4 Physicians' impacts across subgroups

In this section, we explore whether physicians' effects are more pronounced among some subgroups. The literature in economics has studied a variety of heterogeneous effects across different socioeconomic groups, measured by mother's education, age, marital status

[42]In the education context, the teacher value-added literature (e.g., Chetty et al., 2014; Rothstein, 2017) found that an increase in teacher quality of one standard deviation corresponded to an increase in students' test scores of 0.19 standard deviations in math and 0.14 standard deviations in reading. Our results suggest an increase in physician quality of one standard deviation corresponds to a decrease in the probability of being born unhealthy by 6.31%. We find similar effects (6.81%) when we use the average of reading and quantitative score as a proxy for physicians' quality (see Appendix Table A.6). Note that in our context, a one standard deviation increase is almost equivalent to the change from having a physician from the bottom-ranked program to having a physician from a median-ranked program or from having a physician from a median-ranked program to having a physician from the top-ranked program (see Figure 1).

[43]These results are consistent with what has been found in the previous literature and the Colombian context that prematurity is an important determinant of weight at birth (Almond et al., 2005). We find a strong correlation between prematurity and low birth weight in Colombia. Figure A.1 in the Appendix shows a monotonic negative correlation between the probability of low birth weight and the number of gestational weeks for all births in Colombia between 2009 and 2012. The figure presents the local polynomial regression fit of the probability of low birth weight over the number of gestational weeks using all birth records in Colombia from 2009 to 2012.

[44]We also compute cohort-level estimates on mortality. Due to substantial data limitations in mortality records, including missing information on gestation weeks and incomplete LHC data, we were compelled to conduct a cohort-level analysis instead of our preferred birth-level estimates. Given these data limitations, our findings, as detailed in Appendix Table A.7, are expected to be subject to measurement error attenuation. They indicate a negative, but not statistically significant, effect of more qualified physicians on mortality, which aligns with our main results.

Table 4: Main estimates using all sample and average score

	Unhealthy	**LBW**	**Prematurity**	**Apgar < 7**
	Average Health Scores			
	(1)	(2)	(3)	(4)
Coefficient	-0.0060***	-0.0033**	-0.0033**	-0.0027**
Stand. Err.	(0.0020)	(0.0016)	(0.0015)	(0.0013)
Relative effect	-6.31%	-7.71%	-7.97%	-7.16%
Average Dependent Variable	0.095	0.043	0.041	0.037
Number of Observations	256,805			

Notes: This table presents our main estimates. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature or if the Apgar score of the newborn is lower than 7 and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. We interpret the high significance and consistency of these results across the different measures of health at birth as evidence of the important role that skilled physicians play in determining infant's health.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

and gender of the newborn (Almond and Mazumder, 2011; Amarante et al., 2016; Currie and Schwandt, 2016a; Dinkelman, 2017; Eriksson et al., 2010; Hoynes, Page, and Stevens, 2011; Okeke and Abubakar, 2020; Persson and Rossin-Slater, 2018). Similar to other studies focusing on the VSR, our data includes information on the fetus's sex, mother's education, age, marital status, and if the mother is a first-time mother.

We find that the effect of being assigned to a more skilled physician on our main outcome, unhealthy, is slightly more pronounced among first-time mothers and teenage mothers (see Table 5), but we do not find statistically significant differences on the effects across mothers' characteristics. Furthermore, we do not find statistically significant differences across mothers with high and low education, as well as married and single mothers (Appendix Table A.8). Finally, we examine whether the treatment effects vary by the infant's sex. It has been established that male fetuses are more vulnerable to health shocks than female fetuses (Almond and Mazumder, 2011; Currie and Schwandt, 2016a; Eriksson et al., 2010; Kraemer, 2000; Naeye et al., 1971).[45] It is possible that skilled physicians play an important role in mitigating negative shocks on more vulnerable fetuses. Although we find that the reduction in unhealthy was particularly pronounced among male newborns, we do not find any statistical difference between males and females

[45]In medicine and epidemiology, this phenomenon is known as "fragile males" (Cameron, 2004; Eriksson et al., 2010; Kraemer, 2000; Mathews, Johnson, and Neil, 2008; Mizuno, 2000).

(see Appendix Table A.8).

Table 5: Heterogeneity of the effects across mothers' characteristics

	Unhealthy					
	LHC		Mother			
	Higher incidence of Unhealthy (1)	Lower incidence of Unhealthy (2)	First-time (3)	Non-first-time (4)	Teenage mother (5)	Non-teenage mother (6)
	Average score					
Coefficient	-0.0073**	-0.0041*	-0.0070***	-0.0055***	-0.0070***	-0.0058***
Stand. Err.	(0.0031)	(0.0022)	(0.0025)	(0.0018)	(0.0024)	(0.0019)
Relative effect	-6.08%	-4.32%	-6.40%	-5.78%	-6.25%	-6.12%
Average Dependent Variable	0.120	0.095	0.109	0.095	0.113	0.095
Number of Observations	80,992	175,810	152,447	104,357	152,410	104,393

Notes: This table presents the heterogeneity of our estimated results when we divide the sample by LHC and mothers' characteristics. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) for each subgroup. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature or if the Apgar score of the newborn is lower than 7 and zero otherwise. A LHC with "Higher incidence of unhealthy" ("Lower incidence of unhealthy") is a LHC above (below) the 75th percentile of the ex-ante unhealthy proportion distribution. "First-time" refers to the group of of mothers who are giving birth to their first child. "Non-first-time" refers to the complement of the "First-time". A mother is a "Teenage mother" if she is giving birth at age 19 or younger, and is a "Non-teenage mother" otherwise. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. We interpret these results as evidence of a (weak) significant difference between the effect of physicians in LHCs with a high and low incidence of poor health and a lack of evidence of any statistically significant differences in the effects across the observed mothers' characteristics.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.4.1 LHCs' characteristics

Finally, we look at heterogeneity across LHCs' characteristics. We divide the sample between LHCs below (low incidence) and above (high incidence) the 75th percentile of our main outcome—unhealthy—distribution using data from the VSR records for the three years before our sample period (2010-2012). In Table 5, columns 1 and 2, we test the effects associated with physicians assigned to LHCs with an ex-ante low or high incidence of unhealthiness.

We find a (weak) significant difference between the effect of more-skilled physicians on unhealthy in LHCs with a high and low incidence of poor health. The effect is strongly negative and significant in LHCs with a high incidence of poor health (Currie, 2011). The point estimate for physicians in LHCs with high (low) incidence is -0.73 (-0.41) percentage points. Thus, being assigned to a physician with one standard deviation higher skill measure decreases the probability of a child being born unhealthy by 6.08% (4.32%) in a LHC with high (low) incidence of unhealthy. Reassuringly, this evidence suggests that physicians play a more important role in LHCs with a history of poor health outcomes.

These results relate to the wide literature on heterogeneous clinical practices across

LHCs and whether these differences translate into health outcomes. Doyle, Graves, Gruber, and Kleiner (2015) found significant health benefits for older patients who were brought to higher-cost LHCs, Card, Fenizia, and Silver (2019) found that, during the first year of life, newborns who were delivered by C-section were more likely to visit the emergency department, less likely to be readmitted to LHC, and had lower mortality rates. Related contributions include Cutler, Skinner, Stern, and Wennberg (2019) and Finkelstein, Gentzkow, and Williams (2016). Skinner (2011) provides a review of the literature on regional variation in intensity of care or spending.

4.5 Physicians' Value-added

While our empirical setting relates health outcomes at birth to physicians' skills, the estimated coefficients cannot be interpreted as the effect of exogenously increasing physicians' skills while keeping everything else constant. Instead, we identify the effect of being assigned to a more skilled versus a less skilled physician, with all the characteristics that may differ between these two types of professionals. Our results can be useful for policy and decision-makers since test scores are a proxy of skills that can be observed. However, note that test scores may not capture all the factors that affect clinical competence, which would understate the role of the physicians' quality in determining patients' health. To assess this concern, we take advantage of the random assignment of physicians and compute a relative measure of value-added.

Consider the model where child i's potential health at birth outcome when assigned to physician j, denoted by $Y_{j,i}$, can be written as the sum of two components:

$$Y_{j,i} = \mu_j + \alpha_i, \tag{2}$$

where μ_j is the mean potential outcome of children whose mothers got assigned physician j and α_i is the child's latent health at birth. Let $D_{j,i}$ be a dummy variable indicating whether child i's mother was assigned to physician j. The observed health at birth outcome for child i can then be expressed as:

$$\begin{aligned} Y_i &= Y_{s,i} + \sum_{j=1}^{J} (Y_{j,i} - Y_{s,i}) D_{j,i} \\ &= \mu_s + \sum_{j=1}^{J} \beta_j D_{j,i} + \alpha_i, \end{aligned} \tag{3}$$

where the parameter β_j measures the physician j's value-added relative to a reference physician with index value s. In most settings, the match between physicians and patients is at risk of being correlated with other patients' unobserved characteristics, implying that the estimation of equation (3) using OLS would result in potentially biased estimates of the physician's value-added.

Now, consider the projection version of equation (3) but controlling by the draw-by-state fixed effects γ_s,

$$Y_i = \gamma_s + \sum_{j=1}^{J} \beta_j D_{j,i} + \varepsilon_i. \tag{4}$$

Since, in our setting, several cohorts of physicians applied to a specific state and were randomly assigned to LHCs in that state, we have $E[D_{j,i}\varepsilon_i|\gamma_s] = 0$, for all $j = 1, ..., J$, and OLS estimates can identify the causal effect of physicians on children's outcomes. Due to the random nature of physicians assignment to LHCs at the draw-by-state level, we estimate physicians' relative value added by first running a regression of the main health at birth outcome variable on the draw-by-state fixed effects:

$$Y_i = \gamma_s + r_i. \tag{5}$$

We then compute residuals $\hat{r}_i$ from equation (5) and regress them on the J different assignment indicator dummies to recover the estimated physician effect:

$$\hat{r}_i = \sum_{j=1}^{J} \beta_j D_{j,i} + \epsilon_i. \tag{6}$$

Where the $\hat{\beta}_j$, estimated using OLS, is an unbiased estimate of physician j's effect on children's health relative to the draw-by-state average.

The empirical value-added literature typically shrinks the value-added estimates toward a common Bayesian prior (Herrmann, Walsh, and Isenberg, 2016). The benefit of the shrinkage procedure is to produce estimates of value-added for which the estimation error variance is reduced through the dependence on the stable prior. In practice, the prior is specified as the average value-added (Chetty et al., 2014; Kane, Rockoff, and Staiger, 2008).[46] The weight applied to the prior for a physician is an increasing function of the variance with which that value-added is estimated. The following formula describes the empirical shrinkage procedure estimated:

$$\hat{\beta}_j^{EB} = a_j \hat{\beta}_j + (1 - a_j)\bar{\beta}$$

$$a_j = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\lambda}_j}$$

[46] This benefit is particularly valuable in applications where researchers plan to use value-added as an explanatory variable (Chetty et al., 2014; Harris and Sass, 2014; Kane et al., 2008), in which case it reduces attenuation bias.

where $\hat{\sigma}^2$ is the estimated variance of value-added and $\hat{\lambda}_j$ is the squared standard error of $\hat{\beta}_j$. The standard deviation of the shrunken value-added is 0.06σ so that having a physician with value-added at the 85th versus 50th percentile would decrease the probability of being unhealthy by roughly 0.06σ.

Subsequently, we regress the J different estimates of the physicians' relative effects on several physicians' relative characteristics, including the performance in the med-school exit exam, as an attempt to study which physician's characteristics correlate more with the estimated valued-added effects. Columns (1) and (2) of Table 6 show the results of regressing the aforementioned physicians' estimated effects on different sets of physician characteristics. Columns (3) and (4) control for the additional observable characteristics (children and LHC) in equation (5) (e.g., LHC health indicators) to account for the quality of other potential longer-term appointed physicians at the LHC. Table 6 shows that results are similar to the ones presented in Columns (1) and (2).

We interpret the results from Table 6 as evidence of the relevance of the health test scores in predicting physicians' performance. Column (1) shows that the test score is negatively and significantly correlated with the physicians' relative value-added. A one standard deviation increase in the test score is associated with a 0.0093 percentage points improvement in value-added.[47] However, the significant relationship between the scores and the relative value-added could be the result of the test score being correlated with other physicians' characteristics, which could be more relevant and closely associated with physicians' performance. To test this hypothesis, we regress the estimated relative value-added on physicians' test scores and other characteristics that were observed at the same time as test scores, including gender, family socioeconomic characteristics, and some proxies for the quality of the medicine program attended. The results show that, not only does the coefficient on test scores remain significant and statically equivalent to the one in column 1, but also, once we account for the test score, none of the other physicians' observed characteristics has a significant correlation with the physicians' performance. These two results highlight the relevance of the test scores as both a practical observable tool and as an indicator with high predicting power. Finally, as is expected, Columns 3 and 4 show that, given the random assignment, we get similar results when we control for the quality of other potential longer-term appointed physicians.

[47]Since the outcome in equation 5 is the probability of being born unhealthy, a negative value-added has a positive connotation. Also, note that this coefficient should be similar to the one estimated in Table 4, but does not have to be the same; the regression in Table 4 is at the child-level, whereas the regression in Table 5 is at the physician-level.

Table 6: Physicians' observables and their relative value added

	Dependent variables:			
	Value added without controls		Value added with controls	
	(1)	(2)	(3)	(4)
Average health scores	-0.0093***	-0.0084**	-0.0109***	-0.0105***
	(0.0032)	(0.0035)	(0.0039)	(0.0041)
Female		-0.0002		0.0048
		(0.0067)		(0.0068)
Father with tertiary education		0.0068		0.0128
		(0.0078)		(0.0088)
Mother with tertiary education		-0.0010		-0.0060
		(0.0076)		(0.0077)
The father or the mother has a job		0.0040		0.0063
		(0.0096)		(0.0091)
Top program		-0.0100		-0.0065
		(0.0096)		(0.0098)
Top income		-0.0071		-0.0036
		(0.0073)		(0.0078)
Public school		-0.0071		-0.0114
		(0.0086)		(0.0102)
Accredited program		0.0070		0.0121
		(0.0078)		(0.0081)

Notes: This table reports the results of regressing physicians' estimated relative value-added on observable characteristics. Each column from (1) to (4) refers to a different regression. The regressors, listed in the first column, are expressed in relative terms with respect to the by draw and by state average. Column 1 includes only the average health score as a regressor. Column 2 includes other physicians' characteristics as well. Columns 3 and 4 present the results on analogous exercises where relative value-added is estimated as in equation 4 but also using the following observed child and mother characteristics as controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Numbers in parentheses are LHC-level clustered standard errors. We interpret the results from this table as evidence of the distinctive relevance of the health test scores in predicting physicians' performance.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5 Potential Mechanisms

Previous literature has found differences in practice patterns (e.g., between male and female physicians and across geographies) and how these practices affect health outcomes (Tsugawa, Jena, Figueroa, Orav, Blumenthal, and Jha, 2017). Some of these practices, such as the quality of medical advice doctors provide, are unobservable (Das et al., 2008; Leonard and Masatu, 2007), whereas others, such as the number of prenatal consultations, are observable. In this section, we study prenatal consultations as a potential mechanism for observed differences between skilled and unskilled physicians.

5.1 Prenatal consultations

We first explore whether more-skilled physicians increase the number of prenatal consultations, serving as a mechanism to improve the quality of care and health outcomes. Although most of the body of evidence from both economics and medical research shows an important association between prenatal care and both birth weight and prematurity, there are some disagreements (Alexander and Korenbrot, 1995; Amarante et al., 2016; Barber and Gertler, 2010; Carrillo and Feres, 2019; Conway and Deb, 2005; Currie and Grogger, 2002; Grossman and Joyce, 1990; Kramer, 1987; McCormick and Siegel, 2001).

According to WHO (2016) and the Colombian government (Gomez et al., 2013), prenatal care improves the health status of both mother and newborn. In Colombia, the Ministry of Health requires physicians to carry out prenatal monitoring (Gomez et al., 2013). We follow the standard recommended by WHO (2016) for our period of analysis and measure "adequate prenatal care" contact as having at least four visits to the doctor during pregnancy. We do not find evidence that more-skilled doctors reduce the probability that mothers are scheduled for less than four prenatal checkups (see Appendix Table A.9).

We expect that physicians enrolled in the SSO program and assigned to rural areas (outside the metropolitan areas) would be time constrained, as usually they are the only physicians available in those areas.[48] Anecdotal evidence supports this notion, as described in various reports from Colombian medical associations in which physicians refer to the SSO year as an experience during which they had an overwhelming workload and long working hours.[49] In this setting, in which physicians are time-constrained, it comes as no surprise that the overall likelihood of having sufficient prenatal consultations is not significantly affected by the quality of the practitioners. However, we would expect that better physicians could be better at targeting care and more efficiently assigning their resources. Thus, we test whether the more-skilled physicians are targeting their prenatal consultations toward the most vulnerable mothers, measured as those most likely to give birth to an unhealthy baby.

We assume that the probability of an unhealthy baby can be thought of as a prediction problem and take advantage of recent advances in machine learning techniques.[50] We use

[48] Remember that the median number of physicians per LHC in these rural areas is 3.

[49] See, for example, two reports from the Colegio Médico Colombiano (2018) and Universidad del Rosario (2015).

[50] Supervised machine learning seeks to solve the problem of prediction (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015). Athey and Imbens (2017) and Mullainathan and Spiess (2017) emphasize that machine learning is significantly better at making predictions, in part because it is able to use very flexible functional forms and to fit complex data structures without imposing any specific restrictions in advance. According to Mullainathan and Spiess (2017), machine learning algorithms can do significantly better than traditional methods, even with moderate sample sizes and few covariates.

these techniques to generate two groups of predictions about the mothers' probability of giving birth to an unhealthy baby using a set of mother-LHC characteristics that are available for the physician at the time of prenatal care. We apply algorithms that are commonly used in the machine learning literature: random forest and logistic regression models.[51]

The sample is clustered into training and testing subsets of randomly selected LHCs using K-means algorithm. We repeat this procedure—splitting the main sample using K-means—5,000 times. We run Logit and random forest models on the training sets and use the models to predict the probability of giving birth to an unhealthy child on each testing subset.[52] We then divide the test sample into two groups: *low* and *high* predicted probability, defined as mothers with a probability of giving birth to an unhealthy child below and above the 75th percentile, respectively, for each of the two model predictions.[53]

We estimate equation (1) using a dummy that is equal to 1 if the number of prenatal consultations is less than four—as our main outcome—in each of the previously defined groups (i.e., low and high predicted probability of an unhealthy child). Table 7 presents the average coefficient and the standard error for the 5,000 repetitions.[54] Columns (1) and (2) present the results for the sample of mothers with a low predicted probability, and columns (3) and (4) for the sample of mothers with a high predicted probability of giving birth to an unhealthy child. We include the results both with and without controls.

Table 7 shows that regardless of the method we use, more-skilled doctors do not seem to increase the recommended number of antenatal consultations for mothers with a low predicted probability of giving birth to an unhealthy child. Instead, they target prenatal checkups toward the more vulnerable mothers, measured as mothers with a high predicted probability of giving birth to an unhealthy baby. Consistent with our suggested mechanism of physicians being able to target care toward the more vulnerable mothers, we find stronger effects of physicians' quality when we focus on mothers with a higher predicted probability compared to those with lower predicted probability. While the point estimate for the effect of physicians' quality on unhealthy in the lower predicted probability sample is between -0.08 and 0.09 percentage points depending on the prediction used to divide the data, the point estimate for the higher predicted probability group is between

[51]These methods are able to handle many covariates and they provide natural estimators of parameters when these are highly complex. The focus in the machine learning literature is often on working properties of algorithms in specific settings. See Mullainathan and Spiess (2017) for a review of the literature and Breiman (2001) for a description of the methods.

[52]We follow Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018) and re-scale the outcomes and covariates to be between 0 and 1 before training.

[53]Liberman, Neilson, Opazo, and Zimmerman (2018) and Liberman, Medina, Neilson, and Posso (2021) followed a similar strategy when they studied the effects of information deletion and usury rates on consumer credit markets.

[54]Figure A.6 shows that the distribution of the estimated coefficients for all the 5,000 repetitions.

-1.3 and -0.91 percentage points. These estimates suggest that an increase of one standard deviation in physicians' average score decreases the probability that mothers are scheduled for less than four prenatal checkups between 5.59% and 7.98% for mothers with high predicted probability of giving birth to an unhealthy child.

Taken together, the results from this section are consistent with a story of time-constrained physicians not being able to increase the average time spent in prenatal consultations but improving the targeting of care toward the more vulnerable mothers.

Table 7: Antenatal consultations by predicted probability of an unhealthy newborn

	Dependent variable: Antenatal consultations < 4			
	Low predicted probability of Unhealthy		**High predicted probability of Unhealthy**	
	Without controls (1)	With controls (2)	Without controls (3)	With controls (4)
	Panel A. Logit			
Coefficient	0.0009	0.0004	-0.0095***	-0.013***
Stand. Err.	(0.001)	(0.0009)	(0.0026)	(0.0029)
Relative effect	0.55%	0.24%	-5.83%	-7.98%
	Panel B. Random forest			
Coefficient	-0.0007	-0.0008	-0.0091***	-0.0094**
Stand. Err.	(0.0016)	(0.0073)	(0.004)	(0.0046)
Relative effect	-0.44%	-0.49%	-5.59%	-5.77%

Notes: This table reports the differential effects of physicians on antenatal consultations by mother's predicted probability of giving birth to an unhealthy child. To predict the probability of an unhealthy child, we divided our data into training and testing subsets of randomly selected LHCs using a K-mean algorithm. On the training sets, we run Logit and random forest models of the probability of being born unhealthy on our usual set of mother and LHC ex-ante covariates, and use the estimations to predict the probability of giving birth to an unhealthy child on each testing subset. Using the prediction on the testing sample, we divide each subset into high and low predicted probability of giving birth to an unhealthy child, defined as mothers with a probability of an unhealthy child below and above the median, respectively. The coefficients presented represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) on the probability of having insufficient (less than four) antenatal consultations. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. We interpret the non-significant effect for the "Low predicted probability of Unhealthy" group and the significant effect for the "High predicted probability of Unhealthy" group, as evidence consistent with the idea that more skilled physicians are better at targeting care towards the more vulnerable mothers.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.1.1 Effect on unhealthy

We next show—consistent with the idea of better physicians being better at targeting care to the most vulnerable mothers—if being assigned to higher skilled physicians reduces the probability of giving birth to an unhealthy child, particularly among the most vulnerable mothers. Table 8 shows that being assigned to more skilled doctors seem to improve health at birth of children for all mothers (i.e., with a low and high predicted probability of unhealthy babies). However, the effect is more pronounced, regardless of the method we use to split the sample, for mothers with a (ex-ante) high predicted probability of unhealthy babies. In particular, for the more vulnerable mothers, being assigned to a physician with one standard deviation higher health graduation exam scores decreases the probability of an unhealthy newborn by around 9%, while for mothers with (ex-ante) low predicted probability of an unhealthy newborn, the effects are smaller in magnitude, close to 5%.[55]

Table 8: Main outcomes by predicted unhealthiness

	Dependent variable: Unhealthy			
	Low predicted probability of Unhealthy		**High predicted probability of Unhealthy**	
	Without controls (1)	With controls (2)	Without controls (3)	With controls (4)
	Panel A. Logit			
Coefficient	-0.0055***	-0.0052***	-0.0092***	-0.0093***
Stand. Err.	(00003)	(0.0003)	(0.0009)	(0.0009)
Relative effect	-5.79%	-5.47%	-9.69%	-9.79%
	Panel B. Random forest			
Coefficient	-0.0031	-0.0058***	-0.0080***	-0.0084***
Stand. Err.	(0.0021)	(0.0004)	(0.0029)	(0.0012)
Relative effect	-3.34%	-7.16%	-8.57%	-8.85%

Notes: This table reports the differential effects of physicians on the probability of being born unhealthy, by mother's predicted probability of giving birth to an unhealthy child. To predict the probability of an unhealthy child, we divided our data into training and testing subsets of randomly selected LHCs using a K-mean algorithm. On the training sets, we run Logit and random forest models of the probability of being born unhealthy on our usual set of mother and LHC ex-ante covariates, and use the estimations to predict the probability of giving birth to an unhealthy child on each testing subset. Using the prediction on the testing sample, we divide each subset into high and low predicted probability of giving birth to an unhealthy child, defined as mothers with a probability of an unhealthy child below and above the median, respectively. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise. The coefficients presented represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) on "Unhealthy". Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. The results show how, consistent with the idea of better physicians being better at targeting care to the most vulnerable mothers, the negative effects on the probability of unhealthy are particularly pronounced among the more vulnerable mothers.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[55] Figure A.7 presents the distribution of the estimated coefficients for the 5,000 repetitions for the four outcomes studied.

6 Robustness checks

6.1 Proxies for physicians' quality

We run additional specifications in which we use the standardized principal component instead of the standardized average health related scores as a proxy for physicians' quality. In addition, we show that our results are robust to the inclusion of ex-ante LHC characteristics as well as a vector of sociodemographic information of mother-child. Figure A.8 compares the estimated coefficient (relative to the mean), β, in equation (1) using the average (main specification) and the principal component of health scores with and without controls. We see from Figure A.8 that our estimates for unhealthy—and each of the three health measures—are similar if we use the first principal component as a proxy for quality and are robust to the set of controls included in our analysis.[56]

Finally, we standardized, centered, and aggregate the three main health outcomes (LBW, prematurity, and Apgar score) using the inverse covariance index suggested by Anderson (2008) and repeat our main empirical analysis using the index as dependent variable. Appendix Table A.11 presents the results using the covariance index and our main outcome, unhealthy (standardized), as dependent variables. Note that the adjusted standardized coefficients (in standard deviations) are very similar for both specifications.

6.2 Nonlinear model

Note that the average prevalence of the outcomes considered is usually low and around 4%. One concern might be that a linear regression may not fit the data well. To alleviate this concern, we estimate equation (1) using an analogous Logit model and compute the average marginal effect associated with an increase in one standard deviation of physicians' quality. Appendix Table A.12 shows that the marginal effects (signs and magnitudes) are very similar to the ones estimated using a linear regression model.

6.3 Impact across distribution of quality

While ordinary least squares allows us to compute the average effect of physicians' quality, it does not tell us much about the magnitude of this effect across the distribution of physicians' quality. We rank the score into quartiles and estimate equation (1) using a set

[56] Results are reported in Appendix Table A.10, where we use unhealthy, low birth weight, prematurity, and Apgar score as our dependent variables, using the standardized average health-related college examination score and standardized principal component as a proxy for physicians' quality, with and without controls. Also, note that Figure A.8 and Appendix Table A.10 use discrete LHC controls in the regressions that are indicators of low quality. Table A.14 shows that the results are almost identical if we control for the continuous measures of these lagged LHC control variables.

of dummies indicating the score distribution quartile to which physicians belonged. The results are presented in Appendix Table A.13. Columns (1) and (2) present the coefficients associated with the effect of belonging to the second, third, and fourth quartile of the distribution of the average of the health-related scores and the first principal component, respectively, on our main outcome, unhealthy, relative to the first quartile. Although we lack power to find statistically significant differences, we see that the point estimates are negative and monotonically decreasing with respect to the quartile. This suggests that there are potential gains associated with being assigned to more-skilled physician across the whole distribution of skills.

6.4 Alternative measures of quality

We extend our analysis by estimating our main specification using alternative measures of quality. We use the average of the four areas tested in the SABER PRO (health management, public health, reading, quantitative), as well as each individual score as proxies of physicians' quality before the SSO program. We regress unhealthy on the different proxies for physicians' quality, and show in Table A.6 that the scores have a negative effect on unhealthy and are not statistically different from each other. Finally, in Appendix Table A.15, we interact the average score with the university's (program) average score to test if top universities drive the estimated effect. We do not find evidence that top-ranked universities drive the effects presented before.

6.5 Weighted score

While the Ministry of Health (1990, 2001) specifies that the SSO physicians are responsible for the mother's care, including family planning and antenatal consultation, if the randomly assigned doctors are not the entire workforce of the LHCs, the coefficient in our main regression (equation 1) may underestimate the effect of physicians. The underestimation would depend on the fraction of the workforce in each LHC that arrives through the lottery.

To quantitatively explore this idea, we rescale our measure of physicians' quality, weighting the average score of physicians assigned to a LHC by the fraction of the workforce in each LHC that arrives via the lottery. Thus, if a LHC has a small share of doctors that are replaced by the lottery, the difference in mean outcomes caused by the quality difference of the SSO doctors will be much smaller. We excluded municipalities with more than two LHCs per municipality (10 municipalities), because we cannot identify the LHCs where non-SSO doctors work.[57]

[57] Appendix Table A.16 shows that our main results remain almost the same, excluding from our sample

Table A.17 reports the results of the estimation using the weighted score. The table shows that the results are very similar when the weighted score is used to proxy physicians' quality. These results are consistent with the fact that the median number of doctors per LHC is three, and around 95% of the LHCs have less than 20 doctors per LHC.

6.6 Exposure

Finally, since some newborns are more exposed to new physicians than others depending on the gestation period, we split the sample between cohorts of newborns who were fully and partially exposed by the lottery physicians. Table A.18 shows stronger impacts on the four outcomes studied for fully exposed newborns, but the difference is not statically significant at conventional levels.

7 Conclusions

Physicians are a key input in the production function of health at birth. Yet there is little evidence on the effect they can have on birth outcomes. The lack of causal evidence on this topic is related to the selection bias associated with the match between physicians and LHCs (Doyle et al., 2010). In the present study, we provide experimental evidence to answer this difficult question.

In Colombia, medical school graduates must spend the first year of their careers working in the national Mandatory Social Service program (SSO). The SSO program randomly assigns physicians to their first job, providing a test for the effects of being treated by a more-skilled physician. In this paper, we combine administrative records to match physicians in the SSO program, LHCs, vital statistics records, characteristics of the physicians, and mandatory health-specific college graduation exams to measure the skills of the physicians assigned to each LHC and the main health outcomes. Using these datasets, we provide evidence of the covariate balance between LHCs and the quality of physicians. Finally, we provide evidence of the causal relationship between more-skilled physicians and health at birth.

We find that being assigned to a more-skilled physician has a negative and significant effect on the probability of giving birth to an unhealthy child. We estimate that being assigned to a physician with a one standard deviation higher performance in the graduation health test score reduces the probability of giving birth to an unhealthy child by 6.31%.

the ten municipalities with more than two LHCs per municipality, using our main unweighted proxy for physicians' quality. Reassuringly, this restriction delivers results that are very similar to our baseline results (Appendix Table A.10).

Although unhealthy is our main measure of health at birth, the results are robust to other measures such as low birth weight, prematurity and Apgar score.

Furthermore, we explore whether being assigned to a more-skilled physicians increases the number of prenatal consultations, serving as a mechanism to improve the quality of care and health outcomes. According to WHO (2016) and the Colombian government, better and more frequent prenatal care during pregnancy improves health at birth. We find that more-skilled doctors do not schedule mothers for more prenatal checkups. Nonetheless, we provide evidence that these physicians are targeting their prenatal consultations toward the most vulnerable mothers, measured as those with the predicted likelihood of giving birth to an unhealthy baby.

Finally, we present several meaningful placebo tests. The results show the internal validity of our exercise. We conclude that more-skilled physicians play a crucial role in overall health at birth and that governments should consider these findings in developing policies to assign physicians optimally.

References

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review 106*(12), 3730–64.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Adhvaryu, A., A. Nyshadham, T. Molina, and J. Tamayo (2018). Helping children catch up: Early life shocks and the progresa experiment. Technical report, National Bureau of Economic Research.

Administrative Department of Statistics (2005). National census.

Administrative Department of Statistics (2018). Vital statistics records.

Alexander, G. R. and C. C. Korenbrot (1995). The role of prenatal care in preventing low birth weight. *The Future of Children*, 103–120.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics 120*(3), 1031–1083.

Almond, D., J. J. Doyle Jr, A. E. Kowalski, and H. Williams (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics 125*(2), 591–634.

Almond, D. and B. Mazumder (2011). Health capital and the prenatal environment: the effect of ramadan observance during pregnancy. *American Economic Journal: Applied Economics 3*(4), 56–85.

Alsan, M., O. Garrick, and G. Graziani (2019). Does diversity matter for health? experimental evidence from oakland. *American Economic Review 109*(12), 4071–4111.

Amarante, V., M. Manacorda, E. Miguel, and A. Vigorito (2016). Do cash transfers improve birth outcomes? evidence from matched vital statistics, program, and social security data. *American Economic Journal: Economic Policy 8*(2), 1–43.

Anderson, M., C. Dobkin, and T. Gross (2012). The effect of health insurance coverage on the use of medical services. *American Economic Journal: Economic Policy 4*(1), 1–27.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association 103*(484), 1481–1495.

Anderson, M. L., C. Dobkin, and T. Gross (2014). The effect of health insurance on emergency department visits: Evidence from an age-based eligibility threshold. *Review of Economics and Statistics 96*(1), 189–195.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics 132*(2), 871–919.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics 131*(3), 1415–1453.

Aron-Dine, A., L. Einav, A. Finkelstein, and M. Cullen (2015). Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics 97*(4), 725–741.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Baicker, K. and A. Chandra (2004). The productivity of physician specialization: evidence from the medicare program. *American Economic Review 94*(2), 357–361.

Barber, S. L. and P. J. Gertler (2010). Empowering women: how mexico's conditional cash transfer programme raised prenatal care quality and birth weight. *Journal of Development Effectiveness 2*(1), 51–73.

Bardach, N. S., J. J. Wang, S. F. De Leon, S. C. Shih, W. J. Boscardin, L. E. Goldman, and R. A. Dudley (2013). Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *Jama 310*(10), 1051–1059.

Basinga, P., P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy, and C. M. Vermeersch (2011). Effect on maternal and child health services in rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet 377*(9775), 1421–1428.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy 81*(4), 813–846.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Butler, A. S., R. E. Behrman, et al. (2007). *Preterm birth: causes, consequences, and prevention*. National Academies Press.

Callaghan, W. M., M. F. MacDorman, S. A. Rasmussen, C. Qin, and E. M. Lackritz (2006). The contribution of preterm birth to infant mortality rates in the united states. *Pediatrics 118*(4), 1566–1573.

Cameron, E. Z. (2004). Facultative adjustment of mammalian sex ratios in support of the trivers–willard hypothesis: evidence for a mechanism. *Proceedings of the Royal Society of London. Series B: Biological Sciences 271*(1549), 1723–1728.

Card, D., A. Fenizia, and D. Silver (2019). The health impacts of hospital delivery practices. Technical report, National Bureau of Economic Research.

Carrera, M., D. P. Goldman, G. Joyce, and N. Sood (2018). Do physicians respond to the costs and cost-sensitivity of their patients? *American Economic Journal: Economic Policy 10*(1), 113–52.

Carrillo, B. and J. Feres (2019). Provider supply, utilization, and infant health: evidence from a physician distribution policy. *American Economic Journal: Economic Policy 11*(3), 156–96.

Chan Jr, D. C., M. Gentzkow, and C. Yu (2019). Selection with variation in diagnostic skill: Evidence from radiologists. Technical report, National Bureau of Economic Research.

Chandra, A. and D. Staiger (2020). Identifying sources of inefficiency in healthcare. *The Quarterly Journal of Economics 135*(2), 785–843.

Chen, Y. (2021). Team-specific human capital and team performance: Evidence from doctors. *American Economic Review (forthcoming).*

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–2679.

Clemens, J. and J. D. Gottlieb (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review 104*(4), 1320–49.

Colegio Médico Colombiano (2018). Historia del servicio social obligatorio. Retrieved from: https://www.colegiomedicocolombiano.org/web_cmc/upload/docs/Epicrisis-7_web.pdf.

Colombian Institute for Educational Evaluation (2014). Quality evaluation of higher education.

Congress of Colombia (1993, December). Law 100 of 1993. por la cual se crea el sistema de seguridad social integral y se dictan otras disposiciones.

Congress of Colombia (2007, October). Law 1164 of 2007. por la cual se dictan disposiciones en materia del talento humano en salud.

Conway, K. S. and P. Deb (2005). Is prenatal care really ineffective? or, is the 'devil'in the distribution? *Journal of Health Economics 24*(3), 489–513.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature 47*(1), 87–122.

Currie, J. (2011). Inequality at birth: Some causes and consequences. *American Economic Review 101*(3), 1–22.

Currie, J. and D. Almond (2011). Human capital development before age five. In *Handbook of Labor Economics*, Volume 4, pp. 1315–1486. Elsevier.

Currie, J. and J. Grogger (2002). Medicaid expansions and welfare contractions: offsetting effects on prenatal care and infant health? *Journal of Health Economics 21*(2), 313–335.

Currie, J. and J. Gruber (1996). Saving babies: The efficacy and cost of recent changes in the medicaid eligibility of pregnant women. *Journal of Political Economy 104*(6), 1263–1296.

Currie, J. and W. B. MacLeod (2017). Diagnosing expertise: Human capital, decision making, and performance among physicians. *Journal of Labor Economics 35*(1), 1–43.

Currie, J. and W. B. MacLeod (2020). Understanding doctor decision making: The case of depression treatment. *Econometrica 88*(3), 847–878.

Currie, J. and H. Schwandt (2016a). The 9/11 dust cloud and pregnancy outcomes: a reconsideration. *Journal of Human Resources 51*(4), 805–831.

Currie, J. and H. Schwandt (2016b). Mortality inequality: The good news from a county-level approach. *Journal of Economic Perspectives 30*(2), 29–52.

Currie, J. and R. Walker (2011). Traffic congestion and infant health: Evidence from e-zpass. *American Economic Journal: Applied Economics 3*(1), 65–90.

Curtis, J. R., Q. Cai, S. W. Wade, B. S. Stolshek, J. L. Adams, A. Balasubramanian, H. N. Viswanathan, and J. D. Kallich (2013). Osteoporosis medication adherence: physician perceptions vs. patients' utilization. *Bone 55*(1), 1–6.

Cutler, D., J. S. Skinner, A. D. Stern, and D. Wennberg (2019). Physician beliefs and patient preferences: a new look at regional variation in health care spending. *American Economic Journal: Economic Policy 11*(1), 192–221.

Das, J. and J. Hammer (2005). Which doctor? combining vignettes and item response to measure clinical competence. *Journal of Development Economics 78*(2), 348–383.

Das, J. and J. Hammer (2007). Money for nothing: the dire straits of medical practice in delhi, india. *Journal of Development Economics 83*(1), 1–36.

Das, J., J. Hammer, and K. Leonard (2008). The quality of medical advice in low-income countries. *Journal of Economic Perspectives 22*(2), 93–114.

Das, J., A. Holla, A. Mohpal, and K. Muralidharan (2016). Quality and accountability in health care delivery: audit-study evidence from primary care in india. *American Economic Review 106*(12), 3765–99.

Das, J. and T. P. Sohnesen (2007). Variations in doctor effort: Evidence from paraguay: Doctors in paraguay who expended less effort appear to have been paid more than doctors who expended more. *Health Affairs 26*(Suppl2), w324–w337.

Dinkelman, T. (2017). Long-run health repercussions of drought shocks: Evidence from south african homelands. *The Economic Journal 127*(604), 1906–1939.

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010). Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of Health Economics 29*(6), 866–882.

Doyle, J. J., J. A. Graves, J. Gruber, and S. A. Kleiner (2015). Measuring returns to hospital care: Evidence from ambulance referral patterns. *Journal of Political Economy 123*(1), 170–214.

Ehrenstein, V. (2009). Association of apgar scores with death and neurologic disability. *Clinical Epidemiology 1*, 45.

Eriksson, J. G., E. Kajantie, C. Osmond, K. Thornburg, and D. J. Barker (2010). Boys live dangerously in the womb. *American Journal of Human Biology 22*(3), 330–335.

Fernández Ávila, D. G., L. C. Mancipe García, D. C. Fernández Ávila, E. Reyes Sanmiguel, M. C. Díaz, and J. M. Gutiérrez (2011). Analysis of the supply of medicine undergraduate programs in colombia, during the past 30 years. *Revista Colombiana de Reumatología 18*(2), 109–120.

Finkelstein, A., M. Gentzkow, and H. Williams (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics 131*(4), 1681–1726.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Neuse, H. Allen, K. Baicker, and O. H. S. Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly Journal of Economics 127*(3), 1057–1106.

Fletcher, J. M., L. I. Horwitz, and E. Bradley (2014). Estimating the value added of attending physicians on patient outcomes. Technical report, National Bureau of Economic Research.

Gagnon-Bartsch, J., Y. Shem-Tov, et al. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics 13*(3), 1464–1483.

Gomez, P., I. Arevalo, et al. (2013). Guías de práctica clínica para la prevención, detección temprana y tratamiento de las complicaciones del embarazo, parto y puerperio. *Ministerio de Salud y protección social Colombia 84*, 74–82.

Gonzalez, R. M. and D. Gilleskie (2017). Infant mortality rate as a measure of a country's health: a robust method to improve reliability and comparability. *Demography 54*(2), 701–720.

Grossman, M. and T. J. Joyce (1990). Unobservables, pregnancy resolutions, and birth weight production functions in new york city. *Journal of Political Economy 98*(5, Part 1), 983–1007.

Guarin, A., C. Posso, E. Saravia, and J. Tamayo (2023). Healing the gender gap: The impacts of randomized first-job on female physicians.

Harris, D. N. and T. R. Sass (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review 40*, 183–204.

Herrmann, M., E. Walsh, and E. Isenberg (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy 3*(1), 1–10.

Ho, K. and A. Pakes (2014a). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review 104*(12), 3841–84.

Ho, K. and A. Pakes (2014b). Physician payment reform and hospital referrals. *American Economic Review 104*(5), 200–205.

Hoynes, H., M. Page, and A. H. Stevens (2011). Can targeted transfers improve birth outcomes?: Evidence from the introduction of the wic program. *Journal of Public Economics 95*(7-8), 813–827.

Iizuka, T. (2012). Physician agency and adoption of generic pharmaceuticals. *American Economic Review 102*(6), 2826–58.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review 27*(6), 615–631.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–95.

Kraemer, S. (2000). The fragile male. *Bmj 321*(7276), 1609–1612.

Kramer, M. S. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization 65*(5), 663.

Kremer, M. (1993). The o-ring theory of economic development. *The Quarterly Journal of Economics 108*(3), 551–575.

Leonard, K. L. and M. C. Masatu (2007). Variations in the quality of care accessible to rural communities in tanzania: Some quality disparities might be amenable to policies that do not necessarily relate to funding levels. *Health Affairs 26*(Suppl2), w380–w392.

Leonard, K. L., M. C. Masatu, and A. Vialou (2007). Getting doctors to do their best the roles of ability and motivation in health care quality. *Journal of Human Resources 42*(3), 682–700.

Liberman, A., C. Medina, C. Neilson, and C. Posso (2021). Lender market power and the bright side of interest rate caps: Evidence from colombia. Technical report, Unpublished manuscript.

Liberman, A., C. Neilson, L. Opazo, and S. Zimmerman (2018). The equilibrium effects of information deletion: Evidence from consumer credit markets. Technical report, National Bureau of Economic Research.

Lin, W. (2009). Why has the health inequality among infants in the us declined? accounting for the shrinking gap. *Health Economics 18*(7), 823–841.

Mathews, F., P. J. Johnson, and A. Neil (2008). You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. *Proceedings of the Royal Society B: Biological Sciences 275*(1643), 1661–1668.

McCormick, M. C. and J. E. Siegel (2001). Recent evidence on the effectiveness of prenatal care. *Ambulatory Pediatrics 1*(6), 321–325.

Michalopoulos, C., D. Wittenburg, D. A. Israel, and A. Warren (2012). The effects of health care benefits on health care use and health: a randomized trial for disability insurance beneficiaries. *Medical Care*, 764–771.

Ministry of Health (1990, June). Decree 1335 of 1990. por el cual se expide parcialmente el manual general de funciones y requisitos del subsector oficial del sector salud.

Ministry of Health (2001). Reglamento del año de servicio de salud rural.

Ministry of Health (2010, March). Resolution 1058 of 2010. por medio de la cual se reglamenta el servicio social obligatorio para los egresados de los programas de educación superior del área de la salud y se dictan otras disposiciones.

Ministry of Health (2012a, December). Resolution 4503 of 2012. por la cual se modifica el artículo 6 de la resolución 274 de 2011 modificado por el artículo 2 de la resolución 566 de 2012.

Ministry of Health (2012b, March). Resolution 566 of 2012. por la cual se modifica parcialmente la resolución 274 de 2011.

Ministry of Health (2013, May). Resolution 1441 of 2013. por la cual se definen los procedimientos y condiciones que deben cumplir los prestadores de servicios de salud.

Ministry of Health (2014). Reports of professionals registered and assigned to the process of assigning places in the mandatory social service.

Mizuno, R. (2000). The male/female ratio of fetal deaths and births in japan. *The Lancet 356*(9231), 738–739.

Molitor, D. (2018). The evolution of physician practice styles: evidence from cardiologist migration. *American Economic Journal: Economic Policy 10*(1), 326–56.

Moore, E. A., F. Harris, K. R. Laurens, M. J. Green, S. Brinkman, R. K. Lenroot, and V. J. Carr (2014). Birth outcomes and academic achievement in childhood: A population record linkage study. *Journal of Early Childhood Research 12*(3), 234–250.

Moster, D., R. Lie, and T. Markestad (2002). Joint association of apgar scores and early neonatal symptoms with minor disabilities at school age. *Archives of Disease in Childhood-Fetal and Neonatal Edition 86*(1), F16–F21.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Naeye, R. L., L. S. Burt, D. L. Wright, W. A. Blanc, and D. Tatter (1971). Neonatal mortality, the male disadvantage. *Pediatrics 48*(6), 902–906.

Norcini, J. J., J. R. Boulet, A. Opalek, and W. D. Dauphinee (2014). The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Academic Medicine 89*(8), 1157–1162.

Norcini, J. J., R. S. Lipner, and H. R. Kimball (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical education 36*(9), 853–859.

Okeke, E. N. (2023, March). When a doctor falls from the sky: The impact of easing doctor supply constraints on mortality. *American Economic Review 113*(3), 585–627.

Okeke, E. N. and I. S. Abubakar (2020). Healthcare at the beginning of life and child survival: Evidence from a cash transfer experiment in nigeria. *Journal of Development Economics 143*, 102426.

Oreopoulos, P., M. Stabile, R. Walld, and L. L. Roos (2008). Short-, medium-, and long-term consequences of poor infant health an analysis using siblings and twins. *Journal of Human Resources 43*(1), 88–138.

Páez, G., L. Jaramillo, C. Franco, and L. Arregoces (2007). Estudio sobre el modo de gestionar la salud en colombia.

Persson, P. and M. Rossin-Slater (2018). Family ruptures, stress, and the mental health of the next generation. *American Economic Review 108*(4-5), 1214–52.

Pongou, R., B. Kuate Defo, and Z. Tsala Dimbuene (2017). Excess male infant mortality: The gene-institution interactions. *American Economic Review 107*(5), 541–45.

Razaz, N., W. T. Boyce, M. Brownell, D. Jutte, H. Tremlett, R. A. Marrie, and K. Joseph (2016). Five-minute apgar score as a marker for developmental vulnerability at 5 years of age. *Archives of Disease in Childhood-Fetal and Neonatal Edition 101*(2), F114–F120.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review 107*(6), 1656–84.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Schnell, M. and J. Currie (2018). Addressing the opioid epidemic: is there a role for physician education? *American Journal of Health Economics 4*(3), 383–410.

Shimer, R. and L. Smith (2000). Assortative matching and search. *Econometrica 68*(2), 343–369.

Simeonova, E., N. Skipper, and P. R. Thingholm (2020). Physician health management skills and patient outcomes. Technical report, National Bureau of Economic Research.

Skinner, J. (2011). Causes and consequences of regional variations in health care. In *Handbook of health economics*, Volume 2, pp. 45–93. Elsevier.

Tamblyn, R., M. Abrahamowicz, D. Dauphinee, E. Wenghofer, A. Jacques, D. Klass, S. Smee, D. Blackmore, N. Winslade, N. Girard, et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *Jama 298*(9), 993–1001.

Tamblyn, R., M. Abrahamowicz, W. D. Dauphinee, J. A. Hanley, J. Norcini, N. Girard, P. Grand'Maison, and C. Brailovsky (2002). Association between licensure examination scores and practice in primary care. *Jama 288*(23), 3019–3026.

Taylor, H. G., N. Klein, N. M. Minich, and M. Hack (2001). Long-term family outcomes for children with very low birth weights. *Archives of Pediatrics & Adolescent Medicine 155*(2), 155–161.

Tsugawa, Y., A. B. Jena, J. F. Figueroa, E. J. Orav, D. M. Blumenthal, and A. K. Jha (2017). Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. *JAMA Internal Medicine 177*(2), 206–213.

Universidad del Rosario (2015). El año rural: Realidad agridulce para los médicos recién graduados. un relato de quien lo vivió. Retrieved from: https://www.urosario.edu.co/Revista-Nova-Et-Vetera/Vol-1-Ed-2/Cultura/El-ano-rural-Realidad-agridulce-para-los-medicos-r.pdf.

Veddovi, M., D. T. Kenny, F. Gibson, J. Bowen, and D. Starte (2001). The relationship between depressive symptoms following premature birth, mothers' coping style, and knowledge of infant development. *Journal of Reproductive and Infant Psychology 19*(4), 313–323.

Wenghofer, E., D. Klass, M. Abrahamowicz, D. Dauphinee, A. Jacques, S. Smee, D. Blackmore, N. Winslade, K. Reidel, I. Bartman, et al. (2009). Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical education 43*(12), 1166–1173.

WHO (2016). Pregnant women must be able to access the right care at the right time, says who. Retrieved from: https://www.who.int/news/item/07-11-2016-pregnant-women-must-be-able-to-access-the-right-care-at-the-right-time-says-who.

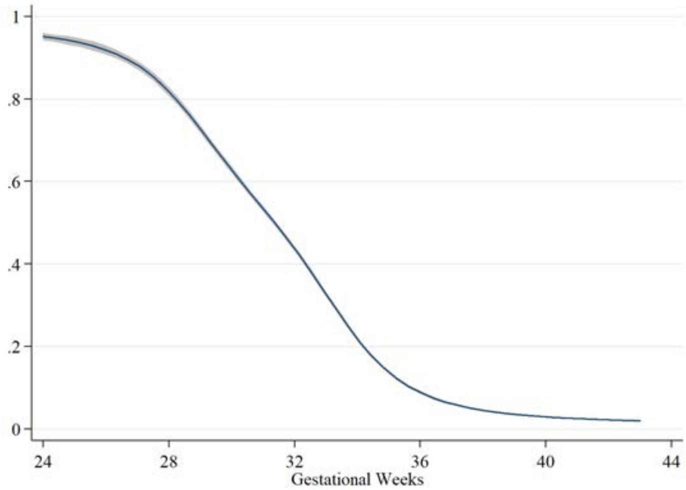
Woodcock, S. D. (2008). Wage differentials in the presence of unobserved worker, firm, and match heterogeneity. *Labour Economics 15*(4), 771–793.

Online Appendix

Not for Publication

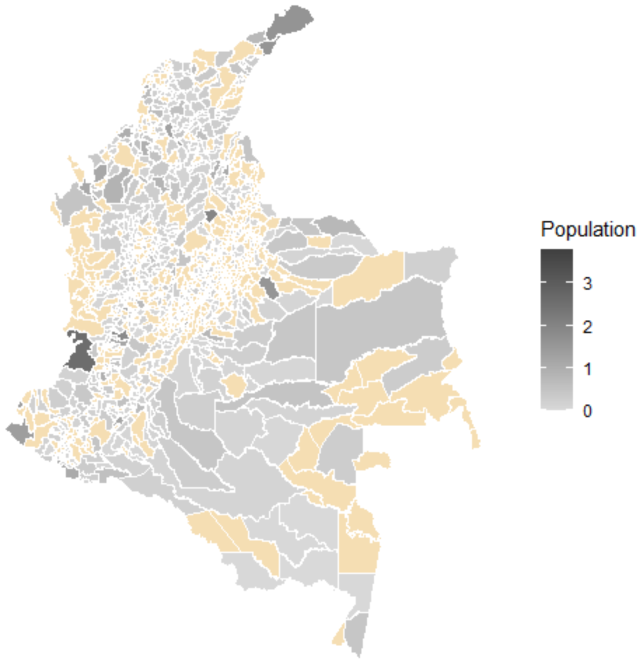
A Appendix

Figure A.1: Probability of low birth weight vs. gestational weeks, 2009-2012



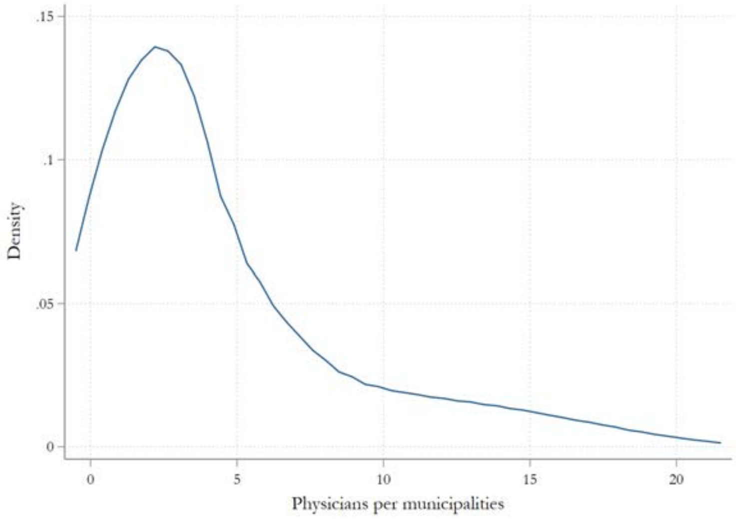
Notes: This figure presents the local polynomial regression fit of the probability of having low birth weight over the number of gestational weeks using all birth records for Colombia from 2009 to 2012.

Figure A.2: Population (per 100,000) for municipalities included in our main sample



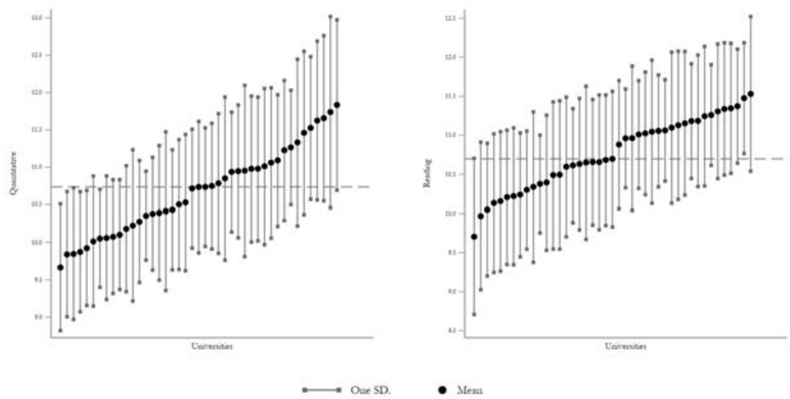
Notes: This figure presents the map of the population per 100,000 people for the municipalities included in our main sample in 2005. The municipalities in orange are not included in our sample or do not have SSO.

Figure A.3: Distribution of physicians per municipalities



Notes: This figure shows the distribution of physicians per municipality for the sample of 590 municipalities with only one LHC. The data spans from January 2012 to December 2012.

Figure A.4: Heterogeneity in quantitative and reading SABER PRO scores



Notes: This figure reports the quantitative and reading test scores for the universities that the physicians in our sample attended. Data accounts for 44 different universities. The figure shows the mean score for each university/program and an interval of one standard deviation to each side of the average. The dashed horizontal line represents the overall percentile 50. The figure shows substantial heterogeneity both within and between programs. For all the fields reported, there is a difference of almost two standard deviations between the averages of the best and the worst programs.

Table A.1: Summary statistics - physicians in the main sample

Covariate	**Mean**	**Standard error**
Sex (female)	0.558	0.497
The household has a private car	0.483	0.500
Number of people in the household	4.025	1.659
Father with tertiary education	0.644	0.479
Mother with tertiary education	0.634	0.482
Socioeconomic strata: 1 or 2 or rural areas	0.292	0.455
Socioeconomic strata: 4, 5 or 6	0.349	0.477
The household has internet	0.831	0.375
Monthly household income: Less than 2 MW	0.229	0.420
Monthly household income: between 2 and 3 MW	0.220	0.414
The father or the mother has a job	0.872	0.335
The household has a washing machine	0.854	0.353
The household has a television	0.859	0.348
The household has a cellphone	0.963	0.188
The house has proper flooring	0.908	0.289
The household has an oven	0.671	0.470
Physician's score on the Health care test	10.426	1.059
Physician's score on the Disease prevention test	10.431	1.010
Physician's score on the Reading test	10.624	1.007
Physician's score on the Math test	10.572	1.123
Physician's average score on SABER PRO	10.513	0.854
Observations	**2,126**	

Notes: This table reports the summary statistics for the physicians included in our main sample. These characteristics are obtained at the time physicians took their SABER PRO exam (before the SSO). "Sex" is a binary variable that takes the value of 1 if the physician is female and zero otherwise; "The household has a private car" that takes the value of 1 if the household of the physician owns a private car at the time the physician took the SABER PRO test and zero otherwise; 'Number of people in the household" counts the number of individuals living in the same house as the physician; "Father with tertiary education" is a binary variable that takes the value of 1 if the physician's father has at least tertiary education and zero otherwise; "Mother with tertiary education" is a binary variable that takes the value of 1 if the physician's mother has at least tertiary education and zero otherwise; "Socioeconomic strata: 1 or 2 or rural areas" takes the value of 1 if the physician's household's socioeconomic strata at the time of the SABER PRO test was 1, 2 or rural and zero otherwise; "Socioeconomic strata: 4, 5 or 6" is a variable that takes the value of 1 if the physician's household's socioeconomic strata at the time of the SABER PRO test was 4, 5 or 6 and zero otherwise; "The household has internet" takes the value of 1 if the physician had internet service at home at the time of the test; "Monthly household income: Less than 2MW" takes the value of 1 if the physician's household had an income lower than 2 times the national monthly minimum wage and zero otherwise; "Monthly household income: between 2 and 3 MW" takes the value of 1 if the physician's household had an income between 2 and 3 times the national monthly minimum wage and zero otherwise; "The father or the mother has a job" takes value 1 if either of the physician's parents have a job; The household has a washing machine, television, cellphone, proper flooring or oven, take value 1 if the household has that characteristic described and zero otherwise; physician's scores are continuous variables of the score obtained on each SABER PRO test subgroup; physician's average score on SABER PRO is the average of the four main components of the test, health care, disease prevention, reading and math.

Table A.2: Covariate balance at LHC level using all the areas tested in the SABER PRO

Covariate	**Coefficient**	**Standar Error**
Unhealthy	0.001	0.001
Low birth weight	0.000	0.001
Prematurity	0.000	0.007
Apgar < 7	0.003	0.009
Antenatal consultations < 4 (Prop.)	0.000	0.003
Female newborn	0.000	0.001
Mother with basic education	-0.002	0.003
Married mother	0.001	0.002
Teenage mother	0.000	0.002
Number of LHCs by municipalities	0.000	0.010
Municipality population	325.7	1,032.3

Notes: This table presents the results of eleven different LHC-by-cohort level regressions of the LHC-level variables, listed in the first column, on the physicians' quality measure (proxied by the average of the four fields of the college examination scores) and the draw-by-state fixed effects. The coefficient and the standard error of the physicians' quality variable are reported in the second and third columns, respectively. Standard errors are clustered at the LHC level. LHCs' characteristics come from the 2010-2012 DANE VSR. "Low birth weight" is the proportion of newborns with low birth weight (weight <2,500 grams); "Prematurity" is the proportion of newborns who were premature (fewer than 37 weeks of gestation); "Apgar < 7" is the proportion of newborns whose Apgar score is lower than 7; "Unhealthy" is the proportion of newborns with at least one of the three previous conditions. "Antenatal consultations < 4" is the proportion of mothers who had less than four antenatal visits (Insufficient antenatal consultations); "Female newborn" is the proportion of female newborns; "Mother with basic education" is the proportion of mothers with 5 or more years of education at the time of the birth; "Married mother" is the proportion of mothers that are married at the time of the birth; and "Teenage mother" is the proportion of mothers that are 19 years old or younger at the time of the birth. We interpret the non-significance of these estimates as evidence in favor of the randomness of the assignment of physicians.

Table A.3: Placebo other years

	Unhealthy		LBW		Prematurity		Apgar < 7	
	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (1)	PCA Health Scores (2)
Panel A. 2 years								
Coefficient	-0.0010	-0.0010	-0.0015	-0.0015	-0.0005	-0.0005	-0.0007	-0.0007
Stand. Err.	(0.0020)	(0.0020)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0014)	(0.0014)
Relative effect	0.10%	0.10%	0.04%	0.04%	0.04%	0.04%	0.04%	0.04%
Panel B. 2.5 years								
Coefficient	-0.0016	-0.0016	-0.0004	-0.0004	-0.0013	-0.0013	-0.0010	-0.0010
Standard Error	(0.0017)	(0.0018)	(0.0010)	(0.0010)	(0.0012)	(0.0013)	(0.0014)	(0.0014)
Relative effect	0.10%	0.10%	0.05%	0.05%	0.04%	0.04%	0.04%	0.04%
Panel C. 3 years								
Coefficient	-0.0022	-0.0022	-0.0010	-0.0011	-0.0011	-0.0012	-0.0012	-0.0012
Standard Error	(0.0018)	(0.0018)	(0.0010)	(0.0010)	(0.0012)	(0.0012)	(0.0014)	(0.0014)
Relative effect	0.11%	0.11%	0.05%	0.05%	0.04%	0.04%	0.04%	0.04%
Panel D. 3.5 years								
Coefficient	-0.0004	-0.0004	-0.0004	-0.0005	-0.0008	-0.0008	0.0000	0.0000
Standard Error	(0.0019)	(0.0019)	(0.0009)	(0.0009)	(0.0013)	(0.0013)	(0.0014)	(0.0014)
Relative effect	0.11%	0.11%	0.05%	0.05%	0.04%	0.04%	0.04%	0.04%

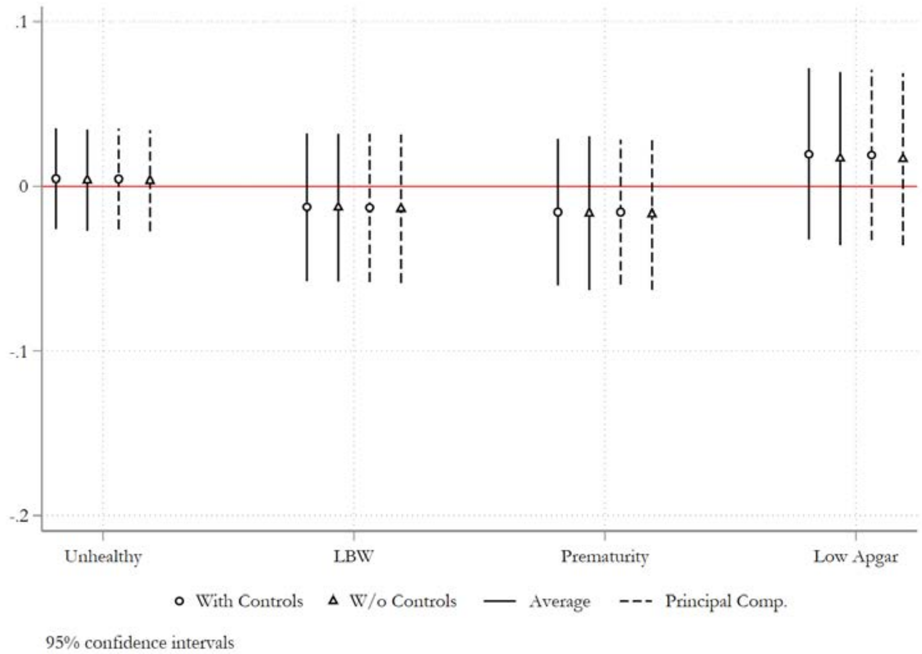
Notes: This table presents the placebo exercise for windows of 3.5, 3, 2.5 and 2 years before the start of the SSO program. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score and the first principal component of the health score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. The results show that regardless of the time window that we use for the calculation of the placebo test, the estimated coefficients are always precisely estimated zeros which we interpret as evidence of the randomness of the assignment of the physicians to the LHCs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Placebo robustness checks

	Unhealthy		LBW		Prematurity		Apgar < 7	
	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (3)	PCA Health Scores (4)	Average Health Scores (5)	PCA Health Scores (6)	Average Health Scores (7)	PCA Health Scores (8)
Panel A. Without controls								
Coefficient	-0.0009	-0.0010	-0.0008	-0.0009	-0.0020	-0.0020	0.0004	0.0003
Stand. Err.	(0.0022)	(0.0022)	(0.0011)	(0.0011)	(0.0016)	(0.0016)	(0.0013)	(0.0013)
Relative effect	-0.79%	-0.83%	-1.76%	-1.85%	-3.76%	-3.79%	0.78%	0.73%
Panel B. With controls								
Coefficient	0.0003	0.0002	-0.0001	-0.0002	-0.0014	-0.0014	0.0008	0.0007
Standard Error	(0.0019)	(0.0019)	(0.0008)	(0.0008)	(0.0015)	(0.0014)	(0.0012)	(0.0012)
Relative effect	0.22%	0.18%	-0.26%	-0.36%	-2.62%	-2.62%	1.66%	1.59%
Average Dependent Variable	0.119		0.046		0.052		0.047	
Number of Observations	262,089							

Notes: This table presents the placebo exercise for the main results. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score and the first principal component of the health scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure A.5: Placebo using all samples and average scores



Notes: This figure shows the results of running an exercise analogous to the one presented in Figure A.8 but moving the arrival date of the physician four years back (years 2009-2012). The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score or the first principal component of the health tests). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Low Apgar" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. These results support the ones presented in Table 3 on the robustness of the estimated zero effect for the placebo tests.

Table A.5: Placebo estimating a Logit model

	Unhealthy		**LBW**		**Prematurity**		**Apgar < 7**	
	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (3)	PCA Health Scores (4)	Average Health Scores (5)	PCA Health Scores (6)	Average Health Scores (7)	PCA Health Scores (8)
				Panel A. Without controls				
Coefficient	-0.0009	-0.0010	-0.0008	-0.0008	-0.0020	-0.0020	0.0004	0.0003
Stand. Err.	(0.0022)	(0.0022)	(0.0010)	(0.0010)	(0.0015)	(0.0015)	(0.0013)	(0.0013)
Relative effect	-0.80%	-0.83%	-1.66%	-1.74%	-3.81%	-3.84%	0.80%	0.74%
				Panel B. With controls				
Coefficient	0.0004	0.0003	-0.0001	-0.0001	-0.0010	-0.0010	0.0008	0.0007
Standard Error	(0.0019)	(0.0019)	(0.0008)	(0.0008)	(0.0015)	(0.0015)	(0.0012)	(0.0012)
Relative effect	0.31%	0.27%	-0.24%	-0.31%	-1.88%	-1.86%	1.64%	1.57%
Average Dependent Variable	0.119		0.047		0.052		0.047	
Number of Observations				261,820				

Notes: This table presents the placebo exercise for the main results using a Logit. The coefficients represent the average marginal effect of being assigned a physician with one standard deviation higher quality (proxied by the average score and the first principal component of the health scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors.

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

Table A.6: Main estimates using all the areas tested in the SABER PRO

	Unhealthy						
	Average All	Average Health Scores	Health Care Score	Prevention Disease Score	Average Academic Scores	Reading Score	Quantitative Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A. Without controls			
Coefficient	-0.0072***	-0.0062***	-0.0058***	-0.0049**	-0.0065***	-0.0023	-0.0022
Stand. Err.	(0.0020)	(0.0021)	(0.0020)	(0.0021)	(0.0021)	(0.0019)	(0.0018)
Relative effect	-7.60%	-6.52%	-6.06%	-5.13%	-6.81%	-2.47%	-2.31%
				Panel B. With controls			
Coefficient	-0.0068***	-0.0059***	-0.0053***	-0.0050***	-0.0059***	-0.0035*	-0.0032*
Standard Error	(0.0018)	(0.0019)	(0.0018)	(0.0019)	(0.0017)	(0.0018)	(0.0017)
Relative effect	-7.12%	-6.24%	-5.63%	-5.24%	-6.26%	-3.64%	-3.41%
Average Dependent Variable				0.095			
Number of Observations				256,805			

Notes: This table presents the main results using all areas tested in the SABER PRO. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by all areas tested in the SABER PRO). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status; number of inhabitants in the municipality; number of LHCs per municipality; area; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using the average of the four areas tested in the SABER PRO (health management, public health, reading, quantitative) as well as each individual (except for reading) score as proxies of the physician's skills before the SSO program. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: Cohort-level mortality estimates

	Fetal deaths		Fetal and neonatal deaths		Infant Mortality Ratio	
	Average Heallth Scores	PCA Heallth Scores	Average Heallth Scores	PCA Heallth Scores	Average Heallth Scores	PCA Heallth Scores
Coefficient	-0.7273	-0.4933	-0.7481	-0.5062	-0.0006	-0.0004
Stand. Err.	(2.3802)	(1.5872)	(2.4006)	(1.6006)	(0.0015)	(0.0010)
Relative effect	-4.94%	-3.35%	-4.76%	-3.22%	-2.35%	-1.56%
Average Dependent Variable	14.73		15.70		0.03	
Number of Observations	1,078		1,078		1,078	

Notes: This table presents the results of six different regressions at the cohort level for three distinct outcomes: fetal deaths, which is the total number of fetal deaths registered in a LHC during the timeframe when the cohort was assigned to that LHC; fetal and neonatal deaths, representing the total number of fetal deaths and mortalities of children under one year old registered in a LHC during the cohort's assignment period (ideally, we would have preferred to focus on shorter-term mortality, but under one year was the most granular definition of infant mortality available in our data); Infant Mortality Ratio, which is the number of fetal and neonatal deaths divided by the sum of births and fetal and neonatal deaths in a LHC during the cohort's assignment period. These variables are regressed on either the cohort's average health score (columns 1, 3, 5) or the cohort's PCA for the health scores (columns 2, 4, 6). The coefficients represent the effect on LHC mortality of being assigned physicians with a quality one standard deviation higher, as proxied by areas tested in the SABER PRO. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Numbers in parentheses are LHC-level clustered standard errors. We restrict to cohorts assigned to LHC where there are at least 5 births during their assignment period, but the results are similar when this threshold is increased/decreased/ignored. While these results are expected to be subject to high measurement error attenuation bias, we still observe negative, albeit not statistically significant, point estimates, which aligns with our main results.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Other heterogeneous effects

	Dependent variable: Unhealthy					
	Mother with low education (1)	Mother with high education (2)	Married mother (3)	Single mother (4)	Female newborns (5)	Male newborns (6)
Panel A. Without controls						
Coefficient	-0.0064***	-0.0056**	-0.0060***	-0.0062***	-0.0053***	-0.0067***
Stand. Err.	(0.0020)	(0.0022)	(0.0019)	(0.0023)	(0.0019)	(0.0021)
Relative effect	-6.48%	-5.93%	-6.97%	-6.52%	-5.65%	-7.09%
Panel B. With controls						
Coefficient	-0.0068***	-0.0059***	-0.0063***	-0.0066***	-0.0055***	-0.0072***
Standard Error	(0.0017)	(0.0017)	(0.0016)	(0.0020)	(0.0016)	(0.0018)
Relative effect	-6.80%	-6.26%	-7.32%	-6.90%	-5.92%	-7.53%
Average Dependent Variable	0.099	0.095	0.086	0.095	0.093	0.095
Number of Observations	101,556					

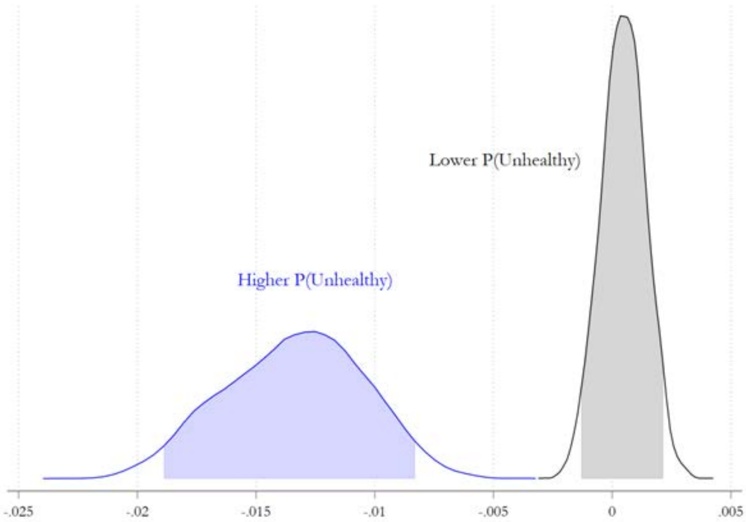
Notes: This table presents the main estimates by mother and gender of the newborn heterogeneous effects. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: Antenatal consultations < 4

	Average Score (1)	PCA Score (2)
	Panel A. Without controls	
Coefficient	-0.0019	-0.0022
Stand. Err.	(0.0070)	(0.0071)
Relative effect	-1.19%	-1.36%
	Panel B. With controls	
Coefficient	-0.0029	-0.0032
Standard Error	(0.0067)	(0.0068)
Relative effect	-1.79%	-1.96%
Average Dependent Variable	0.163	
Number of Observations	256,805	

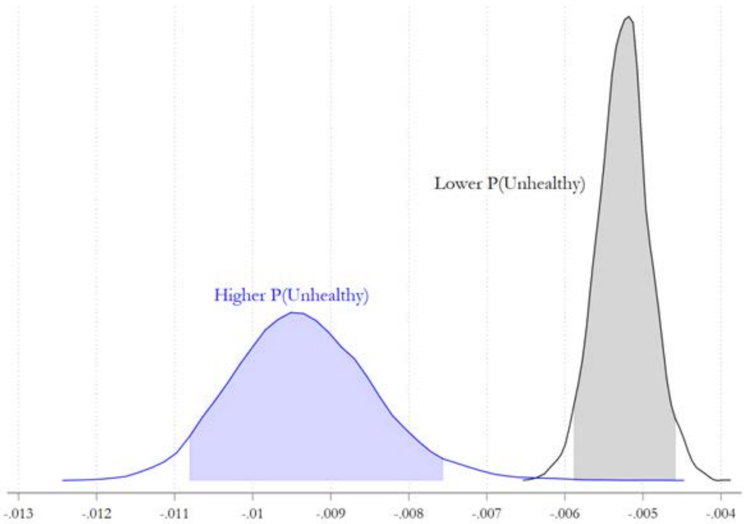
Notes: This table presents the results for antenatal consultations. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score or the first principal component of the health test scores) on the probability that mothers are scheduled for less than four prenatal checkups (Insufficient antenatal consultations). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Antenatal consultations < 4" takes value one if the mother attended to less than 4 consultations while pregnant, an zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors. The results show there is not a significant average effect of more-skilled doctors on the probability that mothers are scheduled for less than four prenatal checkups.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure A.6: Distribution of Logit simulations on antenatal consultations by predicted probability of unhealthy newborn



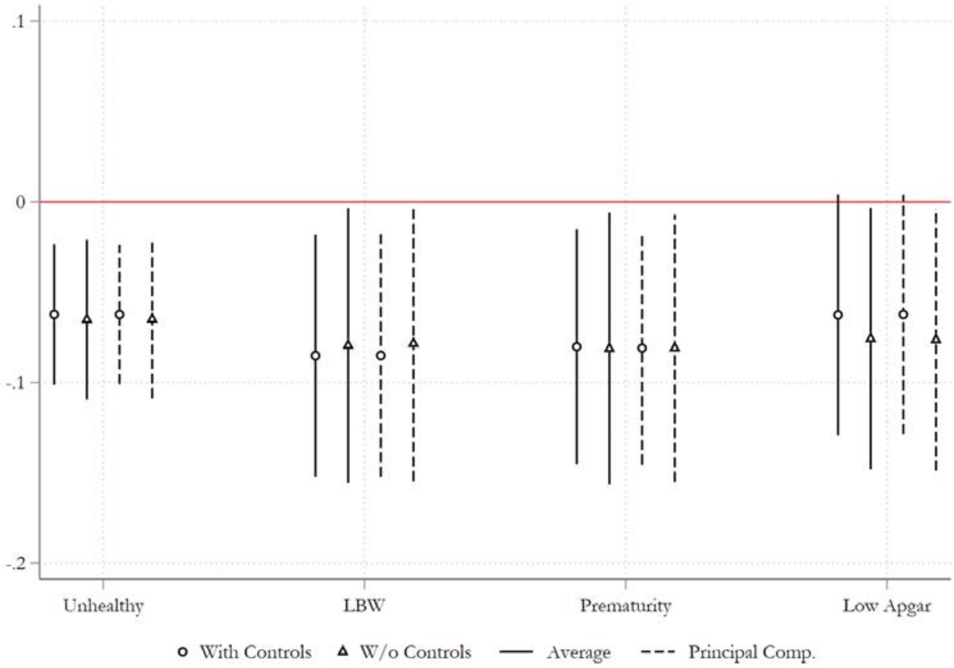
Notes: This figure plots the distribution of the estimated effects of physicians on antenatal consultations by mother's predicted probability of giving birth to an unhealthy child from 5,000 different random repetitions. In each of the 5,000 repetitions, to predict the probability of an unhealthy child, we divided our data into training and testing subsets of randomly selected LHCs using a K-mean algorithm. On the training sets, we run a Logit model of the probability of being born unhealthy on our usual set of mother and LHC ex-ante covariates, and use the estimations to predict the probability of giving birth to an unhealthy child on each testing subset. Using the prediction on the testing sample, we divide each subset into high and low predicted probability of giving birth to an unhealthy child, defined as mothers with a probability of an unhealthy child below and above the median, respectively. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise. The plotted coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) on the probability of having insufficient (less than four) antenatal consultations. All regressions control for draw-by-state fixed effects. The figure shows that there is almost no overlap between the distributions and that most of the mass of the distribution for the coefficient associated with the low predicted Unhealthy is around zero. This is consistent with the idea that more skilled physicians are better at targeting the care towards the more vulnerable mothers.

Figure A.7: Distribution of the coefficient dlogit simulations on the probability of being born unhealthy by the (ex-ante) predicted probability of an unhealthy newborn



Notes: This figure plots the distribution of the estimated effects of physicians on the probability of being born unhealthy by mother's predicted probability of giving birth to an unhealthy child from 5,000 different random repetitions. In each of the 5,000 repetitions, to predict the probability of an unhealthy child, we divided our data into training and testing subsets of randomly selected LHCs using a K-mean algorithm. On the training sets, we run a Logit model of the probability of being born unhealthy on our usual set of mother and LHC ex-ante covariates, and use the estimations to predict the probability of giving birth to an unhealthy child on each testing subset. Using the prediction on the testing sample, we divide each subset into high and low predicted probability of giving birth to an unhealthy child, defined as mothers with a probability of an unhealthy child below and above the median, respectively. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise. The plotted coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) on the probability of having insufficient (less than four) antenatal consultations. All regressions control for draw-by-state fixed effects. The figure shows that there is almost no overlap between the distributions and that the estimated effects of the more skilled physicians are consistently stronger for the population with higher predicted probability of being born unhealthy.

Figure A.8: Main estimates using all sample



Notes: This figure presents the coefficients for the relative effect of being assigned a physician with one standard deviation higher quality (proxied by the average score or the first principal component of the health test scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. “Unhealthy” is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; “LBW” is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; “Prematurity” is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; “Low Apgar” is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Regressions for the coefficients labeled as “With controls” also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure physicians’ skills (Averages vs. principal components). Standard errors are clustered at the LHC level. 95% confidence intervals.

Table A.10: Main estimates without and with controls

	Unhealthy		**LBW**		**Prematurity**		**Apgar < 7**	
	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (3)	PCA Health Scores (4)	Average Health Scores (5)	PCA Health Scores (6)	Average Health Scores (7)	PCA Health Scores (8)
				Panel A. Without controls				
Coefficient	-0.0060***	-0.0060***	-0.0033**	-0.0032**	-0.0033**	-0.0032**	-0.0027**	-0.0027**
Stand. Err.	(0.0020)	(0.0020)	(0.0016)	(0.0016)	(0.0015)	(0.0015)	(0.0013)	(0.0013)
Relative effect	-6.31%	-6.28%	-7.71%	-7.59%	-7.97%	-7.92%	-7.16%	-7.21%
				Panel B. With controls				
Coefficient	-0.0057***	-0.0057***	-0.0035**	-0.0035**	-0.0032**	-0.0032**	-0.0022*	-0.0022*
Stand. Err.	(0.0017)	(0.0017)	(0.0014)	(0.0014)	(0.0013)	(0.0013)	(0.0012)	(0.0012)
Relative effect	-6.03%	-6.01%	-8.23%	-8.20%	-7.77%	-7.82%	-5.95%	-5.91%
Average Dependent Variable	0.095		0.043		0.041		0.037	
Number of Observations				256,805				

Notes: This table presents the main results with and without controls. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score of the first principal component of the scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure quality. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: Main results using covariance index (Anderson, 2008)

	Unhealthy Cov index		**Unhealthy standarized**	
	Average Scores	PCA Scores	Average Scores	PCA Scores
	Panel A. Without controls			
Coefficient	-0.0160***	-0.0160***	-0.0211***	-0.0210***
Stand. Err.	(0.0055)	(0.0055)	(0.0072)	(0.0072)
Relative effect	-2.35%	-2.35%	-2.11%	-2.10%
	Panel B. With controls			
Coefficient	-0.0153***	-0.0153***	-0.0202***	-0.0202***
Standard Error	(0.0050)	(0.0050)	(0.0064)	(0.0064)
Relative effect	-2.24%	-2.24%	-2.02%	-2.02%
Number of Observations	256,805			

Notes: This table presents the main results using covariance index. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; All regressions control for draw-by-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using the covariance index as an outcome instead of unhealthy. The results are also robust to the inclusion/exclusion of controls and how we measure quality. Numbers in parentheses are LHC-level clustered standard errors. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.12: Main estimates using a Logit model

	Unhealthy		LBW		Prematurity		Apgar < 7	
	Average Health Scores (1)	PCA Health Scores (2)	Average Health Scores (3)	PCA Health Scores (4)	Average Health Scores (5)	PCA Health Scores (6)	Average Health Scores (7)	PCA Health Scores (8)
Panel A. Without controls								
Coefficient	-0.0061***	-0.0061***	-0.0032**	-0.0032**	-0.0034**	-0.0034**	-0.0028**	-0.0028**
Stand. Err.	(0.0020)	(0.0020)	(0.0014)	(0.0014)	(0.0015)	(0.0015)	(0.0014)	(0.0014)
Relative effect	-6.42%	-6.39%	-7.52%	-7.40%	-8.26%	-8.22%	-7.52%	-7.56%
Panel B. With controls								
Coefficient	-0.0056***	-0.0056***	-0.0036***	-0.0036***	-0.0034***	-0.0034***	-0.0024**	-0.0023**
Standard Error	(0.0017)	(0.0017)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Relative effect	-5.91%	-5.91%	-8.36%	-8.36%	-8.27%	-8.36%	-6.32%	-6.26%
Average Dependent Variable	0.095		0.043		0.041		0.037	
Number of Observations	256,602							

Notes: This table presents the main results using a Logit model. The coefficients represent the average marginal effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. “Unhealthy” is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; “LBW” is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; “Prematurity” is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; “Apgar < 7” is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Regressions for the coefficients labeled as “With controls” also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using an analogous Logit model and compute the average marginal effect associated with an increase in one standard deviation of the skill measure. The results are also robust to the inclusion/exclusion of controls and how we measure quality. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.13: Main estimates linearity

		Unhealthy	
		Average Health Scores (1)	PCA Health Scores (2)
Quartile 2	Coefficient	-0.0066*	-0.0070*
	Stand. Err.	(0.0036)	(0.0040)
	Relative effect	-6.95%	-7.36%
Quartile 3	Coefficient	-0.0082**	-0.0079**
	Stand. Err.	(0.0039)	(0.0040)
	Relative effect	-8.64%	-8.33%
Quartile 4	Coefficient	-0.0133***	-0.0134***
	Stand. Err.	(0.0036)	(0.0035)
	Relative effect	-13.98%	-14.09%

Notes: This table presents estimates using the quartiles of the quality distribution. The coefficients represent the effect of being assigned a physician of the quartiles 2, 3, or 4 of the physicians' quality distribution compared to being assigned a physician from the first quartile. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise. All regressions control for draw-by-state fixed effects. Numbers in parentheses are LHC-level clustered standard errors. While not all the coefficients are statistically different from each other, we do observe increases in the point estimates associated with higher quartiles and cannot discard linearity of the effects.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.14: Main estimates without, with dummy and continuous controls

	Unhealthy		LBW		Prematurity		APGAR <7	
	Average Scores	PCA Scores	Average Scores	PCA Scores	Average Scores	PCA Scores	Average Scores	PCA Scores
				Panel A. Without controls				
Coefficient	-0.0062***	-0.0062***	-0.0034**	-0.0033**	-0.0033**	-0.0033**	-0.0028**	-0.0029**
Stand. Err.	(0.0021)	(0.0021)	(0.0016)	(0.0017)	(0.0016)	(0.0015)	(0.0014)	(0.0014)
Relative effect	-6.52%	-6.50%	-7.95%	-7.85%	-8.13%	-8.10%	-7.58%	-7.64%
				Panel B. With dummy controls				
Coefficient	-0.0059***	-0.0059***	-0.0036**	-0.0036**	-0.0033**	-0.0033**	-0.0023*	-0.0023*
Standard Error	(0.0019)	(0.0019)	(0.0014)	(0.0015)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
Relative effect	-6.24%	-6.24%	-8.52%	-8.51%	-8.02%	-8.10%	-6.26%	-6.24%
				Panel C. With continuous controls				
Coefficient	-0.0060***	-0.0060***	-0.0030**	-0.0030**	-0.0028**	-0.0028**	-0.0032***	-0.0032***
Standard Error	(0.0017)	(0.0017)	(0.0014)	(0.0014)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Relative effect	-6.34%	-6.35%	-7.06%	-7.06%	-6.75%	-6.88%	-8.49%	-8.44%
Average Dependent Variable	0.095	0.095	0.043	0.043	0.041	0.041	0.037	0.037
Number of Observations				256,805				

Notes: This table presents the main results without controls, and with dummy and continuous controls. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score and the first principal component of the health test scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-by-state fixed effects. Regressions for the coefficients labeled as "With dummy controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Regressions for the coefficients labeled as "With continuous controls" include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; the LHC's low birth weight average measured in 2010-2012; the LHC's prematurity percentage measured in 2010-2012; and the LHC's Apgar average measured in 2010-2012 and zero otherwise . These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.15: Interaction between cohort scores and program scores

		Unhealthy	**LBW**	**Prematurity**	**Apgar < 7**
Average Health Score	Coeff.	-0.0068***	-0.0028**	-0.0024	-0.0043**
	Stand. Err.	(0.0025)	(0.0014)	(0.0016)	(0.0020)
	Relative effect	-7.20%	-6.51%	-5.80%	-11.50%
Program Average	Coeff.	0.0014	0.0000	-0.0017	0.0027
	Standard Error	(0.0028)	(0.0016)	(0.0017)	(0.0021)
	Relative effect	1.45%	-0.05%	-4.09%	7.27%
Av. Health Sc. x Prog. Av.	Coeff.	-0.0001	0.0011	0.0005	-0.0006
	Standard Error	(0.0017)	(0.0013)	(0.0008)	(0.0014)
	Relative effect	-0.13%	2.64%	1.11%	-1.57%
Average Dependent Variable		0.095	0.043	0.041	0.037
Number of Observations			256,805		

Notes: This table presents the main results using the interaction between cohort and program scores. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score) and higher average program quality. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. These results show that the effects presented in Table 4 are not driven by top-ranked universities. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.16: Main results using municipalities with one LHC

	Unhealthy		**LBW**		**Prematurity**		**Apgar < 7**	
	Average Health Scores	PCA Health Scores	Average Health Scores	PCA Health Scores	Average Health Scores	PCA Health Scores	Average Health Scores	PCA Health Scores
				Panel A. Without controls				
Coefficient	-0.0064***	-0.0064***	-0.0034*	-0.0033*	-0.0037**	-0.0037**	-0.0029**	-0.0029**
Stand. Err.	(0.0022)	(0.0022)	(0.0018)	(0.0018)	(0.0017)	(0.0016)	(0.0013)	(0.0013)
Relative effect	-6.66%	-6.66%	-7.78%	-7.64%	-8.80%	-8.80%	-7.63%	-7.76%
				Panel B. With controls				
Coefficient	-0.0055***	-0.0055***	-0.0033**	-0.0033**	-0.0033**	-0.0034**	-0.0021	-0.0021
Standard Error	(0.0020)	(0.0020)	(0.0016)	(0.0016)	(0.0014)	(0.0014)	(0.0013)	(0.0013)
Relative effect	-5.75%	-5.77%	-7.57%	-7.54%	-8.01%	-8.13%	-5.62%	-5.66%
Average Dependent Variable	0.096	0.096	0.043	0.043	0.042	0.042	0.037	0.037
Number of Observations					237,082			

Notes: This table presents the main results using municipalities with only one LHC. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. The table shows that the results presented in Table 4 are almost identical if we exclude from our main sample the ten municipalities with more than two LHCs per municipality. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.17: Main results using the weighted score without and with controls

	Unhealthy	**LBW**	**Prematurity**	**Apgar < 7**
	Average Health Scores			
	(1)	(2)	(3)	(4)
	Panel A. Without controls			
Coefficient	-0.0059***	-0.0032**	-0.0033**	-0.0026**
Stand. Err.	(0.0020)	(0.0015)	(0.0015)	(0.0013)
Relative effect	-6.14%	-7.44%	-7.85%	-7.03%
	Panel B. With controls			
Coefficient	-0.0057***	-0.0035**	-0.0032**	-0.0022*
Standard Error	(0.0017)	(0.0014)	(0.0013)	(0.0012)
Relative effect	-5.75%	-7.57%	-8.01%	-5.62%
Average Dependent Variable	0.096	0.043	0.042	0.037
Number of Observations	237,082			

Notes: This table presents the main results using weighted score. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. "Unhealthy" is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7, and zero otherwise; "LBW" is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; "Prematurity" is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; "Apgar < 7" is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the sex of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is 19 years old or younger and zero otherwise; marital status, number of inhabitants in the municipality; number of LHCs per municipality; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. The table shows that the results are very similar when the weighted score is used as a proxy of physicians' skills. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are LHC-level clustered standard errors.

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

Table A.18: Main estimates by levels of exposure

	Unhealthy		**LBW**		**Prematurity**		**APGAR <7**	
	Exp>=100	**Exp<100**	**Exp>=100**	**Exp<100**	**Exp>=100**	**Exp<100**	**Exp>=100**	**Exp<100**
				Panel A. Without controls				
Coefficient	-0.0039*	-0.0029**	-0.0023*	-0.0013	-0.0030*	-0.0018**	-0.0019	-0.0011
Stand. Err.	(0.0022)	(0.0013)	(0.0013)	(0.0008)	(0.0018)	(0.0008)	(0.0016)	(0.0008)
Relative effect	-3.67%	-3.28%	-4.51%	-3.21%	-5.68%	-4.94%	-5.03%	-3.30%
				Panel B. With controls				
Coefficient	-0.0041**	-0.0029**	-0.0022*	-0.0013*	-0.0031*	-0.0019***	-0.0021	-0.0010
Standard Error	(0.0020)	(0.0012)	(0.0012)	(0.0007)	(0.0016)	(0.0008)	(0.0015)	(0.0007)
Relative effect	-3.84%	-3.26%	-4.37%	-3.28%	-5.84%	-5.32%	-5.42%	-2.82%
Average Dependent Variable	0.106	0.089	0.051	0.040	0.053	0.037	0.039	0.035
Number of Observations	70,021	400,216	70,021	400,216	70,021	400,216	70,021	400,216

Notes: This table presents the main results splitting the sample between children whose gestation period was completely included in the SSO period of the cohort (Exp>=100) and those who were only partially treated by the cohort with and without controls. The coefficients represent the effect of being assigned a physician with one standard deviation higher quality (proxied by the average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as “With controls” also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of LHCs per municipality; area; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of low birth weight measured in 2010-2012 and zero otherwise; an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise; and an indicator variable that takes the value of 1 if the LHC is above the 75th percentile of the distribution of the Apgar score measured in 2010-2012 and zero otherwise. Results are marginally stronger for newborns with higher exposure, however we are unable to identify statistically significant differences.Numbers in parentheses are LHC-level clustered standard errors.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$